LEVERAGING FUTURE RELATIONSHIP REASONING FOR VEHICLE TRAJECTORY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the interaction between multiple agents is crucial for realistic and plausible vehicle trajectory prediction. Accordingly, existing methods tried to model and predict the interaction using observed past trajectories of agents with pooling, attention, or graph-based methods. However, we observed that they easily fail under complex road structures. It is because they do not explicitly utilize the map information for predicting the relationship, and they only model the relationship between vehicles in a deterministic manner, not in a stochastic manner. In this paper, we propose a new method to model a stochastic future relationship among agents using the lane information contained in the map. Our method first predicts a probability of lane-level waypoint occupancy of vehicles. Then the proposed method utilizes the temporal probability of passing the adjacent lanes to learn the interaction between agents. In addition, we model the interaction using probabilistic distribution. This distribution is learned by the posterior distribution of interaction from GT future trajectory. We validate our method on popular trajectory prediction datasets: nuScenes and Argoverse. The code will be available in public upon acceptance.

1 INTRODUCTION

For safe autonomous driving, predicting the future trajectory of a vehicle is very important. Early heuristic prediction models utilized only the past trajectory of the target vehicle (Zeng et al. (2020); Rhinehart et al. (2018)). Recently, thanks to deep learning, more accurate prediction can be made by considering not only its own state but also the relationship with the High-Definition (HD) map (Liang et al. (2020); Zeng et al. (2021)) or surrounding agents (Lee et al. (2017); Chandra et al. (2020)). Since surrounding vehicles are not stationary, predicting relationships with them is much more complicated and has become essential for realistic trajectory prediction. In addition, since an individual driver controls each vehicle, its interaction has stochastic nature.

Previous works utilized past trajectories of the surrounding vehicles to model interaction by pooling, operating multi-head attention, or adapting spatio-temporal graph methods. However, we observed that they easily fail under complex road structures. On the left of Fig. 1, the past trajectories of agents are shown, and on the right is a plot showing the attention weight among agents obtained by the previous method (Mercat et al. (2020)) that learned the interaction among agents using multi-head attention (MHA). Since agent 0 and agent 4 are expected to join in the future, the attention weight between the two is expected to be high. However, the model predicts low attention weight between them. It shows the difficulty of reasoning future relationships between agents based only on past trajectories. The reasoning should have been much easier if it leveraged the road structure.

First, let us recall the decision-making process of human drivers. They first set their goal and roughly infer how the surrounding vehicles will behave *in the future* while they head to their own goal. After that, they infer the interaction with other agents by inferring how likely the future path of other vehicles will overlap the path set by themselves. We define the interaction from this process as a "*Future Relationship*" to mimic the decision-making process of human drivers. We use the following approaches to model and train the Future Relationship, as shown in Fig. 3.

First, because two vehicles in adjacent locations and headings will have much influence on each other, a future motion of each vehicle should be obtained. Since vehicles mainly move along lanes, we utilize lane information as strong prior for representing the rough future motion of vehicles. Because

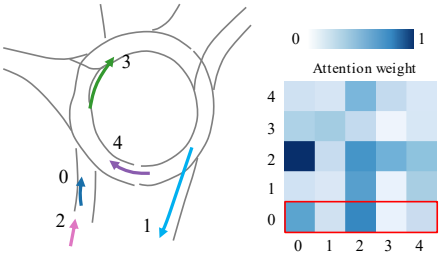

Figure 1: Past motions and their attention map from previous work (Mercat et al. (2020)). Weak relationship is inferred between agents that will highly interact in the future: agent 0 and 4.

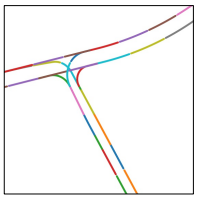

Figure 2: Lane segments represented in different colors.

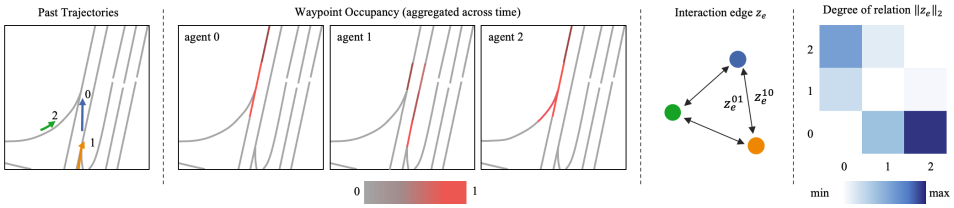

Figure 3: Key concept of the proposed method. Interaction between agents is defined as an edge. The more probability two agents will pass an adjacent lane in the future, the more interaction is expected.

lanes contain both location and direction, rough future motion can be represented as waypoint occupancy. The waypoint occupancy is the probability of a vehicle passing a specific lane segment at every timestep. Fig. 3 shows the waypoint occupancy of each agent. The probability that the vehicle passes that lane during the prediction horizon is drawn using the tone of red color.

Second, based on the waypoint occupancy, we infer the Future Relationship in probabilistic distribution. In most vehicle trajectory prediction tasks, the interaction between agents is still made in a deterministic manner. However, we tackle that interaction between vehicles is highly stochastic, and there can be multiple possible interactions. The deterministic relation inference averages out diverse interactions, interrupting socially-aware trajectory prediction. Therefore, we define Future Relationship in Gaussian Mixture (GM) distribution. Motivated by Neural Relational Inference (NRI) (Kipf et al. (2018)), we propose a method to train the diverse interaction distribution explicitly.

In summary, our contributions are:

1) We propose to infer Future Relationship using lane information, assuming that vehicles generally follow the lane.

2) We propose the model to infer the distribution of interaction as Gaussian Mixture and propose a new learning method to make the model learn the distribution from the posterior.

3) The proposed method is validated on popular real-world vehicle trajectory datasets: nuScenes and Argoverse. In both datasets, there was a remarkable improvement in prediction performance and achieved state-of-the-art performance in nuScenes dataset.

2 Related Work

2.1 Goal-conditioned trajectory prediction

Predicting future trajectories at once is a challenging task. Instead, the first goal-conditional prediction, which oversamples goal candidates and selects a final set of goals, is helpful and has shown state-of-the-art performance (Zhao et al. (2021); Gu et al. (2021); Phan-Minh et al. (2020); Chai et al. (2019); Zhang et al. (2021)), especially in long-range prediction tasks. CoverNet (Phan-Minh et al. (2020)) and MultiPath (Chai et al. (2019)), which quantize the trajectory space to a set of anchors, often generate map-agnostic trajectories that cross non-drivable areas because the surrounding map is not considered. Recent studies have exploited the map information to obtain more performant goal

candidates based on the assumption that vehicles follow lanes. TNT (Zhao et al. (2021)) uses goal points sampled from a lane centerline; GoalNet (Zhang et al. (2021)) uses lane segments as anchors and predicts multiple trajectories based on each anchor. Similarly, we set lane segments with their connectivity with nearby segments as goal candidates to model the inherent uncertainty of trajectory prediction. The relationship between lane segments, such as predecessor and successor, provides global connectivity, enabling the model to predict realistic and plausible trajectories.

2.2 INTERACTION MODELING

Trajectory prediction considering the interaction between agents helps to predict socially aware trajectory. In the very early stage, modeling interaction using a global pooling layer (Deo & Trivedi (2018); Gupta et al. (2018)) were proposed. In another way, researchers attempt to obtain interaction through attention-based (Ngiam et al. (2022); Mercat et al. (2020); Vemula et al. (2018)) and GNN-based method (Carrasco et al. (2021); Cao et al. (2021); Zeng et al. (2021); Casas et al. (2020); Liang et al. (2020); Gao et al. (2020)). In the previous method, interactions between agents are learned only with regression loss, which is insufficient to represent dynamic and rapidly changing situations. Besides, a line of works using Neural Relational Inference (NRI) that explicitly predicts and learn interaction using a latent interaction graph was proposed (Kipf et al. (2018)). EvolveGraph (Li et al. (2020)) utilizes two interaction graphs, static and dynamic, and NRI-MPM (Chen et al. (2021)) uses a relation interaction mechanism and spatio-temporal message passing mechanism. Similarly, we apply the NRI-based method to predict future trajectory explicitly in complex road structures.

2.3 MULTI-MODAL TRAJECTORY PREDICTION

Due to the nature of the trajectory prediction; stochasticity, there is no unique answer, but several futures are possible. Therefore, recent works use deep generative models such as GAN (Goodfellow et al. (2014)) and VAE (Kingma & Welling (2013)). GAN-based (Gupta et al. (2018); Kosaraju et al. (2019); Li et al. (2021b)) and CVAE-based models (Ivanovic & Pavone (2019); Salzmann et al. (2020); Tang & Salakhutdinov (2019)) predict multiple futures by sampling multiple latent vectors. As it is important to sample a meaningful latent vector for predicting diverse and plausible future trajectories, trying to define a well-organized latent space became a natural choice (Ma et al. (2021); Bae et al. (2022)). The most closely related work to ours is GRIN (Li et al. (2021a)), which argues that multi-modality in trajectory prediction comes from two sources: i) personal intention and ii) social relations with other agents. GRIN only uses the past trajectories of agents to model the latent space of personal intention and social relations. However, our model propose to utilize lane information for modeling social relations based on the occupancy. Leveraging the future occupancy makes the model interpretable and easy to be trained since the pseudo labels of occupancy can be obtained from ground-truth trajectories. Also, we sample the personal intention of a driver from a set of lane segments feature by following the standard goal-conditioned method (Zhao et al. (2020)).

3 FORMULATION

In each scene, the observed past trajectory and GT future trajectory of N vehicles are denoted as $\mathbf{x}^- = \{\mathrm{x}\}_{-t_p:0}^{1:N}$ and $\mathbf{x}^+ = \{\mathrm{x}\}_{1:t_f}^{1:N}$ where $t_p + 1$ and t_f are past/future timesteps. We use lane information from the HD map, expressed as M segmented lane polylines (Zhao et al. (2021); Varadarajan et al. (2022)). The lane information is represented as a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\ell}, \mathbf{e})$ where the node $\boldsymbol{\ell} = \{\ell\}^{1:M}$ is the set of directed sequences of xy coordinates. The segmented lane nodes are shown in Fig. 2, where different colors denote different individual segments (ℓ^m). The edge is defined as five types: $\mathbf{e} = \{e_{pre}, e_{suc}, e_{left}, e_{right}, e_{inter}\}^{1:M}$ which are predecessor, successor, left/right neighbor and in-the-same-intersection. We denote overall input as $\mathbf{X} = \{\mathbf{x}^-, \boldsymbol{\mathcal{G}}\}$. The output is denoted as $\mathbf{Y}$, which is the set of F samples of xy coordinates of the target vehicles. We also define the future lane occupancy (waypoint occupancy) denoted as $\boldsymbol{\tau}_t = \prod_{n,m}^{N,M} p_t^n(\ell^m)$, where $p_t^n(\ell^m)$ is the probability of the m-th lane segment being occupied by the n-th vehicle at future time step t. The waypoint occupancy during future timestemps (t_f) predicted from $\mathbf{X}$ is denoted by $\boldsymbol{\tau}_{1:t_f}^-$ and GT waypoint occupancy obtained from GT future trajectory is denoted by $\boldsymbol{\tau}_{1:t_f}^+$.

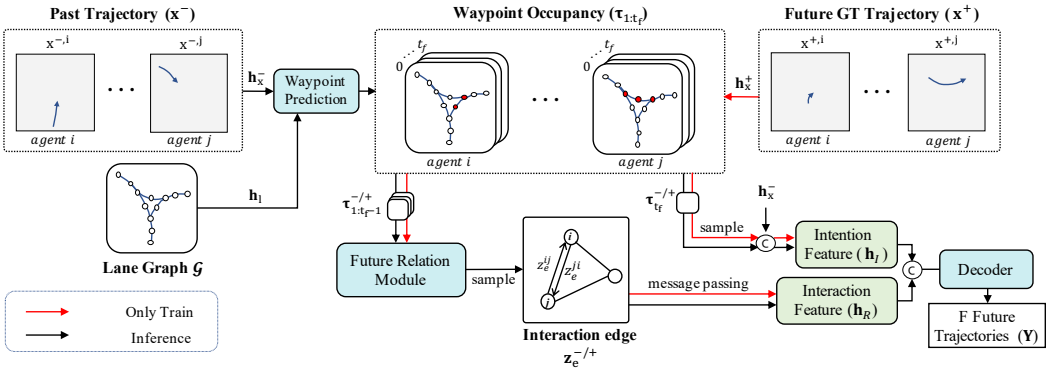

Figure 4: Overall structure. With prediction from past motions or directly from future motions, waypoint occupancy ($\boldsymbol{\tau}_{1:t_f}$) is obtained. Goal lane features sampled from $\boldsymbol{\tau}_{t_f}$ and past motion feature (h_x^-) is aggregated to obtain the intention feature. $\boldsymbol{\tau}_{1:t_f-1}$ is fed to FRM then interaction edge among agents are sampled. With the edge, message passing is performed to obtain the interaction feature. Decoder predicts F future trajectories from the concatenation of intention and interaction feature.

4 METHOD

We focus on modeling the "Future Relationship" between agents. The naive method to infer the Future Relationship is predicting all future vehicle trajectories and then using the similarity among the trajectories. However, this method needs to perform prediction twice, which is inefficient and redundant. Also, the criteria for calculating the similarity between trajectories are unclear. Inspired by the idea that the vehicles mainly follow lanes, we predict the interaction between vehicles using the lane information as a strong prior of interaction modeling. The key idea is that *if two agents are expected to pass an adjacent lane, they will have a high chance of interacting in the future*.

The overall structure is shown in Fig. 4. First, we predict the waypoint occupancy, the probability of the lane segment where the vehicle will pass during the future time step 1:t_f (Sec. 4.1). Then, from each vehicle's waypoint occupancy, the FRM infers interaction features between agents as Gaussian Mixture (Sec. 4.2). Next, a message passing through inferred interaction features is performed to obtain interaction feature. Finally, in the decoding stage (Sec. 4.3), the decoder predicts future trajectories from the aggregation of the interaction feature and intention feature. Here, the intention feature is the concatenation of past motion and goal lane features. Our method follows the CVAE scheme (Sec. 4.2.1, 4.2.2), where the condition corresponds to the intention feature, and the latent code corresponds to the interaction feature. We describe the details in the following sections.

4.1 WAYPOINT OCCUPANCY

Past trajectories $\mathbf{x}^-$ and lane graph $\mathcal{G}$ are encoded into past motion and lane features: $\boldsymbol{h}_\mathrm{x}^-$ and $\boldsymbol{h}_\ell$. The same motion encoder is used to obtain future motion features $\boldsymbol{h}_\mathrm{x}^+$. Similar to TNT (Zhao et al. (2021)), we predict waypoint occupancy in categorical distribution from past motion and lane features.

$$\boldsymbol{\tau}_{1:t_f}^- = \mathrm{softmax}(\mathrm{MLP}([\boldsymbol{h}_\mathrm{x}^-, \boldsymbol{h}_\ell])) \quad \in \mathbb{R}^{N \times M \times (t_f - 1)} \tag{1}$$

where $[\cdot, \cdot]$ is concatenation. As waypoint occupancy is a probability, we apply a softmax so that it becomes $\sum^M \tau_t^m = 1$. Then goal feature ($\boldsymbol{h}_g^-$) is sampled from $\boldsymbol{h}_\ell$ following waypoint occupancy at t_f ($\boldsymbol{\tau}_{t_f}^-$). The concatenation of $\boldsymbol{h}_g^-$ and $\boldsymbol{h}_\mathrm{x}^-$ *is interpreted as the agents' intentions* to reach, representing the intention feature ($\boldsymbol{h}_I^-$). Besides, $\boldsymbol{\tau}_{1:t_f-1}^-$ is a temporal feature containing the probabilistic future motion, so it is passed to FRM.

For GT waypoint occupancy ($\boldsymbol{\tau}_{1:t_f}^+$), we can obtain it directly from GT future trajectory. We can get binary lane occupancy of vehicles as we know the position and heading of vehicles at future timesteps. The details about getting GT waypoint occupancy are described in the supplementary material. Similarly, the unique GT goal feature($\boldsymbol{h}_g^+$) also can be obtained by indexing from $\boldsymbol{h}_\ell$.

4.2 FUTURE RELATIONSHIP MODULE (FRM)

Attaining Future Relationship is depicted in Fig. 5. At first, waypoint occupancy is smoothed by Graph Convolutional Network (GCN) (Welling & Kipf (2016)). It is because when a vehicle passes a specific lane, the vehicle will affect other vehicles that pass the adjacent lane, not necessarily the same lane. Therefore, we apply different smoothing for each lane connectivity (predecessor, successor, neighbor, in-the-same-intersection) by employing 2-hop GCN layers. Each layer is expressed as:

$$\tilde{\boldsymbol{\tau}}_{1:t_f-1} = \sum_{\substack{e \in \{succ, pred, \\ right, left, inter\}}} \sigma\left(D_e^{-1} A_e \boldsymbol{\tau}_{1:t_f-1} W_e\right) \tag{2}$$

Here, σ is softmax followed by ReLU activation, $D_e \in \mathbb{R}^{M \times M}$ is a degree matrix, $A_e \in \mathbb{R}^{M \times M}$ is an adjacency matrix, and $W_e \in \mathbb{R}^{t_f-1 \times d}$ is a weight matrix for each edge type.

Next, with the obtained waypoint occupancy, we compute the inter-agent proximity (PR $= \{pr^{ij}\}^{1:N,1:N}$) between vehicle pairs which is the temporal probability that two agents would occupy adjacent lanes. The PR can be obtained with a cross product of $\tilde{\boldsymbol{\tau}}_{1:t_f-1}$ across the lane axis. It is visually drawn in the middle of Fig. 5.

$$\text{PR} = \tilde{\boldsymbol{\tau}}_{1:t_f-1} \cdot (\tilde{\boldsymbol{\tau}}_{1:t_f-1})^\top \quad \in \mathbb{R}^{N \times N \times (t_f-1)} \tag{3}$$

4.2.1 PRIOR OF THE RELATION

Our method follows CVAE scheme as mentioned above, the prior distribution of relation should be computed. We utilize past motion features ($\boldsymbol{h}_{\text{x}}^-$) and predicted inter-agent proximity (PR$^-$) to obtain the prior. In this subsection, we omit superscript - for simplification. There are two design factors for our interaction modeling: i) interaction should reflect diverse and stochastic properties, and ii) interaction occurs in every pair of vehicles. As a result, the prior of relation is assumed to be Gaussian Mixture (GM) per agent pairs. Then, we define interaction edge e^{ij} between agent i and agent j as a d-dimensional feature that has GM distribution ($p_\theta(\boldsymbol{e}|\mathbf{X}) \sim \prod_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}\boldsymbol{\sigma}_k^2)$) following GMVAE (Dilokthanakul et al. (2016)). The distribution parameters ($\boldsymbol{\mu}_K, \boldsymbol{\sigma}_K, \boldsymbol{\pi}_K = \{\boldsymbol{\mu}_K^{ij}, \boldsymbol{\sigma}_K^{ij}, \boldsymbol{\pi}_K^{ij}\}^{1:N,1:N}$) are obtained from the neural network F_θ:

$$\boldsymbol{\mu}_K^{ij}, \boldsymbol{\sigma}_K^{ij}, \boldsymbol{\pi}_K^{ij} = F_\theta([pr^{ij}, h_{\text{x}}^i, h_{\text{x}}^j]) \quad \in \mathbb{R}^{Kd}, \mathbb{R}^{Kd}, \mathbb{R}^K \tag{4}$$

F_θ is composed of MLP layers and 1-d conv layer to capture temporal dependency (Deo & Trivedi (2018)). Then, we perform sampling twice: one for the interaction mode k (from $\boldsymbol{\pi}_K$) and one for ϵ (from Gaussian noise). It allows having K distinct interactions where each interaction follows Gaussian distribution:

$$\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k = \text{argmax}_k(\boldsymbol{\pi}_K + \boldsymbol{g}), \quad \boldsymbol{g} \sim \text{Gumbel}(0, 1) \tag{5}$$

$$\boldsymbol{z_e} = \boldsymbol{\mu}_k + \boldsymbol{\sigma}_k \boldsymbol{\epsilon} \quad \in \mathbb{R}^{N \times N \times d}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

We then compute interaction feature ($\boldsymbol{h}_R^- = \{h_R^i\}^{1:N}$) by message passing from sampled interaction edge. Because we define interaction as a d-dimensional feature, the message function is composed of summation followed by Hadamard product ($\otimes$) of interaction edge and the other agent features:

$$h_R^i = \sigma'(\frac{1}{N-1} \sum_{j \neq i}^{N} z_e^{ij} \otimes F_p(h_{\text{x}}^j)) \quad \in \mathbb{R}^d \tag{7}$$

Stochasticity is obtained by giving flexibility within a limited range by predicted mean and standard deviation. It is because, unlike a symmetric and deterministic interaction between particles, such as pull or push (Kuramoto (1975)), each driver can asymmetrically consider stochastic interaction.

4.2.2 CVAE POSTERIOR

To obtain the posterior distribution of relation, the future motion feature ($\boldsymbol{h}_{\text{x}}^+$) and GT waypoint occupancy ($\boldsymbol{\tau}^+$) are used. Similarly, we omit superscript + in this subsection for simplification. Inter-agent proximity is obtained by following same procedure in Eqs. (2)-(3). The difference from the prior is that the posterior is modeled in a single Gaussian distribution. Thus, F_θ is replaced with F_ϕ which outputs parameters of Gaussian distribution ($\boldsymbol{\mu}, \boldsymbol{\sigma} = \{\mu^{ij}, \sigma^{ij}\}^{1:N,1:N}$):

$$\mu^{ij}, \sigma^{ij} = F_\phi([pr^{ij}, h_{\text{x}}^i, h_{\text{x}}^j]) \quad \in \mathbb{R}^d, \mathbb{R}^d \tag{8}$$

Then we sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and interaction edge is obtained: $\boldsymbol{z_e} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$. Following the same message passing, interaction features ($\boldsymbol{h}_R^+$) is obtained.

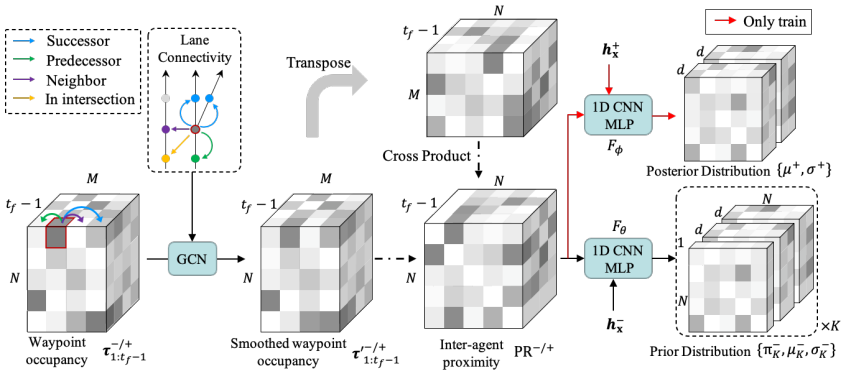

Figure 5: Future Relationship Module. During inference, inferred waypoint occupancy of agents ($\boldsymbol{\tau}^{-}_{1:t_f-1}$) is fed to GCN. Then, it is cross-produced by itself to get inter-agent proximity (PR^{-}). Prior of interaction ($\boldsymbol{\pi}_K, \boldsymbol{\mu}_K, \boldsymbol{\sigma}_K$) is then obtained. During training, GT waypoint occupancy ($\boldsymbol{\tau}^{+}_{1:t_f-1}$) is fed to get posterior of interaction ($\boldsymbol{\mu}, \boldsymbol{\sigma}$). Prior is modeled in Gaussian Mixture, and Posterior in Gaussian distribution.

4.3 Decoder

The decoder predicts future trajectories from the aggregation of the interaction feature ($\boldsymbol{h}_R$) and intention feature ($\boldsymbol{h}_I$). GT goal feature ($\boldsymbol{h}_g^{+}$) is concatenated to the past motion feature ($\boldsymbol{h}_{\mathrm{x}}^{-}$) to get a unique GT intention feature. The GT intention feature is repeated F times, and we sample the interaction feature ($\boldsymbol{h}_R^{+}$) F times from the posterior distribution. During inference, the goal feature ($\boldsymbol{h}_g^{-}$) is sampled F times. The intention feature is obtained by concatenation of sampled goal features and the past motion feature ($\boldsymbol{h}_{\mathrm{x}}^{-}$). The interaction feature ($\boldsymbol{h}_R^{-}$) is sampled F times from prior distribution. The decoder is modeled by 2-layer MLP and predicts the sequence of x,y coordinates from the concatenation of the intention and interaction feature.

4.4 Training

Because the ground ($\boldsymbol{\tau}^{+}$) is available for waypoint occupancy, we train $\boldsymbol{\tau}^{-}$ with supervision. Since the waypoint occupancy is modeled as a classification per agent, we learn it using negative log-likelihood (NLL): $\mathcal{L}_{nll} = -\boldsymbol{\tau}^{+}_{1:t_f} log(\boldsymbol{\tau}^{-}_{1:t_f})$.

In contrast, because interaction $\boldsymbol{z}_e$ is unobservable, we optimize the negative evidence lower bound (ELBO) to learn the distribution of interaction following the CVAE scheme.

$$ELBO = -\mathbb{E}[log(p(\mathbf{x}^{+} \mid \mathbf{X}, \boldsymbol{z}_e, \boldsymbol{\tau})] + KL[q_\pi(\boldsymbol{z}_e \mid \mathbf{Y}, \boldsymbol{\tau}^{+}) \parallel p_\theta(\boldsymbol{z}_e \mid \mathbf{X}, \boldsymbol{\tau}^{-})] \tag{9}$$

where q_π is an approximate posterior distribution of relation, and p_θ is the prior. Because our model only allows interaction posterior to be Gaussian distribution, the Kullback–Leibler (KL) divergence term can be simplified below. The detailed derivation with the reparameterization trick is included in the supplementary material.

$$\mathcal{L}_{KL} = -KL[q_\pi \parallel p_\theta] \approx log \sum_k \boldsymbol{\pi}_k \exp(-KL[q_\phi \parallel p_{\theta,k}]) \tag{10}$$

The common drawback of NRI method is the "degenerate" problem which leads the decoder to ignore the relation edge during training. So we train the network to that two features, intention and interaction, to have different roles. Because the GT trajectory is conditioned on the GT goal feature, we use the GT goal feature to compute the reconstruction term. Therefore, training in that way restricts the role of interaction edge to be narrowed down to momentary motion: $\mathcal{L}_{recon} = \min_{\boldsymbol{e}}\{\mathbb{E}[log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{e}, \boldsymbol{\tau}^{+})]\}$.

The final loss is the sum of the three losses, and they are trained jointly: $\mathcal{L} = \mathcal{L}_{nll} + \mathcal{L}_{KL} + \mathcal{L}_{recon}$.

Table 1: Comparison on nuScenes test set. Best in **bold**, second best in underline.

Paper	$mADE_5$	$mADE_{10}$	MR_5	MR_{10}	$mFDE_1$
Trajectron++ Salzmann et al. (2020)	1.88	1.51	0.70	0.57	9.52
P2T Deo & Trivedi (2020)	1.45	1.16	0.64	0.46	10.5
AgentFormer Yuan et al. (2021)	1.86	1.45	-	-	-
LaPred Kim et al. (2021)	1.47	1.12	0.53	0.46	8.37
MultiPath Chai et al. (2020)	1.44	1.14	-	-	7.69
GOHOME Gilles et al. (2022a)	1.42	1.15	0.57	0.47	6.99
Autobot Girgis et al. (2021)	1.37	1.03	0.62	0.44	8.19
THOMAS Gilles et al. (2022b)	1.33	1.04	0.55	0.42	6.71
PGP Deo et al. (2022)	1.27	0.94	0.52	0.34	7.17
Ours	**1.18**	**0.88**	**0.48**	**0.30**	**6.59**

Table 2: Comparison on Argoverse val/test set. Best in **bold**, second best in underline.

Paper	Val set		Test set	
	$mADE_6$	$mFDE_6$	$mADE_6$	$mFDE_6$
TNT Zhao et al. (2021)	0.73	1.29	0.94	1.54
DenseTNT Gu et al. (2021)	0.73	1.05	0.88	1.28
LaneRCNN Zeng et al. (2021)	0.77	1.19	0.90	1.45
TPCN Ye et al. (2021)	0.73	1.15	0.87	1.38
Autobot Girgis et al. (2021)	0.73	1.10	0.89	1.41
mmTransformer Liu et al. (2021)	0.72	1.21	0.84	1.34
SceneTransformer Varadarajan et al. (2022)	-	-	0.80	1.23
Multipath++ Varadarajan et al. (2022)	-	-	0.79	1.21
HiVT Zhou et al. (2022)	0.66	0.96	**0.77**	**1.17**
Baseline	0.71	1.03	0.86	1.30
Ours	0.68	0.99	0.82	1.27

Table 3: Impact of prediction time to the proposed modeling in terms of $mADE_1/mADE_6$.

	Baseline	Ours	Improvement
nuScenes (6sec)	3.23/1.17	2.89/1.10	**10.5**%/6.0%
nuScenes (3sec)	1.26/0.50	1.19/0.48	5.6%/4.0%
Argoverse (3sec)	1.41/0.71	1.33/0.68	5.7%/4.2%

Table 4: Ablation studies on nuScenes validation set.

	F=1 mADE/mFDE	F=5 mADE/mFDE
Impact of model design		
Baseline	3.23/7.60	1.26/2.49
Ours w/o FR	3.21/7.59	1.26/2.50
Ours w/o GCN	3.04/6.94	1.22/2.41
Ours w/ Sym	2.99/6.78	1.22/2.35
Importance of multimodal stochastic interaction		
Ours w/ GP	2.98/6.78	1.20/2.33
Ours w/ Deterministic	2.96/6.80	1.28/2.52
Ours (Full)	**2.89/6.61**	**1.19/2.30**

5 Experiment and Result

We train and evaluate on nuScenes (Caesar et al. (2020)) and Argoverse (Chang et al. (2019)) datasets. The nuScenes/Argoverse datasets provide the 2/2 seconds of past, and 6/3 seconds of future trajectory is predicted with 0.5/0.1 sampling time. Training/validation/test sets consist of real-world driving scenes of 32186/8560/9041 in nuScenes and 205942/39472/78143 in Argoverse. For the baseline model, we follow TNT for goal conditioned model, and MHA encodes interaction from past trajectories. For implementation and training details, refer to the supplementary material.

5.1 Quantitative result

Our method outperforms the SoTA models in all metrics used in the nuScenes benchmark, as shown in Tab. 1. The proposed model outperforms the runner-up method, PGP (Deo et al. (2022)), by a large margin. The superiority implies that explicit interaction modeling via inferring waypoint occupancy helps scene understanding of our model compared to the implicit interaction modeling of PGP. When more samples are predicted, $mADE_{10}$ and MR_{10} showed improvements of 5.3% and 8.8%. THOMAS (Gilles et al. (2022b)) had first placed in $mFDE_1$ by proposing a recombination module that post-processes the marginal predictions into the joint predictions aware of other agents. Our model shows higher performance in $mFDE_1$ than THOMAS, indicating better interaction modeling ability without post-processing. This is possible because inferring future relationships helps to understand the future interaction with other agents; details are provided in the ablation study.

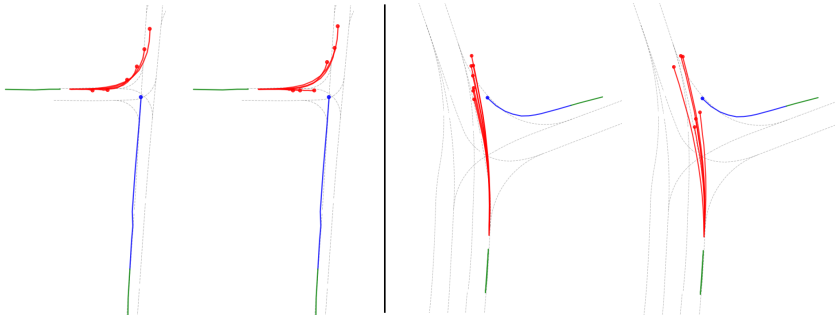

Figure 6: Qualitative results of the proposed method. The green solid line is past trajectories, the red line is 6 predicted samples by baseline (left) and our method (right), the blue line is GT future trajectory of other vehicles. Lane centerlines are in gray dashed lines. In complex road scene, baseline generates spatially uniform samples regardless of interaction with surrounding vehicles. On the other hand, our method generates diverse yet interaction-aware samples: wait or surpass other vehicles that would join in the future.

We also validated our method on the Argoverse dataset. Our model does not achieve SotA performance but still makes remarkable performance improvement in both valid/test sets. Moreover, except for the HiVT, our method make competitive performance in mADE. Please note that our model (0.82) is still comparable to SceneTransformer (0.80) and Multipath++ (0.79) in the test set results. In the case of HiVT, it uses the surrounding vehicles' trajectories for training. It results in increasing the training data, so a direct comparion to HiVT is rather unfair.

We do not achieve SotA in Argoverse because the proposed method is less effective than in nuScenes. We conclude that the relatively small performance gain on Argoverse compared to nuScenes comes from the different configurations of datasets. A model should predict the 6-second future in nuScenes while it only needs to predict the 3-second future in Argoverse. Intuitively, the impact of interaction gets more significant in long-range prediction task. We thus measure the performance gain on nuScenes when the prediction horizon is the same as that of Argoverse and report the results in Tab. 3. In 6-second prediction, the proposed interaction inference method improved mADE_1 by more than 10%, but its effect halves in a 3-second prediction which is a similar effect on Argoverse. This result shows that our interaction inference method is more effective in longer-range prediction tasks.

5.2 Qualitative Result

In Fig. 6, prediction samples (F=6) of the baseline and our method are depicted. To simply check the effect of our method, we took the samples where two agents are in and plotted the prediction of a single agent per scene. Green, blue and red solid lines represent past trajectories, future trajectories of other agents, and prediction samples of target agents. In two scenes, each target agent sets its intention to where the other agent is likely to pass in the future. The prediction samples from our model (right) generate samples that understand and leverage the interaction with other agents. Unlike the baseline method that ignores other agents and generates spatially uniform trajectories, our model considers other agents to surpass or wait for them. Moreover, not only considering multi-modal interaction, we allow stochasticity within a single mode of interaction. As a result, it can be seen that our method generates diverse yet interaction-aware samples.

In addition, our method can consider stochastic interaction when multiple agents are present. Similarly, we predicted the trajectories of a single agent per scene, and we fixed the intention feature of the target vehicle. Two relation edges are sampled with fixed intention, and corresponding predicted trajectories and degree of relation ($\|\boldsymbol{z}_e\|$) are plotted. In the first row, target agent (0) infers large interaction with agent 2 in sample 1. Because agent 2 is moving in the same direction and is predicted to go ahead, the model generates an accelerating trajectory to follow agent 2. In sample 2, in contrast, the interaction between agent 1 is sampled large because they are expected to be in the same intersection. In this case, the model generates decelerating trajectory considering the future motion of agent 1. In addition, with fixed intention, it shows that all generated trajectories are properly on goal lane segments. It means that the proposed training method effectively restricts the role of interaction features affecting momentary motion, as we defined.

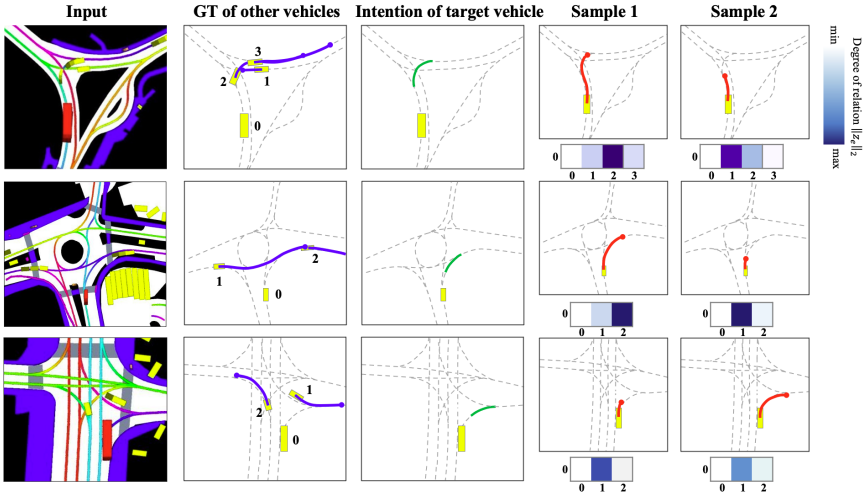

Figure 7: Qualitative results of the proposed method in multi-agent scene.

5.3 Ablation Study

Ablation on the impact of model design is shown in the upper part in Tab. 4. **Ours w/o FR** does not consider Future Relationship in interaction modeling. This model only utilizes past trajectories to infer the relation and does not predict waypoint occupancy. After all, its performance is almost identical to the baseline, which uses MHA of past trajectories to model interaction. This result shows that leveraging Future Relationship indeed helps to infer interaction. **Ours w/o GCN** does not smooth predicted/GT waypoint occupancy. It causes inaccurate inter-agent proximity (PR) estimation especially in the posterior distribution. Because GT waypoint occupancy is a binary value, computing PR from it could cause inaccurate proximity and brings performance decline. In the proposed model, we allow asymmetric relation between two agents. **Ours w/ Sym** applies hard symmetric relation (Li et al. (2019)), and it shows that our asymmetric design is more suitable for modeling the driver relation.

The importance of multi-modal stochastic relation modeling is shown in the lower part of Tab. 4. **Ours w/ GP** models the prior distribution of relation as Gaussian distribution instead of GM. As it only considers a single modality of agent relation, the performance drops compared to when considering multiple modalities. **Ours w/ Deterministic** only predicts the mean of relation feature in Eq. 4. Although it can model multi-modal relation, the diversity is prone to be limited compared to the stochastic counterpart especially when the sample size F is large. The result also shows that deterministic relation modeling severely degrades prediction performance when F is large. On the other hand, **Ours w/ GP** shows relatively less drop in performance as it does not lose stochasticity even after removing the GM prior.

6 Conclusion

In this paper, we propose Future Relationship to learn the interaction between vehicles for trajectory prediction effectively. Beyond utilizing only past trajectory to infer interaction, Future Relationship has advantages in explicitly using lane information. We show that lane information enables proper interaction inference even in complex road structures. In addition, we propose the model that obtains interaction probabilistically, generating diverse yet socially plausible trajectory samples. We trained our model with the CVAE scheme and validated in popular real-world trajectory prediction datasets. The proposed method achieves SotA performance in long-range prediction task, nuScnes, and brings remarkable performance improvement in short-range prediction task, Argoverse. Because considering Future Relationship is novel, we expect that using more sophisticated training methods (Ye et al. (2022); Zhou et al. (2022)) or better baseline model (such as GANet (Wang et al. (2022)) will also lead to performance improvements in our model.

REFERENCES

Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6477–6487, 2022.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628. IEEE Computer Society, 2020.

Defu Cao, Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1839–1845. IEEE, 2021.

Sandra Carrasco, D Fernández Llorca, and MA Sotelo. Scout: Socially-consistent and understandable graph attention network for trajectory prediction of vehicles and vrus. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1501–1508. IEEE, 2021.

Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *2020 European Conference on Computer Vision (ECCV)*. Springer, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *CoRL*, 2019.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning*, pp. 86–99. PMLR, 2020.

Rohan Chandra, Tianrui Guan, Srujan Panuganti, Trisha Mittal, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. *IEEE Robotics and Automation Letters*, 5(3):4882–4890, 2020.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8740–8749. IEEE Computer Society, 2019.

Siyuan Chen, Jiahai Wang, and Guoqing Li. Neural relational inference with efficient message passing mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7055–7063, 2021.

Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1549–15498. IEEE, 2018.

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

Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pp. 203–212. PMLR, 2022.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 9107–9114. IEEE, 2022a.

Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Thomas: Trajectory heatmap output with learned multi-agent sampling. In *International Conference on Learning Representations*, 2022b.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *2021 Advances in neural information processing systems (NeurIPS)*, 2014.

Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15283–15292. IEEE, 2021.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2255–2264, 2018.

Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2375–2384, 2019.

ByeoungDo Kim, Seokhwan Lee, Seong Hyeon Park, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, and Jun Won Choi. Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pp. 14636–14645. IEEE Computer Vision and Pattern Recognition, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pp. 2688–2697. PMLR, 2018.

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International symposium on mathematical problems in theoretical physics*, pp. 420–422. Springer, 1975.

Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 336–345, 2017.

Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020.

Longyuan Li, Jian Yao, Li Wenliang, Tong He, Tianjun Xiao, Junchi Yan, David Wipf, and Zheng Zhang. Grin: Generative relation and intention network for multi-agent trajectory prediction. *Advances in Neural Information Processing Systems*, 34:27107–27118, 2021a.

Xiao Li, Guy Rosman, Igor Gilitschenski, Cristian-Ioan Vasile, Jonathan A DeCastro, Sertac Karaman, and Daniela Rus. Vehicle trajectory prediction using generative adversarial network with temporal logic syntax tree features. *IEEE Robotics and Automation Letters*, 6(2):3459–3466, 2021b.

Yaguang Li, Chuizheng Meng, Cyrus Shahabi, and Yan Liu. Structure-informed graph auto-encoder for relational inference and simulation. In *ICML Workshop on Learning and Reasoning with Graph-Structured Data*, volume 8, pp. 2, 2019.

Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *2020 European Conference on Computer Vision (ECCV)*. Springer, 2020.

Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7573–7582. IEEE Computer Society, 2021.

Yecheng Jason Ma, Jeevana Priya Inala, Dinesh Jayaraman, and Osbert Bastani. Likelihood-based diverse sampling for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13279–13288, 2021.

Jean Mercat, Thomas Gilles, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Multi-head attention for multi-modal joint vehicle motion forecasting. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9638–9644. IEEE, 2020.

Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David J Weiss, Ben Sapp, Zhifeng Chen, and Jonathon Shlens. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *2022 International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=Wm3EA5OlHsG`.

Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14062–14071. IEEE, 2020.

Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. r2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *European Conference on Computer Vision*, pp. 794–811. Springer, 2018.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2020.

Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7814–7821. IEEE, 2022.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Mingkun Wang, Xinge Zhu, Changqian Yu, Wei Li, Yuexin Ma, Ruochun Jin, Xiaoguang Ren, Dongchun Ren, Mingxu Wang, and Wenjing Yang. Ganet: Goal area network for motion forecasting. *arXiv preprint arXiv:2209.09723*, 2022.

Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11318–11327, 2021.

Maosheng Ye, Jiamiao Xu, Xunnong Xu, Tongyi Cao, and Qifeng Chen. Dcms: Motion forecasting with dual consistency and multi-pseudo-target supervision. *arXiv preprint arXiv:2204.05859*, 2022.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9813–9823, 2021.

Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. Dsdnet: Deep structured self-driving network. In *European conference on computer vision*, pp. 156–172. Springer, 2020.

Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 532–539. IEEE, 2021.

Lingyao Zhang, Po-Hsun Su, Jerrick Hoang, Galen Clark Haynes, and Micol Marchetti-Bowick. Map-adaptive goal-based trajectory prediction. In *Conference on Robot Learning*, pp. 1371–1383. PMLR, 2021.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. *arXiv preprint arXiv:2008.08294*, 2020.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pp. 895–904. PMLR, 2021.

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*, pp. 8823–8833, 2022.