HPNet: Dynamic Trajectory Forecasting with Historical Prediction Attention

Xiaolong Tang[1,2] Meina Kan[1,2] Shiguang Shan[1,2] Zhilong Ji[3] Jinfeng Bai[3] Xilin CHEN[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3] Tomorrow Advancing Life
{tangxiaolong22s, kanmeina, sgshan, xlchen}@ict.ac.cn

Abstract

Predicting the trajectories of road agents is essential for autonomous driving systems. The recent mainstream methods follow a static paradigm, which predicts the future trajectory by using a fixed duration of historical frames. These methods make the predictions independently even at adjacent time steps, which leads to potential instability and temporal inconsistency. As successive time steps have largely overlapping historical frames, their forecasting should have intrinsic correlation, such as overlapping predicted trajectories should be consistent, or be different but share the same motion goal depending on the road situation. Motivated by this, in this work, we introduce HPNet, a novel dynamic trajectory forecasting method. Aiming for stable and accurate trajectory forecasting, our method leverages not only historical frames including maps and agent states, but also historical predictions. Specifically, we newly design a Historical Prediction Attention module to automatically encode the dynamic relationship between successive predictions. Besides, it also extends the attention range beyond the currently visible window benefitting from the use of historical predictions. The proposed Historical Prediction Attention together with the Agent Attention and Mode Attention is further formulated as the Triple Factorized Attention module, serving as the core design of HPNet. Experiments on the Argoverse and INTERACTION datasets show that HPNet achieves state-of-the-art performance, and generates accurate and stable future trajectories. Our code are available at https://github.com/XiaolongTang23/HPNet.

1. Introduction

Accurate and reliable trajectory prediction of road agents such as cars and pedestrians is critical to the decision-making and safety of autonomous driving systems. However, trajectory prediction is extremely challenging. On the one hand, an agent's motion is influenced not only by road geometry and rules but also by surrounding agents. On the other hand, the agent's intentions are unknown, leading to high levels of future uncertainty.

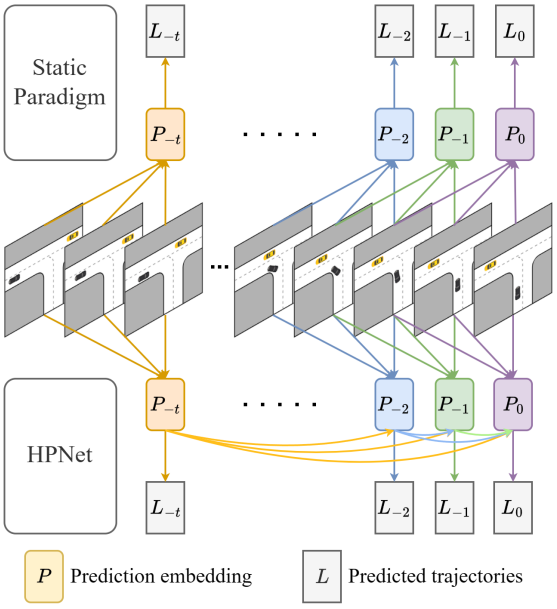


Figure 1. The difference between previous methods and ours. Previous methods (upper) treat trajectory prediction as a static task, predicting future trajectories based on a fixed-length sequence of historical frames. They independently forecast trajectories even at adjacent timesteps, despite the considerable overlap in input data. In contrast, HPNet (lower) views trajectory prediction as a dynamic task. It not only leverages historical frames but also historical prediction embeddings to forecast trajectories.

Recently, researches such as Macformer [10], HiVT [44], and Multipath++ [32] have achieved notable results by employing intricately designed network architectures to seamlessly fuse heterogeneous information including agent history, agent-agent interactions, and agent-map interactions. Wayformer [25] further explores the unified architecture for fusing the heterogeneous information. In addition, to account for future uncertainties, recent work [3, 17, 20, 24–27, 32, 34, 36, 41, 43, 45] have shifted to-

wards generating multi-modal future trajectories rather than single trajectory, since road agents may make varying decisions even in identical scenarios. Anchor-based approaches [3, 17, 34, 41, 43] utilize multiple candidate goals or predefined paths as anchors to indicate various potential future, thereby facilitating the generation of multi-modal trajectories. More recently, models [24, 25, 32, 36, 45] adopt learnable queries to generate multi-modal predictions, achieving promising results.

While the existing methods have achieved great advancements in prediction accuracy, they mostly treat trajectory prediction as a static task by using a fixed number of historical frames to predict future trajectories. As shown in Fig. 1, successive predictions are inherently independent, although their input overlaps considerably. *This static paradigm of trajectory prediction may lead to instability and temporal inconsistency in successive predictions, which is not conducive to the autonomous driving system making safe and reliable decisions.* Aiming for more stable prediction, DCMS [38] proposes to model trajectory prediction as a dynamic problem. It explicitly considers the correlation between successive predictions and imposes a temporal consistency constraint that requires the overlapping parts of predicted trajectories at adjacent time steps to be identical. Moreover, QCNet [45] introduces a query-centric encoding paradigm to encode location-dependent features and location-independent features separately, thereby circumventing redundant encoding in successive predictions. These preliminary studies show the effectiveness and rationality of modeling trajectory prediction as a dynamic task.

These works [38, 45] inspire us to think that the intrinsic relationship between successive predictions should be more general, not just consistent. For example, the overlapping portions may remain consistent as that in DCMS [38] or change slightly. Even more, when an agent is navigating through a congested multi-way intersection, the successive predictions may be quite different but still share the same motion goal as shown in Fig. 4 (b). So, in this work, we present a novel dynamic trajectory prediction method, HPNet. It models the dynamic relationship between successive predictions as the process of History Prediction Attention. Specifically, HPNet consists of three components: Spatio-Temporal Context Encoding, Triple Factorized Attention, and Multimodal Output. Initially, the mode queries aggregate spatio-temporal context to form preliminary prediction embeddings. Subsequently, Triple Factorized Attention, comprising Agent Attention, Historical Prediction Attention, and Mode Attention, models the interactions between agents, predictions, and modes, respectively, to obtain more informative prediction embeddings. Finally, the embeddings are decoded as multimodal future trajectories in the last module.

Our method has two clear advantages: First, our model establishes a general relationship between successive predictions, using the historical predictions as references to improve stability and increase accuracy. Second, in online inference, existing static attention-based methods are constrained within the fixed visible historical range, due to limited dataset size or computational resources. Instead, our approach can achieve a larger visible range (i.e. longer attention) without increasing computational overhead, which is beneficial for better accuracy in practical applications.

2. Related Work

Attention Mechanism. Transformer [33] has achieved notable success in fields such as natural language processing and computer vision. This is largely attributed to the attention mechanism, which considers the whole context and focuses on the important parts of the input data. Recently, many methods [6, 7, 15, 16, 19, 24–26, 30, 36, 39, 42, 44–46] employ attention to process agent history sequence or model agent-agent and agent-lane interactions, and have achieved great success in trajectory prediction. The input for the trajectory prediction task usually encompasses data across temporal and spatial dimensions. Instead of flattening the input together into one joint self-attention [6, 36, 39], applying attention to each axis [16, 24–26, 44, 45] may be more suitable for trajectory prediction task, which leads to better semantic consistency as well as lower computational complexity. In our work, the key module Triple Factorized Attention also follows the latter attention approach, consisting of Historical Prediction Attention, Agent Attention, and Mode Attention to model the interaction of agents in three different dimensions.

Multimodal Output. To model future uncertainty distributions, probabilistic methods [2, 9, 18, 28, 29, 39] firstly propose to use generative models (*e.g.*, GAN, VAE) to obtain multimodal outputs through multiple samplings. However, in these methods, the number of sampling iterations required to produce reliable results is indeterminate, making them unreliable for the later decision process of autonomous driving. Afterwards, deterministic methods [3, 17, 20, 22–27, 31, 32, 34–38, 43–45] propose to eliminate the need of multiple sampling and directly output multimodal future trajectories in a single shot, yielding more accurate trajectory prediction. Among them, Anchor-based approaches usually employ two types of anchors, candidate targets [17, 20, 34, 43] and predefined paths [3, 27, 41]. These anchors indicate a variety of potential future trajectories, thus effectively improving the prediction accuracy of multimodal trajectories. However, the performance of these methods heavily depends on the quality of the anchors, and sometimes bad anchors can lead to irreparably bad results. So, inspired by DETR [1], several methods propose to use learnable queries [24, 25, 31, 32, 36, 45] rather than anchors to facilitate multimodal output. In these methods, each

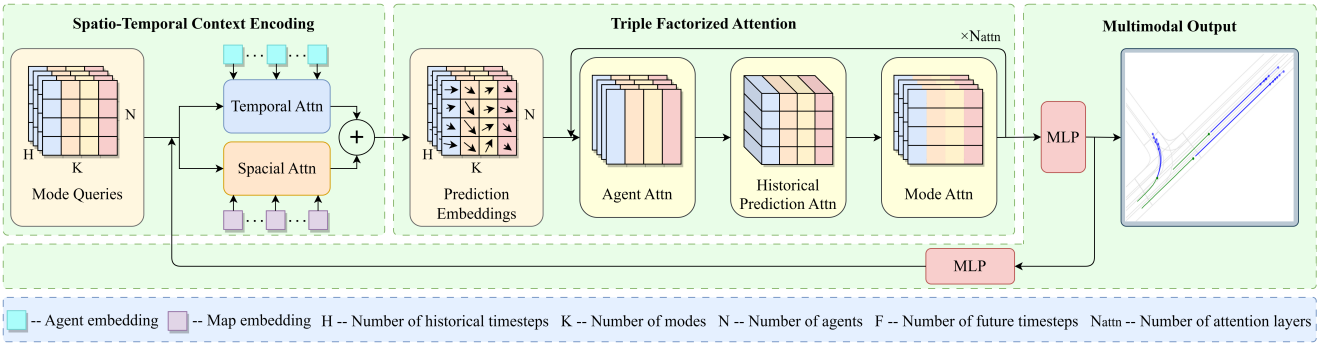


Figure 2. An overview of HPNet. The proposed HPNet encompasses three components: Spatio-Temporal Context Encoding, Triple Factorized Attention, and Multimodal Output. Firstly, it combines agent and lane features with mode queries to create initial prediction embeddings. Subsequently, Triple Factorized Attention — comprising Agent Attention, our proposed Historical Prediction Attention, and Mode Attention — refine these prediction embeddings. Finally, the prediction embeddings are decoded by an MLP to obtain the predicted trajectories. The predicted trajectories are fed into this pipeline again to enhance the precision of predictions.

mode query can adaptively generate a potential trajectory for every sample based on its context, which is more flexible than pre-defined fixed anchors, leading to promising performance in trajectory prediction. Lately, there appear some approaches [7, 31, 36, 45, 46] to combine query-based and anchor-based methods. For example, ProphNet [36] generates specific anchors based on the sample and feeds the anchor information into learnable mode queries to help produce multimodal trajectories. QCNet [45] adaptively generates trajectory proposals through anchor-free mode queries and then refines these proposals based on the context using anchor-based mode queries.

3. Method

Trajectory forecasting aims to predict the future trajectories of any agent given its historical status. Specifically, given a fixed length sequence of history status frames $\{f_{-T+1}, f_{-T+2}, ..., f_0\}$, the goal is to predict K different modal trajectories for N agents as below:

$$L_0 = \{L_{0,n,k}\}_{n\in[1,N],k\in[1,K]}, \tag{1}$$

where $f_t = \{a_t^{1\sim N}, \mathcal{M}\}$, $a_t^{1\sim N}$ represents the features of all agents in the scene at time t, and $\mathcal{M}$ denotes the high-definition (HD) map including M lane segments. And each trajectory contains future locations for the next F time steps:

$$L_{0,n,k} = \{l_{1,n,k}, l_{2,n,k}, ..., l_{F,n,k}\}, \tag{2}$$

where $l_{i,n,k} \in \mathbb{R}^2$ represents the predicted position at time step i of mode k for agent n. Simultaneously, a probability score is usually obtained for each predicted trajectory to indicate its likelihood of being the path that the agent will actually follow.

An overview of our proposed HPNet is illustrated in Fig. 2. As shown, our model consists of three parts: Spatio-Temporal Context Encoding, Triple Factorized Attention, and Multimodal Output. Firstly, the spatio-temporal features of agents and lanes are aggregated with learnable mode queries to generate prediction embeddings that can preliminarily predict future trajectories. Then, Triple Factorized Attention including Agent Attention, Historical Prediction Attention, and Mode Attention are conducted to refine the prediction embeddings. Among them, Agent Attention models interactions between agents, Mode Attention models interactions across different modes (*i.e.*, different predicted paths), and Historical Prediction Attention is a novel module we propose to dynamically model the intrinsic correlation between current and historical predictions. Finally, The prediction embeddings are decoded by a MLP to obtain the predicted trajectories, which are fed into the whole pipeline again to enhance the precision of forecasts.

3.1. Spatio-Temporal Context Encoding

HPNet is based on Graph Neural Networks (GNNs) and adopts relative spatio-temporal position encoding [19, 42, 45, 46]. It encodes the location-independent features of agents and maps into node embeddings while encoding the relative spatio-temporal positions into edge embeddings.

Encoding Agent Features. The agent features include spatial position, motion state, and semantic attributes of each agent at each time step. Each agent at each time step is adopted as a node in the graph, and its features represented as $a_t^n = \{p_x^{t,n}, p_y^{t,n}, \theta^{t,n}, v_x^{t,n}, v_y^{t,n}, c_a^{t,n}\}$, where $(p_x^{t,n}, p_y^{t,n})$ is the location, $\theta^{t,n}$ is the orientation, $(v_x^{t,n}, v_y^{t,n})$ is the velocity, and $c_a^{t,n}$ is the attribute. For each agent at each time step, we take its location as the origin of a local polar coordinate system and its orientation as the positive direction. In this reference frame, the velocity $(v_x^{t,n}, v_y^{t,n})$ is represented as $(v^{t,n}, \varphi^{t,n})$ where $v^{t,n}$ is the velocity magnitude and $\varphi^{t,n}$ is the direction of velocity. A two-layer MLP is adopted to encode the location-independent features as agent embed-

dings $E_a^{t,n} = \mathrm{MLP}(v^{t,n}, \varphi^{t,n}, c_a^{t,n})$, where $E_a^{t,n} \in \mathbb{R}^D$, D is the encoding dimension.

Encoding Map Features. The map features include spatial position, length, and semantic attributes of each lane segment. The lane segments are adopted as nodes in the graph, each of which comprises a set of centerlines along with some attributes. The location and orientation of the midpoint of centerlines are used to represent the position and orientation of each lane segment. The lengths of the lane segments l_m and their attributes c_m serve as node features, which are encoded through a two-layer MLP as map embeddings $E_m = \mathrm{MLP}(l_m, c_m)$, where $E_m \in \mathbb{R}^{M \times D}$. Similar to LaneGCN [23], to capture the topological structure of the map, the lane nodes are connected based on adjacent, predecessor, and successor relationships. Interaction between lane nodes is then accomplished via self-attention.

Encoding Relative Spatio-Temporal Position. The relative spatio-temporal positions between nodes are used as the features for the edges. All nodes in the graph are encoded using features in their local polar coordinates, so the edges represent the transformation relationships between different local polar coordinates. The edge features can be denoted as $\{d_e, \phi_e, \psi_e, \delta_e\}$, where d_e represents the distance from the source node to the target node, ϕ_e indicates the orientation of the edge in the target node's reference frame, ψ_e denotes the relative orientation between the source and target nodes, and δ_e signifies the time difference between them. Similarly, a two-layer MLP is adopted to encode these features into edge embeddings $E_e = \mathrm{MLP}(d_e, \phi_e, \psi_e, \delta_e)$, where $E_e \in \mathbb{R}^{Y \times D}$, Y is the number of edges.

Spatio-Temporal Attention. Spatio-temporal attention comprises two parallel cross-attention modules. Temporal Attention aggregates the historical embeddings of agents, while Spatial Attention models agent-lane interaction. We assign the same learnable mode queries to each agent at each time step, denoted as $\{q_{t,n,k}\}_{t \in [1-T,0], n \in [1,N], k \in [1,K]}$. Each mode query, as a node in the graph, the spatio-temporal position of which is identical to the corresponding agent. For each mode query node, Spatial Attention with lane nodes is performed within a certain spatial radius R_1 and Temporal Attention with agent nodes within a specified time span I_1, respectively. Edges participate in this process via concatenating with the source node [19, 42, 44–46]:

$$q_{t,n,k}^S = \mathrm{MHA}(q_{t,n,k}, [E_m, E_e], [E_m, E_e]), \tag{3}$$

$$q_{t,n,k}^T = \mathrm{MHA}(q_{t,n,k}, [E_a^{t-I_1 \sim t,n}, E_e], [E_a^{t-I_1 \sim t,n}, E_e]), \tag{4}$$

where $\mathrm{MHA}(a, b, c)$ denotes the multi-head attention with a as query, b as key, and c as value.The results of the two cross-attention modules are then summed to generate the prediction embeddings:

$$P_{t,n,k} = q_{t,n,k}^T + q_{t,n,k}^S. \tag{5}$$

Subsequently, the prediction embeddings generated from Eq. (5) are passed through Triple Factorized Attention. Triple Factorized Attention comprises Agent Attention, Historical Prediction Attention, and Mode Attention, allowing each prediction embedding to directly or indirectly 'talk' to the embeddings of different agents, different time steps, and different modes. In Sec. 3.2, Sec. 3.3, and Sec. 3.4, we introduce Agent Attention, Historical Prediction Attention, and Mode Attention, respectively.

3.2. Agent Attention

In the Agent Attention module, self-attention is accomplished across agents for each mode and each time step on these prediction embeddings:

$$P_{t,n,k}^A = \mathrm{MHA}(P_{t,n,k}, [P_{t,n',k}, E_e], [P_{t,n',k}, E_e]), \tag{6}$$

where n' denotes all agents within a certain radius R_2 of the n-th agent under the same mode and time step. On the one hand, agent attention models the interactions among agents within their respective spatio-temporal contexts. On the other hand, it can also be conceived as the interaction between the future trajectories of different agents, thereby mitigating potential collisions.

3.3. Historical Prediction Attention

After aggregating historical agent states, agent-lane interactions, and agent-agent interactions, previous methods typically begin predicting future trajectories. However, we observe that the current and historical predictions are usually correlated, while most existing methods neglect this. For example, when an agent is moving steadily on a straight path, the overlapping segments of successive predictions should be almost the same or vary minimally. When an agent traverses a busy multi-lane crossroads, the successive predictions may be quite different but still share the same motion goal as shown in Fig. 4 (b). Our experiments show that this kind of correlation between successive predictions is critical not only for the stability of the predictions but also for accuracy.

Consequently, to further improve the stability and accuracy of trajectory prediction, we design this novel Historical Prediction Attention. It incorporates historical predictions to inform the current forecast, modeling the dynamic correlation between successive predictions by the attention mechanism. Specifically, each prediction embedding performs self-attention with historical prediction embeddings within the temporal span I_2 for each agent and each mode:

$$P_{t,n,k}^{HP} = \mathrm{MHA}(P_{t,n,k}^A, [P_{t-I_2 \sim t,n,k}^A, E_e], [P_{t-I_2 \sim t,n,k}^A, E_e]). \tag{7}$$

Here, the prediction embeddings rather than final historical prediction trajectories are used to model the dynamic

relationship, because the latter will change the training process from parallel execution to serial execution, greatly increasing the time required for training.

Besides improving the accuracy and stability of prediction, this attention in Eq. (7) can also absorb longer history information beyond the currently visible window, *i.e.*, extends the attention range. To be specific, without Historical Prediction Attention, the observation window of P_t^{HP} is confined to the interval $[t - I_1, t]$ as it only uses the I_1 previous frames, limited to the time span I_1. In contrast, if $I_2 = I_1$, the observation window of Historical Prediction Attention is twice as long, *i.e.*, $[t - I_1 - I_1, t]$. In detail, the current prediction uses the historical I_1 prediction embedding, and thus the observation window is $[t - I_1, t]$ w.r.t. the prediction embedding. However, the prediction embedding at the farthest timestep $t - I_1$ actually already absorbs attention information from previous $[t - I_1 - I_1, t - I_1]$ frames. Therefore, the actual observation window of Historical Prediction Attention is the sum of the two intervals, *i.e.*, $[t - I_1 - I_1, t]$. In general case that $I_2 \neq I_1$, the actual observation window of Historical Prediction Attention is $[t - I_2 - I_1, t]$, which is also longer than that of most existing methods $[t - I_1, t]$. *The longer attention range of Historical Prediction Attention can provide more beneficial information for better trajectory prediction without additional computational cost.*

3.4. Mode Attention

After the historical prediction attention, self-attention is then applied across different modes of prediction embeddings for each agent and each time step, modeling mode-mode interactions between different future trajectories to enhance multimodal outputs:

$$P_{t,n,k}^{M} = \mathrm{MHA}(P_{t,n,k}^{HP}, [P_{t,n,1\sim K}^{HP}, E_e], [P_{t,n,1\sim K}^{HP}, E_e]). \tag{8}$$

After Eq. (8), the Triple Factorized Attention is accomplished, inducing enhanced prediction embedding. The Triple Factorized Attention is repeated $N_{attn} = 2$ times so that all prediction embeddings can fully interact with each other for more accurate prediction.

3.5. Multimodal Output

Finally, all the prediction embeddings are decoded through a two-layer MLP to obtain multiple future locations:

$$L_{t,n,k}^{1} = \mathrm{MLP}(P_{t,n,k}^{M}), \tag{9}$$

where $L_{t,n,k}^{1} \in \mathbb{R}^{F\times 2}$. To further enhance the output trajectories, following QCNet [45], $L_{t,n,k}^{1}$ is taken as the input of the whole pipeline to further refine the predicted trajectories. In detail, $L_{t,n,k}^{1}$ is taken as trajectory proposals and encoded into mode queries by another two-layer MLP. These encoded mode queries replace the learnable mode queries as inputs of Spatio-Temporal Attention to re-aggregate the spatio-temporal context and perform Triple Factorized Attention again. This refinement process produces a trajectory refinement $\Delta L_{t,n,k}$ and probability scores $\hat{\pi}_{t,n,k}$.

Then, the final predicted trajectory is obtained by summing the trajectory proposal and the trajectory refinement:

$$L_{t,n,k}^{2} = L_{t,n,k}^{1} + \Delta L_{t,n,k}. \tag{10}$$

3.6. Training Objective

Following the existing works [8, 11, 24, 25, 36, 44, 45], we adopt the winner-takes-all [21] strategy to optimize our model. For marginal prediction, the $k_{t,n}$-th mode to be optimized is determined based on the minimum endpoint displacement between the predicted trajectory $\{L_{t,n,k}^{1}\}_{k\in[1,K]}$ and the ground truth $G_{t,n} = \{g_{t+1,n}, g_{t+2,n}, ..., g_{t+F,n}\}$:

$$k_{t,n} = \underset{k\in[1,K]}{\arg\min}(l_{t+F,n,k}^{1} - g_{t+F,n}). \tag{11}$$

Then, the regression loss function contains two Huber losses for the trajectory proposals and the refined final trajectories, respectively:

$$\mathcal{L}_{reg1}^{t,n} = \mathcal{L}_{Huber}(L_{t,n,k_{t,n}}^{1}, G_{t,n}), \tag{12}$$

$$\mathcal{L}_{reg2}^{t,n} = \mathcal{L}_{Huber}(L_{t,n,k_{t,n}}^{2}, G_{t,n}). \tag{13}$$

Besides, the probability scores are optimized by using the cross-entropy loss function:

$$\mathcal{L}_{cls}^{t,n} = \mathcal{L}_{CE}(\{\hat{\pi}_{t,n,k}\}_{k\in[1,K]}, k_{t,n}). \tag{14}$$

Overall, the total loss function of the whole model is formulated as follows:

$$\mathcal{L} = \frac{1}{TN}\sum_{t=-T+1}^{0}\sum_{n=1}^{N}(L_{reg1}^{t,n} + L_{reg2}^{t,n} + L_{cls}^{t,n}). \tag{15}$$

For joint prediction, we treat the prediction of all agents in the same mode as a predicted future and the joint endpoint displacement determines the mode to be optimized. Please refer to the supplementary material for a detailed explanation of the training objective for joint prediction.

4. Experiments

4.1. Experimental Setup

Dataset. We conduct the experiments on the Argoverse [4] and INTERACTION [40] datasets. Both are based on real-world driving scenarios, providing high-definition maps and detailed motion information, sampled at a frequency of 10Hz. On the Argoverse dataset, we assessed HPNet's capability for marginal trajectory prediction. Conversely, on

	Method	**b-minFDE↓**	minFDE↓	MR↓	minADE↓
Single model	LaneGCN [23]	2.0539	1.3622	0.1620	0.8703
	mmTransformer [24]	2.0328	1.3383	0.1540	0.8436
	DenseTNT [17]	1.9759	1.2815	0.1258	0.8817
	THOMAS [15]	1.9736	1.4388	0.1038	0.9423
	TPCN [37]	1.9286	1.2442	0.1333	0.8153
	SceneTransformer [26]	1.8868	1.2321	0.1255	0.8026
	HiVT [44]	1.8422	1.1693	0.1267	<u>0.7735</u>
	GANet [35]	<u>1.7899</u>	<u>1.1605</u>	<u>0.1179</u>	0.8060
	HPNet(single model)	**1.7375**	**1.0986**	**0.1067**	**0.7612**
Ensembled model	HOME+GOHOME [13, 14]	1.8601	1.2919	**0.0846**	0.8904
	Multipath++ [32]	1.7932	1.2144	0.1324	0.7897
	Macformer [10]	1.7667	1.2141	0.1272	0.8121
	DCMS [38]	1.7564	1.1350	0.1094	0.7659
	HeteroGCN [12]	1.7512	1.1602	0.1168	0.7890
	Wayformer [25]	1.7408	1.1616	0.1186	0.7676
	ProphNet [36]	1.6942	1.1337	0.1101	0.7623
	QCNet [45]	<u>1.6934</u>	**1.0666**	<u>0.1056</u>	**0.7340**
	HPNet(ensembled model)	**1.6768**	<u>1.0856</u>	0.1075	<u>0.7478</u>

Table 1. Comparison of HPNet with the state-of-the-art methods on the Argoverse test set, where b-minFDE is the official ranking metric. For each metric, the best result is in **bold** and the second best result is <u>underlined</u>.

Method	minJointFDE↓	minJointADE↓
AutoBot [16]	1.0148	0.3123
THOMAS [15]	0.9679	0.4164
Trai-MAE [5]	0.9660	0.3066
HDGT [38]	0.9580	0.3030
FJMP [45]	<u>0.9218</u>	<u>0.2752</u>
HPNet(single model)	**0.8231**	**0.2548**

Table 2. Comparison of HPNet with the state-of-the-art methods on the INTERACTION test set. For each metric, the best result is in **bold** and the second best result is <u>underlined</u>.

the INTERACTION dataset, renowned for its complex driving scenarios and detailed multi-agent interactions, we examined HPNet's effectiveness in joint prediction.

Metrics. For our evaluation, we employed official trajectory forecasting metrics, encompassing minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), Miss Rate (MR), and Brier-minimum Final Displacement Error (b-minFDE) for Argoverse. MinADE measures the average ℓ_2-norm distance across predicted and actual trajectory points, while minFDE examines the ℓ_2-norm distance at the trajectory's endpoint. MR assesses instances where predictions stray more than 2.0 meters from the actual endpoint, gauging model reliability. Lastly, brier-minFDE extends minFDE by incorporating the probability part $(1 - \hat{\pi})^2$, providing insights into the model's confidence in its best prediction. For the INTERACTION dataset, we employed minJointADE and minJointFDE metrics to assess joint trajectory prediction performance. MinJointADE evaluates the average ℓ_2-norm distance across all agents' predicted and actual trajectories, while minJointFDE focuses on the ℓ_2-norm distance at the final time step for all agents. To explore the model's capability in capturing multimodal outputs, we set $K = 6$ for both marginal prediction and joint prediction.

4.2. Comparison with State-of-the-art

Results on Argoverse. The marginal trajectory prediction results on Argoverse are reported in Tab. 1. Our HPNet achieves the best results on all indicators among all single models. Compared to GANet in second place, the improvement is up to 0.052 in b-minFDE, 0.062 in minFDE, and 0.045 in minADE. Moreover, following [32, 36, 38, 45], HPNet is further compared in the setting of model ensembling. As can be seen, our HPNet also performs best in the official ranking metric. Compared to the single model, ensembled HPNet only yields a reduction of 0.013 in minFDE. This is mainly because the HPNet's predictions are more stable, and thus the improvement from ensembling is smaller than other methods. Overall, our HPNet achieves state-of-the-art performance, verifying the superiority of our HPNet.

Results on INTERACTION. Tab. 2 shows the results of our method on the INTERACTION multi-agent track. We

Spatio-Temporal Attention		Triple Factorized Attention			Metrics			
Spatial	Temporal	Agent	Historical Prediction	Mode	b-minFDE↓	minFDE↓	MR↓	minADE↓
✓	✓				1.832	1.203	0.126	0.771
✓	✓		✓	✓	1.711	1.084	0.102	0.722
✓	✓	✓		✓	1.527	0.909	0.075	0.661
✓	✓	✓	✓		1.531	0.894	0.073	0.645
✓	✓	✓	✓	✓	1.506	0.871	0.069	0.638

Table 3. Ablation study of Triple Factorized Attention. Experiments are performed on the Argoverse validation set.

achieved state-of-the-art performance on this benchmark, achieving substantial gains over the second-ranked FJMP, with improvements of 0.099 in minJointFDE and 0.020 in minJointADE. This indicates that our HPNet can be used simply and effectively for joint trajectory prediction.

4.3. Ablation Study

We first conduct ablation studies on Triple Factorized Attention to analyze the importance of Agent Attention, Historical Prediction Attention, and Mode Attention in our proposed HPNet. Then we explore the impact of Historical Prediction Attention on prediction accuracy and stability. Lastly, we examine the influence of Historical Prediction Attention on the reaction timeliness.

Component Study of Triple Factorized Attention. As shown in Tab. 3, the model with all components achieves 1.506 in terms of b-minFDE, which is the best result on the validation set. If removing Triple Factorized Attention, the performance in terms of b-minFDE drops by 0.326, implying the importance of the Triple Factorized Attention module in the overall model architecture. If removing the Agent Attention, Historical Prediction Attention, and Mode Attention, the performance in terms of b-minFDE drops by 0.205, 0.021, and 0.025, respectively. This indicates the effectiveness of all three attention modules, among which the Agent Attention between agents and surrounding agents has the most important influence and is indispensable for prediction. Besides, our proposed Historical Prediction Attention also plays an important role with obvious improvements on four metrics, which clearly illustrates the necessity of considering the relationship between successive predictions.

The Impact of Historical Prediction Attention on Accuracy and Stability. Our proposed Historical Prediction Attention is expected to improve the accuracy and stability of trajectory prediction by considering the relationship between current and historical predictions. To investigate whether this expectation is achieved, we conduct comparative experiments on two models: the HPNet and its baseline model without Historical Prediction Attention. Predictions are made across 10 time steps, ranging from 20 to 30, with each prediction utilizing a visible history flame window of

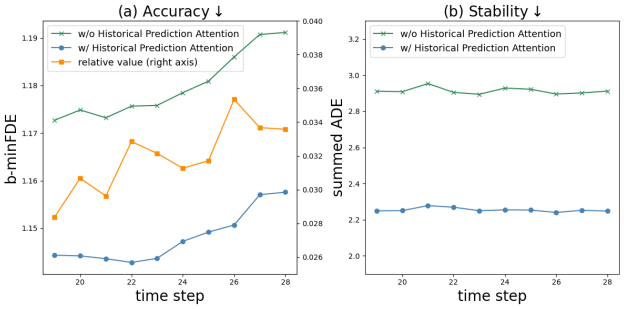
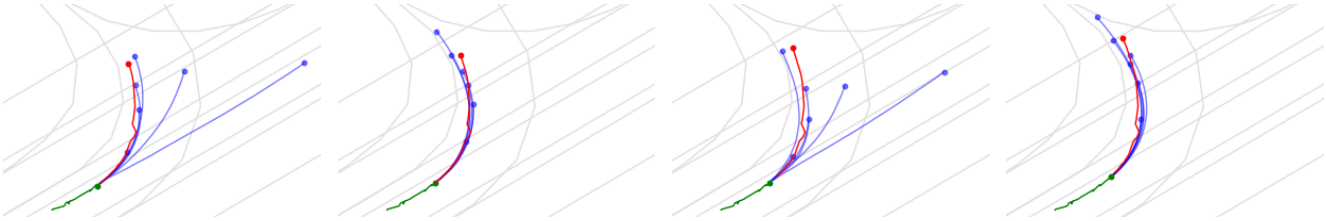


Figure 3. Comparison of prediction Accuracy (b-minFDE↓) and Stability (summed ADE↓) of our HPNet and its baseline without Historical Prediction Attention on the Argoverse validation set.

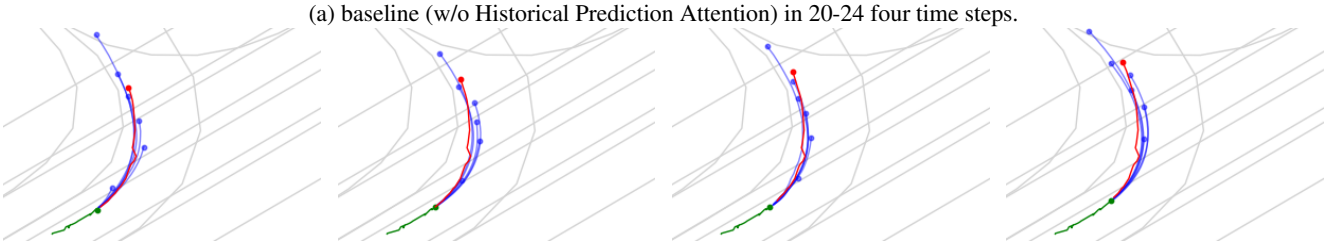
20 time steps and a historical prediction window of equal length. The accuracy of the predictions is quantified by using the b-minFDE metric. The stability is assessed by the summed ADE of the overlapping segments of matched trajectory pairs at current and previous time steps, where the matched trajectory pairs are obtained via the Hungarian matching algorithm.

As shown in Fig. 3 (a), for all predicted time steps, the performance of HPNet in terms of b-minFDE is better than that of the baseline. This superior performance indicates that Historical Prediction Attention indeed improves the accuracy of trajectory prediction. Besides, it is also observed that the accuracy of both our HPNet and the baseline declines along the temporal axis, and this is mainly because of the appearance of new agents in later time steps that are not present in the first 20 frames. Even so, the relative improvement of our HPNet over the baseline (*i.e.*, the orange dashed line) becomes larger over time. When the time step becomes longer, a big difference is that the actual visible historical window of our HPNet is beyond 20 time steps as analyzed in Sec. 3.3, while that of the baseline is always fixed to 20 time steps. Therefore, this significant relative improvement over time clearly verifies the benefit from the longer attention window of Historical Prediction Attention.

As shown in Fig. 3 (b), for all predicted time steps, the



(a) baseline (w/o Historical Prediction Attention) in 20-24 four time steps.



(b) HPNet (w/ Historical Prediction Attention) in 20-24 four time steps.

Figure 4. Qualitative results on the Argoverse validation set. Baseline (a) alternately forecasts one motion goal (*i.e.*, turn left) and two motion goals (*i.e.*, turn left and go straight). In contrast, HPNet (b) consistently and reliably predicts the same motion goal (*i.e.*, turn left). The lanes, historical trajectory, ground truth trajectory, and six predicted trajectories are indicated in grey, green, red, and blue, respectively.

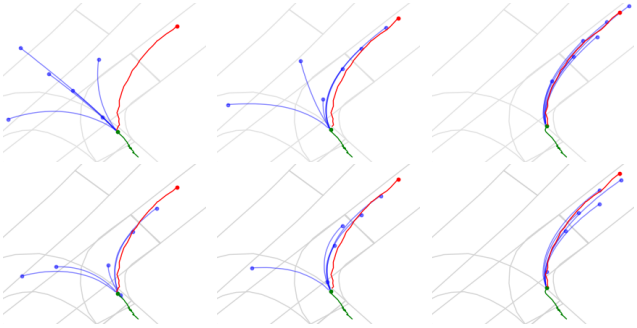


Figure 5. Predictions of HPNet (lower) and baseline (upper).

summed ADE of HPNet between successive time steps is about 2.25, while the summed ADE of the baseline is about 2.90 which is much larger. This indicates that Historical Prediction Attention indeed makes predicted trajectories much more stable. We show an example in Fig. 4 to make the comparison more intuitive. As shown, the agent chooses to turn left at the intersection. However, the baseline may be due to the agent's pause in the middle moment, alternately predicting one motion goal (*i.e.*, turn left) and two motion goals (*i.e.*, turn left and go straight). In contrast, HPNet consistently and reliably predicted the same motion goal (*i.e.*, turn left). At the same time, unlike DCMS [38], the successive predictions of HPNet only share the same motion goal in complex road conditions, without forced overlapping waypoints. These stable prediction results enable subsequent modules to produce stable and time-consistent safe driving decisions. More qualitative results can be found in the supplementary material.

The Influence of Historical Prediction Attention on Reaction Timeliness. While Historical Prediction Attention enhances forecasting stability by leveraging historical predictions, does it hurt the reaction timeliness? Our answer is no. This is mainly benefited from the attention mechanism. When an abrupt change occurs (*e.g.*, a sudden right turn in Fig. 5), the similarity between the current and historical prediction embeddings decreases, leading to reduced weights for historical predictions. Consequently, the impact of past predictions on the current moment dynamically diminishes. Fig. 5 illustrates a qualitative example where an agent at an intersection is observed at three consecutive moments. When the agent shows no specific intention, HPNet stably and accurately forecasts the possibilities of turning left or right, outperforming the baseline. At the sudden right turn (in the final moment), HPNet quickly adjusts to predict right turns only, with no delay compared to the baseline.

5. Conclusion

In this paper, we propose a novel dynamic trajectory prediction method, HPNet. A Historical Prediction Attention module is designed to model the dynamic relationship between successive predictions. It employs historical prediction embeddings to guide current forecast, making the predicted trajectories more accurate and stable. Experiments on the Argoverse and INTERACTION datasets demonstrate that our proposed HPNet achieves state-of-the-art performance, and also proves that Historical Prediction Attention can effectively improve accuracy and stability.

Acknowledgment. This work was partially supported by the National Key R&D Program of China (No. 2020AAA0104500) and the Natural Science Foundation of China (Nos. U2336213 and 62122074).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[2] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[3] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning (CoRL)*, 2020. 1, 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[5] Hao Chen, Jiaze Wang, Kun Shao, Furui Liu, Jianye Hao, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. Traj-mae: Masked autoencoders for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6

[6] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[7] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[8] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 5

[9] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning (CoRL)*, 2022. 2

[10] Chen Feng, Hangning Zhou, Huadong Lin, Zhigang Zhang, Ziyao Xu, Chi Zhang, Boyu Zhou, and Shaojie Shen. Macformer: Map-agent coupled transformer for real-time and robust trajectory prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2023. 1, 6

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[12] Xing Gao, Xiaogang Jia, Yikang Li, and Hongkai Xiong. Dynamic scenario representation learning for motion forecasting with heterogeneous graph convolutional recurrent networks. *IEEE Robotics and Automation Letters (RA-L)*, 2023. 6

[13] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Home: Heatmap output for future motion estimation. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021. 6

[14] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 6

[15] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Thomas: Trajectory heatmap output with learned multi-agent sampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2, 6

[16] 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. *arXiv preprint arXiv:2104.00563*, 2021. 2, 6

[17] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6

[18] 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 (CVPR)*, 2018. 2

[19] Xiaosong Jia, Penghao Wu, Li Chen, Yu Liu, Hongyang Li, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2, 3, 4

[20] ByeoungDo Kim, Seong Hyeon Park, Seokhwan Lee, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, Jeong Soo Kim, and Jun Won Choi. Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[21] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 5

[22] Jinmin Li, Tao Dai, Mingyan Zhu, Bin Chen, Zhi Wang, and Shu-Tao Xia. Fsr: A general frequency-oriented framework to accelerate image super-resolution networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2

[23] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 4, 6

[24] 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 (CVPR)*, 2021. 1, 2, 5, 6

[25] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 5, 6

[26] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2, 6

[27] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[28] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[30] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[31] 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 (NIPS)*, 2022. 2, 3

[32] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 1, 2, 6

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[34] Jingke Wang, Tengju Ye, Ziqing Gu, and Junbo Chen. Ltp: Lane-based trajectory prediction for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[35] 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. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 6

[36] Xishun Wang, Tong Su, Fang Da, and Xiaodong Yang. Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 6

[37] 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 (CVPR)*, 2021. 6

[38] Maosheng Ye, Jiamiao Xu, Xunnong Xu, Tengfei Wang, Tongyi Cao, and Qifeng Chen. Dcms: Motion forecasting with dual consistency and multi-pseudo-target supervision. *arXiv preprint arXiv:2204.05859*, 2022. 2, 6, 8

[39] 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 (ICCV)*, 2021. 2

[40] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 5

[41] 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 (CoRL)*, 2021. 1, 2

[42] Zhejun Zhang, Alexander Liniger, Christos Sakaridis, Fisher Yu, and Luc V Gool. Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. 2024. 2, 3, 4

[43] 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 (CoRL)*, 2021. 1, 2

[44] 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 (CVPR)*, 2022. 1, 2, 4, 5, 6

[45] 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 (CVPR)*, 2023. 1, 2, 3, 5, 6

[46] Yiyao Zhu, Di Luan, and Shaojie Shen. Biff: Bi-level future fusion with polyline-based coordinate for interactive trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4