DMP: Destination-driven Motion Prediction with Prior Fusion

Ting Yu Lingxin Jiang Wei Li
Inceptio Technology
ting.yu@inceptio.ai, lingxinjiang@gmail.com, wei.li@inceptio.ai

Abstract

Some recent works have demonstrated the remarkable accuracy achieved by the destination-driven approach in multimodal motion prediction, particularly with attention-based architecture. However, there are certain limitations regarding the performance of attention in the decoder network and the trajectory selection approach during inference. To address these limitations, we propose three enhancements: prior fusion and local map attention in the decoder network and score propagation in the trajectory selection. The proposed methods demonstrate competitive performance in terms of Mean Average Precision (mAP) and other evaluation metrics in the 2023 Waymo Motion Prediction Challenge.

1. Introduction

Recently, the task of multi-modal prediction has emerged as a crucial task [1, 4, 6, 8, 12, 14] in the field of autonomous driving. This task suffers significant challenges due to multiple factors, including uncertain behavior, multimodality in agent intentions, and only one ground truth.

Recent research demonstrates the promising capabilities of anchor-based methods which explicit to potential modality. Previous research has introduced several anchor-based methods, which can be mainly classified into two different categories. Some methods [7, 15] adopt the goals to serve as anchors, which are either sampled from the road or computed from the training dataset. Subsequently, waypoints are generated. Other methods [2, 13] utilize path-based anchors and employ regression techniques to refine the trajectory predictions toward the final desired paths.

These methods have achieved success in motion prediction. However, following the approach presented in [7], conventional goal-based strategies require a dense set of goals, which suffer from efficiency. To address this issue, some works adopt a limited number of initial goals, for example, generated by k-means clustering on trajectory endpoints in the training dataset, such as [12], and then refine the future trajectories with stacked layers of attention, which also results in an extremely deep and large model with substantial learning weights. Moreover, the simple stack of attention cannot capture the interaction between agents' motion and context features effectively. Hence, to overcome these limitations, we propose an effective destination-driven approach, namely DMP.

Specifically, we adopt a prior guided goal generator for goal prediction efficiency. Then, the prior fused query embedding method and local map attention approach are proposed to enhance the capacity of context features for prediction. Finally, we employ the score propagation method in trajectory selection during the inference.

2. Method

This section provides a comprehensive description of our proposed approach, DMP, which on a high level is similar to that of TNT [15]. Following the recent works [9, 10, 12], we use the transformer in our scene context encoder and decoder. The overview of our network architecture is illustrated in Fig. 1, including a hierarchical vector-based encoder for scene context encoding, a three-stage decoder for motion prediction, and a post-processing module for trajectory selection.

2.1. Encoder

Considering a scenario, DMP predicts the multimodal trajectory of the target agent based on multi-sourced contexts, including the motion history information of agents and the surrounding road graph.

Input representation. We employ an agent-centric strategy, following the representation in [12, 13, 15], wherein we transform both the agents' history information and the road graph to the target agent coordinate system. Following [5], we adopt a vector-based representation for all input objects, including agents' histories and map polylines, followed by a polyline encoder for object-level feature extraction. Specifically, for the Waymo Open Motion dataset [3], in the agent polyline encoder, the input features include object center position, heading, bounding box, and velocity, while the map polyline encoder incorporates polyline points sampled with a fixed interval, polyline type, direction, and traffic light

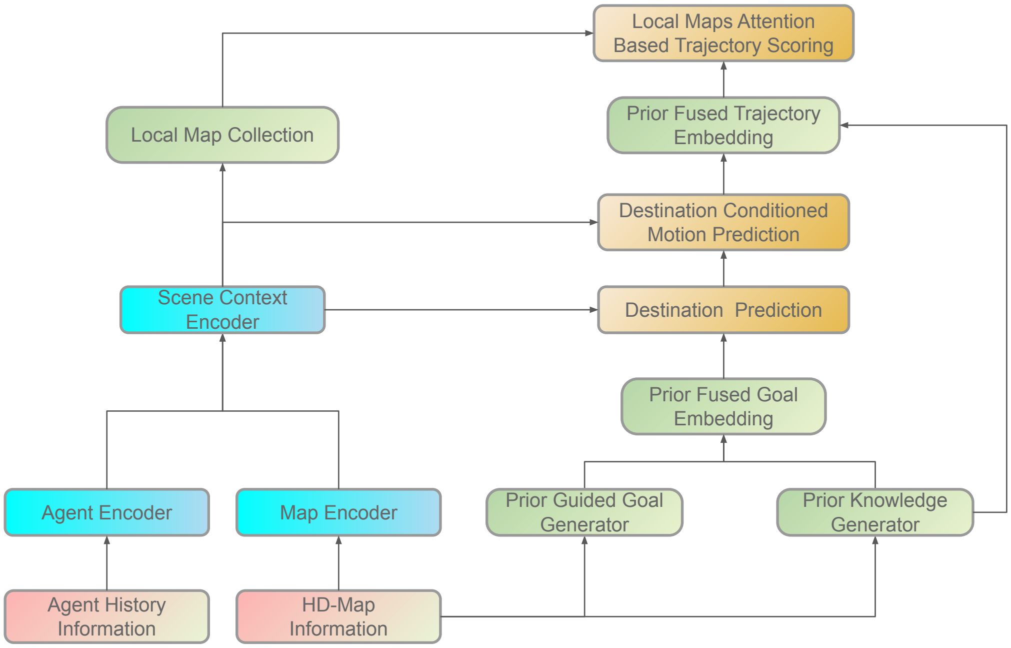
Local Maps Attention Based Trajectory Scoring
Local Map Collection
Prior Fused Trajectory Embedding
Destination Conditioned Motion Prediction
Scene Context Encoder
Destination Prediction
Prior Fused Goal Embedding
Agent Encoder
Map Encoder
Prior Guided Goal Generator
Prior Knowledge Generator
Agent History Information
HD-Map Information

Figure 1. The network architecture of our proposed Destination-driven Motion Prediction with Prior Fusion.

information.

Scene context encoder. In contrast to the previous intricate design present in [12, 13], we employ an original transformer encoder, with training efficiency, to model the interactions among the encoded objects.

2.2. Decoder

DMP shares a high-level similarity with the decoder approach employed in TNT [15]. As mentioned before, the transformer is also used in decoders. We present the key modifications and techniques utilized in our decoder, highlighting the effectiveness in modeling the interaction of future motions and context features.

Prior Guided Goal Generator. We follow the practice [7, 15] to generate goals on road maps. The difference is that instead of sample goals with a fixed interval, we use a dynamic interval method computed as follows:

$$I_d = I_c * v * \alpha, \tag{1}$$

where I_d and I_c denote the dynamic interval and default interval, respectively. v represents the velocity of the target agent at the observation timestamp and α is the scale factor.

Prior Fused Query Embedding. In the decoder, we employ a prior equipped query embedding, in all three stages. Our proposed prior fusion method aims to enhance the capability of context features for both goal prediction and trajectory scoring. In practice, in reference to the lane at the observed timestamp, the lanes are firstly classified into four categories: exit lanes, left change lanes, right change lanes, and forbidden lanes. These classifications indicate the accessibility of each lane according to the given lane topography and the maximum allowed number of exit or lane change actions. Exit lanes can be reached without any lane changes, left change lanes require at least one left lane change action, right change lanes require at least one right lane change action, and forbidden lanes cannot be accessed. Additionally, an extra class off-road, is employed to describe that the goal is not on the road(parking lots, etc). Then the annotation is encoded with one-hot representation and concatenated with goals or trajectories as query embedding. For off-map goals and trajectories, the distance to the closest lane is also encoded into query embedding as another prior.

Local Map Attention aims to focus on the interaction of future motions and local road maps. Specifically, in contrast to the dynamic map collection approach employed in [12], our method focuses on maps within a specified distance range from the trajectories. To achieve this, we apply an attention mask to exclude maps that fall outside the predefined range. This approach facilitates the interaction between trajectories and local maps.

2.3. Propagation and Selection

As the number of scored trajectories is M in the final stage of the network, an efficient trajectory selection algorithm has been employed during inference to determine the output trajectories. Inspired by the object detection work presented in [11], instead of the non-maximum suppression method employed in [15], we introduce the score propagation cluster approach in trajectory selection. Specifically, we consider a raw trajectory set denoted as

Table 1. Detailed evaluation metrics of our approach on the test set of Waymo Open Motion Dataset [3].

Category	mAP ↑	SoftmAP ↑	minADE ↓	minFDE ↓	Miss Rate ↓
Vehicle	0.4725	0.4826	0.7701	11.5400	0.1529
Pedestrian	0.4172	0.4224	0.3741	0.7882	0.0848
Cyclist	0.4005	0.4037	0.7436	1.4885	0.2043
Avg	0.4301	0.4362	0.6293	1.2723	0.1473

$T = \{t_1, t_2, \cdots, t_M\}$, which represents the output from the network. Additionally, we define a trajectory distance set, $D = \{d_{1,1}, \cdots, d_{1,N}, \cdots, d_{N,1}, \cdots, d_{N,N}\}$, where each element $d_{i,j} \in D$ represents the distance between trajectory pairs $(t_i, t_j \in T)$. The formulation of distance is defined as follows:

$$d_{i,j} = \frac{\sum_{k=N-10}^{N} ||t_{i,k} - t_{j,k}||_2}{10}, \tag{2}$$

where N represents the waypoint size, with $t_{i,k}$ and $t_{j,k}$ denote the k-th waypoint of trajectory i and j, respectively. We construct a set of graphs denoted as $G = \{g_1, g_2, \cdots\}$ based on the given D. Following the algorithm presented in [11], the score is updated by incorporating positive and negative messages. Finally, the top K trajectories with the highest scores are selected as the inference output.

3. Results

Table 1 presents the detailed per-class performance of our approach in the 2023 Waymo Open Dataset Motion Prediction Challenge. Our approach exhibits strong performance across various classes, especially in the vehicle and cyclist categories.

4. Conclusion

In this work, we present an effective approach for multimodal motion prediction. By adopting velocity-instructed goal generation, prior-fused embedding, local map attention, and score propagation, our proposed approach further improves the performance of the destination-driven method.

References

[1] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1793–1805. PMLR, 14–18 Dec 2023. 1

[2] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning*, 2019. 1

[3] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 1, 3

[4] 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*, pages 11525–11533, 2020. 1

[5] 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*, pages 11525–11533, 2020. 1

[6] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1

[7] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets, 2021. 1, 2

[8] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021. 1

[9] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021. 1

[10] 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 *International Conference on Learning Representations*, 2022. 1

[11] Yichun Shen, Wanli Jiang, Zhen Xu, Rundong Li, Junghyun Kwon, and Siyi Li. Confidence propagation cluster: Unleash full potential of object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1161, 2022. 2, 3

[12] 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*, 2022. 1, 2

[13] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S. Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, and Benjamin Sapp. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction, 2021. 1, 2

[14] 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*, pages 21995–22003, 2023. 1

[15] 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*, pages 895–904. PMLR, 2021. 1, 2