SMTP: Self-supervised Mapless Trajectory Prediction

Jianwei Ren

Abstract

Trajectory prediction is crucial for the safety of autonomous vehicles. Previous works rely on explicit modeling and representations of high-definition (HD) maps, which however limits the effective utilization of comprehensive scene information. This study introduces a framework that directly leverages bird's eye view (BEV) feature, rather than the decoded superficially informative map elements (e.g., rasterized or vectorized lanes). To improve the interaction between trajectory queries and scene representation, this framework reconstructs masked trajectory in a self-supervised manner. The experiments will be conducted on Argoverse 2 and nuPlan to demonstrate its effectiveness.

1. Introduction

Trajectory prediction is a challenging yet important task for safety of autonomous driving, since vehicles need to take into account the surrounding environment and the behavior of other agents to formulate appropriate plans. Recently, many learning-based methods [7] have demonstrated its effectiveness in modeling the scene-agent and agent-agent interactions.

Conventional studies [8, 9, 21, 26] often encode the provided high-definition (HD) maps for the perception of the scene context. As online mapping [19, 25] has progressed, the key elements of HD maps can now be well detected from the bird's eye view (BEV) features, derived from the surrounding multi-view images. Leveraging this advancement, recent end-to-end methods [13–15, 22] integrate online rendering of map elements into their frameworks, utilizing representations whether rasterized or vectorized. Although explicit map elements offer strong priors, they may eliminate a considerable amount of contextual information that is rather captured by intermediate features, i.e. BEV, as highlighted in [11].

This research proposal presents a query-based framework that optimizes the leveraging of BEV features, for the purpose of multi-agent multimodal trajectory prediction. The end-to-end framework adopts off-the-shelf BEV generation schemes like BEVFormer [17] and LSS [20], and subsequently engages the features in direct interaction with agents, without explicit element modeling. To improve the interaction between trajectory query and scene context, it performs a BERT-like [6] self-supervised reconstruction task. Specifically, prior to the integration of the query and BEV features into the transformer decoder, temporal random masking is applied to the representative query for each agent. Training signal will be derived from the recovery of the intact entire trajectory.

The experiments of this method will be conducted on the Argoverse 2 Motion Forecasting Dataset [23] and nuPlan database [2] to demonstrate its performance.

2. Related Work

End-to-End Architectures have proven their efficiency and superiority in the domain of trajectory prediction. These studies [10, 11, 13, 14, 18] follow the paradigm of simultaneously conducting perception and prediction tasks. Most of them rely on explicit maps to achieve better scene perception. While ViP3D [10] obtains ground truth map during the perception stage, UniAD [13], OccNet [22], PIP [14], and VAD [15] detect traffic elements from BEV features using DETR-like [3] designs. UniAD [13] and OccNet [22] exploit rasterized semantic and occupancy maps, whereas PIP [14] and VAD [15] adopt vectorized representations.

A recent study [11] advocates for the direct utilization of BEV features to prevent inefficiency and information loss. While this method introduces three strategies to enhance the model's ability to leverage BEV features, it fails to present a coherent framework.

Self-supervised Learning has shown considerable success in trajectory prediction. As a pioneer, VectorNet [7] regards graph completion as an auxiliary task for better interactions among nodes. SSL-Lanes [1] designs pretext tasks by capturing the relationships between the map and agents. In PreTraM [24], the insight of cross-modal similarity serves as the foundation for contrastive learning.

Recently, inspired by MAE [12], Forecast-MAE [5], Traj-MAE [4], and SEPT [16] perform mask-reconstruction tasks on trajectories and road maps to achieve efficient scene representation. This work also takes the notion of masked autoencoder. Through the prediction of missing trajectory segments, the transformer decoder is expected to re-

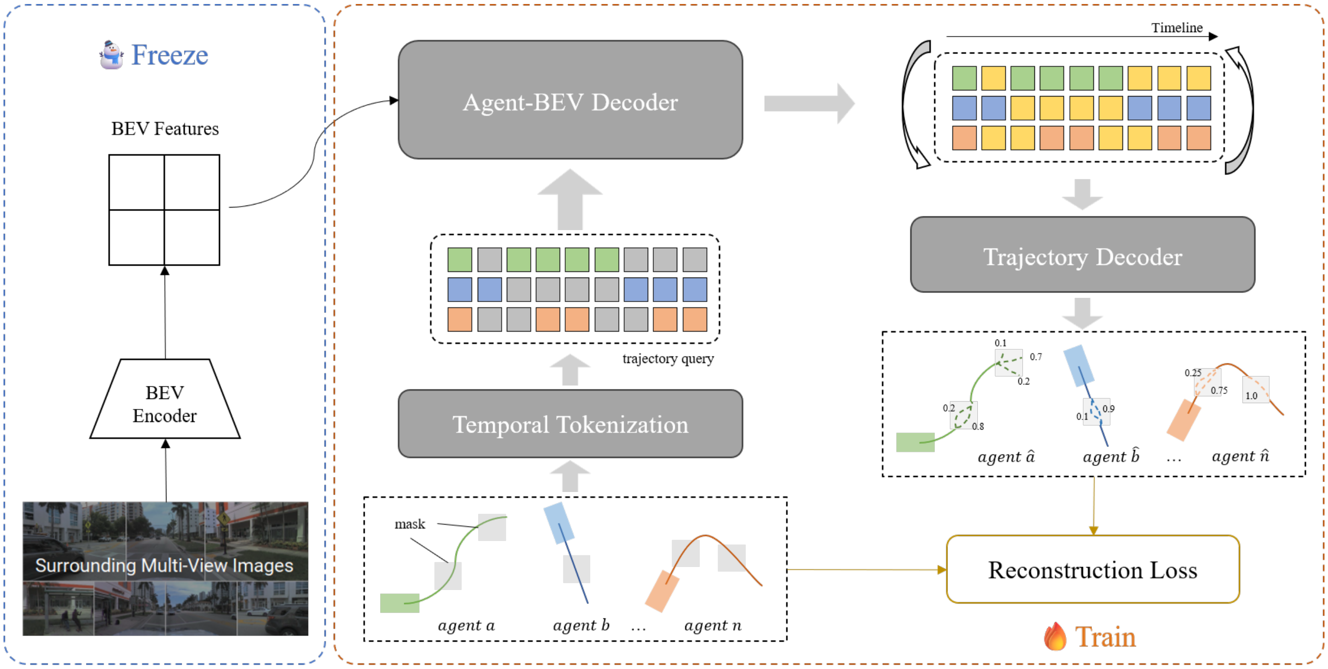

Figure 1. Overview of SMTP.

inforce latent features within BEV that benefit motion forecasting.

3. Method

3.1. Overview

The overall framework of SMTP is illustrated in Fig. 1. SMTP adopts a pretrained off-the-shelf encoder such as BEVFormer [17] to acquire BEV features. During training, the complete motions of N agents are provided. The trajectories are first subjected to random masking, and then tokenized while preserving the temporal integrity. The tokens, also identified as queries, engage in spatial cross-attention with BEV features to explore the underlying correlation. This spatial modeling of agent-BEV interactions is embedded into a transformer decoder, which then inpaints the motion features based on the unmasked segments. The aforementioned procedure converts the temporal trajectory of each agent into the spatial domain, thus allowing the model to concentrate on improving its spatial modeling ability. These reconstructed embeddings will pass through another decoder to output explicit multi-modal trajectories and probabilities. The decoder fully considers the social interactions and adjusts the pattern of fusion among features to obtain predictions for each agent.

3.2. Problem Formulation

This framework takes surrounding multi-view images and N agents' trajectories as input. The BEV features generated from vision are denoted as $X \subset \mathbb{R}^{H\times W\times C}$, where H, W and C represent height, width and channel respectively. The trajectory of the i-th agent on the entire timeline T is denoted as $P_i \subset \mathbb{R}^{T\times 2}$. During preprocessing, random masking will be applied on P_i in a predefined ratio α. Subsequently, the temporal-preserving module transforms trajectories of agents into query $Q \subset \mathbb{R}^{(N\times T)\times C}$. Given the final reconstruction after prediction head as $\hat{P}_i \subset \mathbb{R}^{T\times 2}$, the training loss $\mathcal{L}$ is designed in the form of $L1$ norm.

During inference, the framework masks the "future" segments to satisfy the task's definition.

3.3. Dataset and Metrics

Dataset. This method will be evaluated on the Argoverse 2 [23]. It includes 199,908, 24,988, and 24,984 samples for training, validation, and testing. Each sample contains 5-seconds history motions and 6-second future trajectories for each agent, with a sampling rate of 10 Hz.

The evaluation in nuPlan [2] is also scheduled to explore the potential application of this method in closed-loop scenarios.

Metrics. ADE is defined as the average displacement error between ground-truth trajectories and predicted trajectories over all time steps. FDE is defined as displacement error between ground-truth trajectories and predicted trajectories at the final time step. While this framework forecast up to 6 trajectories for each agent, minADE and minFDE are adopted. Miss rate (MR) refers to the percentage of the best-predicted trajectories whose FDE is within a threshold (2 m).

4. Conclusion

This research proposal outlines a plausible end-to-end framework for trajectory prediction. The framework suggests exploiting BEV features directly from the online mapping process, rather than the decoded traffic elements or the given HD maps. The motivation is that BEV features inherently contain all necessary information, potentially rendering downstream decoding redundant. To enhance interaction between BEV features and agents, the framework adopts a MAE-like self-supervised training paradigm. This study offers a novel solution to existing multi-agent multi-modal trajectory prediction.

References

[1] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. In *Conference on Robot Learning*, pages 1793–1805. PMLR, 2023. 1

[2] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 1, 2

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

[4] 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*, pages 8351–8362, 2023. 1

[5] 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*, pages 8679–8689, 2023. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[7] 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

[8] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Home: Heatmap output for future motion estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 500–507. IEEE, 2021. 1

[9] 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*, pages 15303–15312, 2021. 1

[10] 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

[11] Xunjiang Gu, Guanyu Song, Igor Gilitschenski, Marco Pavone, and Boris Ivanovic. Accelerating online mapping and behavior prediction via direct bev feature attention. *arXiv preprint arXiv:2407.06683*, 2024. 1

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[13] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[14] Bo Jiang, Shaoyu Chen, Xinggang Wang, Bencheng Liao, Tianheng Cheng, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, and Chang Huang. Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction. *arXiv preprint arXiv:2212.02181*, 2022. 1

[15] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 1

[16] Zhiqian Lan, Yuxuan Jiang, Yao Mu, Chen Chen, and Shengbo Eben Li. Sept: Towards efficient scene representation learning for motion prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2

[18] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 1

[19] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022. 1

[20] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 1

[21] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022. 1

[22] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 1

[23] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1, 2

[24] Chenfeng Xu, Tian Li, Chen Tang, Lingfeng Sun, Kurt Keutzer, Masayoshi Tomizuka, Alireza Fathi, and Wei Zhan. Pretram: Self-supervised pre-training via connecting trajectory and map. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 1

[25] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024. 1

[26] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 1