Advancing Online HD Map Construction with Foundation Models and Multi-Task Contextual Learning

Ivanov Leonid Dmitry Yudin
Moscow Institute of Physics and Technology
Moscow, Russia
ivanov.la@mipt.ru, yudin.da@mipt.ru

Abstract

In autonomous driving, high-definition (HD) maps and semantic maps in bird's-eye view (BEV) are essential for accurate localization, planning, and decision-making. This paper introduces an enhanced framework for online HD map generation. We significantly boost feature representation quality by incorporating powerful foundation models for encoding both camera images and LiDAR point clouds. To further enrich the model's understanding of the environment and improve prediction quality, we integrate auxiliary prediction heads for semantic segmentation of the BEV representation. This multi-task learning approach provides richer contextual supervision, leading to a more comprehensive scene representation and ultimately resulting in higher accuracy and improved quality of the predicted vectorized HD maps. The source code is available at https://github.com/LIvanoff/repo_name.

1. Introduction

High-definition (HD) semantic maps are among the most important components of autonomous driving technology. The use of semantic information about the environment, such as lane boundaries, road dividers and pedestrian crossings, plays a crucial role in vehicle navigation, motion prediction, and planning. Traditional methods of HD map construction, such as Simultaneous Localization and Mapping (SLAM) and manual map annotation, require significant time and human effort. These limitations are driving a shift towards online, learning-based methods that leverage onboard sensor data.

A pioneering work in the field of learning-based methods for HD map construction is the HDMapNet [12], which perform semantic segmentation in BEV space by producing rasterized maps. Several existing methods [9, 24, 27, 35] allow the construct vectorized HD maps through post-processing of rasterized images. However, these approaches often demonstrate inefficiencies, both in the quality of the resulting vectorized maps and in the speed of the post-processing pipeline. Alternatively, recent approaches showcase the benefit of leveraging rasterized maps not as the primary output requiring vectorization, but as an efficient intermediate representation or prior. A notable example is HRMapNet [34], which effectively utilizes a historical rasterized map as a low-cost source of complementary information. This historical context significantly improves the performance of the main online vectorized map generation module.

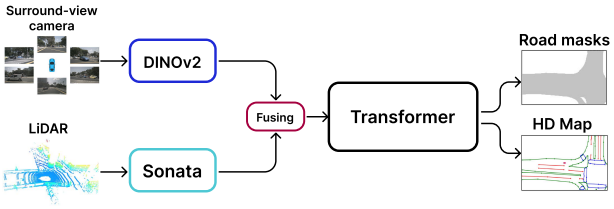

Figure 1. The general pipeline of the proposed multimodal method predicts a vector HD map and raster masks of the road.

To enhance the quality of vectorized map prediction and address the limitations of traditional segmentation methods, recent approaches have shifted towards directly generating sets of points for map construction. These methods [6, 8, 11, 13–16, 18], leveraging transformer decoders similar to DETR [3], allow for the immediate generation of structured maps without the need for additional post-processing.

Existing online HD map methods often employ standard backbones like ResNet [10], potentially limiting representation power compared to modern foundation models. We propose leveraging the robust features from large-scale pre-trained models like DINOv2 and Sonata to enhance the perception module for map generation. This novel application aims to improve map accuracy by building on a stronger feature foundation.

In this paper, we introduce an efficient framework de-

signed to significantly improve the accuracy and robustness of online vectorized HD map generation. Our approach makes two key contributions. First, departing from standard task-specific encoders, we harness the representational power of large-scale pre-trained foundation models, namely DINOv2 [20] and Sonata [30], for initial image and point cloud feature extraction, which, to our knowledge, is a novel application in this domain. Second, to further enhance the model's understanding of the driving environment, we introduce auxiliary heads for BEV semantic segmentation of driveable area on different level.

To summarize, this paper makes the following contributions:

- We propose CFMapNet, an online end-to-end HD map construction framework. It leverages the MapQR query mechanism for efficient map element instance modeling and introduces auxiliary BEV semantic segmentation heads, enabling the model to learn a richer understanding of the scene context through multi-task supervision.
- We pioneer the integration and demonstrate the effectiveness of using pre-trained vision foundation models (DINOv2 for camera images and Sonata for LiDAR point clouds) as powerful feature extractors for this task.

2. Related works

2.1. Online HD Map construction

Online HD map construction seeks to generate maps directly from onboard sensors, mitigating issues of cost and freshness associated with offline methods. Early approache relied on BEV semantic segmentation [12], but required complex post-processing to obtain vectorized outputs, often losing precision and instance details [11]. This led to the development of end-to-end methods that directly predict vectorized representations.

To address these limitations, subsequent research focused on directly predicting vectorized map elements. VectorMapNet [16] was a pioneering end-to-end method using a two-stage approach with an auto-regressive decoder to predict points sequentially for each map instance. However, its sequential nature can be slow and potentially prone to error accumulation.

2.2. Transformer decoder

Inspired by the success of DETR [4] in object detection, MapTR [13] introduced a paradigm shift by modeling map elements as sets of points and using a parallel transformer decoder. This has improved the efficiency and speed of the end-to-end approach. MapTRv2 [14] further refined this by introducing decoupled self-attention for efficiency, auxiliary losses for better supervision, and support for directed map elements like lane centerlines. Other contemporary works have explored different representations or mechanisms. BeMapNet [21] and PivotNet [8] utilize Bézier curves or pivot points, respectively, for a more compact representation of map elements compared to raw point sequences. Addressing temporal consistency, StreamMapNet [33] and MapTracker [6] introduced streaming architectures that propagate information (BEV features or queries) across time frames. Methods like HRMapNet [34] and GlobalMapNet [25] leverage historical or global map information (often rasterized or vectorized) as an explicit prior to enhance current predictions. More recently, MapQR [18] focused on improving the query design within the DETR-like framework using a scatter-and-gather mechanism for better information probing and instance modeling. Our work builds upon these DETR-like approaches, particularly improving the query mechanism and feature representation.

2.3. Foundation model

In recent years, there has been a paradigm shift with the advent of Foundation models. These models are typically large-scale neural networks pre-trained on vast amounts of broad, often unlabeled, data using SSL techniques. Their key characteristic is the ability to serve as a strong "foundation" for a wide variety of downstream tasks, often achieving remarkable performance with minimal task-specific fine-tuning (few-shot or zero-shot learning) or by simply providing powerful, general-purpose feature representations.

In Computer Vision This trend marked a move beyond supervised pre-training on datasets like ImageNet [22]. Models like CLIP [19] learned joint image-text embeddings from web-scale data, enabling impressive zero-shot image classification. MAE (Masked Autoencoders) [26] and BEiT [1] adapted the masked language modeling idea from BERT [7] to vision, learning rich representations by reconstructing masked patches of an image. DINO [5] and DINOv2 [20] leverage self-distillation and self-supervised learning to extract transferable visual features for dense prediction tasks without fine-tuning. Pre-training on large, diverse datasets makes them strong general-purpose vision backbones. Previously the BEVCar [23] method used a ViT-Adapter for frozen DINOv2 for semantic segmentation and 3D object detection.

For 3D Point Cloud Self-supervised learning is challenging due to the "geometric shortcut," where models learn trivial spatial cues, hindering semantic feature learning, particularly evident in poor linear probing results [30]. While early masking by PointBERT [32] or contrastive [31] methods existed, Sonata [29], using a PointTransformerV3 (PTv3) [28] backbone and point self-distillation, specifically targets this issue. By employing techniques to obscure geometry, Sonata achieves state-of-the-art linear probing results, signifying robust feature learning suitable for downstream tasks.

Previous online HD map construction predominantly used standard backbones like ResNet [10]. We propose leveraging features from large-scale foundation models like Sonata and DINOv2. To our knowledge, this is the first application of such models as direct feature extractors for online vectorized HD map construction. We aim to improve map accuracy and robustness by building upon this stronger perceptual foundation.

3. Method

3.1. Overall architecture

Our proposed method, CFMapNet introduces significant enhancements in feature representation and contextual understanding through multi-task learning and the integration of foundation models. The overall architecture, depicted in Fig. 1, retains the core components: a feature extraction backbones, a fusing module for combining two modalities, and a transformer for map element prediction, but incorporates key modifications.

3.2. Foundation Model Backbones

To capture richer and more generalizable features from sensor inputs, we replace standard backbones with powerful pre-trained foundation models. For camera inputs, the conventional ResNet [10] is substituted with a frozen DINOv2 [20] model, leveraging its strong visual representations learned from large-scale self-supervised pre-training. In configurations utilizing LiDAR data, the standard SECOND encoder is replaced with the Sonata [30] foundation model, known for its ability to learn reliable geometric and semantic features from point clouds while mitigating the "geometric shortcut" problem. These foundation models serve as the perceptual cornerstone of our framework.

3.3. Enhanced BEV Encoder with Dual Queries

A key novelty lies in enhancing the BEV representation learning. Instead of solely optimizing the BEV encoder for predicting vectorized map elements (thin lines like lane dividers or road boundaries), we introduce an auxiliary task to explicitly learn the spatial extent and semantics of the drivable area. To facilitate this multi-faceted feature learning within the BEV encoder (or view transformation module), we potentially introduce two types of queries:

Vector Queries which similar to MapQR, these queries are designed to probe BEV features relevant for localizing precise vector representations of map elements.

Raster Queries These queries are introduced to encourage the BEV features to capture broader spatial context and areal information, specifically focusing on identifying the drivable road surface plane. This dual-query mechanism encourages the BEV encoder to produce features beneficial for both precise vector prediction and broader semantic area understanding.

3.4. Auxiliary Road Surface Segmentation Head

Building upon the enhanced BEV features, we add an auxiliary segmentation head. Unlike segmentation heads predicting fine boundary lines, this head is trained to predict a BEV mask representing the drivable road surface area. This auxiliary task provides direct supervision for understanding the main drivable region, forcing the shared BEV encoder to learn features that distinguish the road plane from surrounding elements (sidewalks, buildings, etc.). This contextual information implicitly aids the primary task of accurately predicting vectorized HD map elements handled by the main MapQR-style decoder.

3.5. Vector Map Decoder

For the primary task of generating vectorized HD map elements, we retain the core MapQR decoder architecture [18] with its efficient scatter-and-gather query mechanism and positional embeddings. This decoder takes the contextually enriched BEV features from our enhanced encoder and the instance queries to predict the final set of vectorized map elements (lane dividers, road boundaries, pedestrian crossings, etc.).

4. Experiments

We compare the proposed method CFMapNet with SOTA methods, including MapTR series [13, 14] PivotNet [8] and our baseline MapQR [18].

4.1. Settings

We evaluate our method on public datasets nuScenes [2]. These dataset provide surrounding images captured by self-driving cars in hundreds of city scenes. Each sample contains 6 RGB images and 200K points in LiDAR's point cloud. nuScenes provides 2D vectorized map elements as ground truth.

Evaluation Metric Following previous works [12,14], three static map categories are used for performance evaluation, i.e., lane divider, pedestrian crossing, and road boundary. Chamfer distance is used to determine whether the prediction matches the ground truth. Besides the thresholds 0.5m, 1.0m, 1.5m used in MapTR [14], the results are also evaluated with smaller thresholds 0.2m, 0.5m, 1.0m used in BeMapNet [21]. The corresponding mAP metrics are denoted as mAP2 and mAP1, respectively. Training and Inference Details. We follow most settings as in the MapTR series [13, 14], and the modifications will be emphasized. The BEV feature is set to be 200×100 to perceive [30m, 30m] for rear to front, and [15m, 15m] for left to right. We use 100 instance queries to detect map element instances,

Table 1. Comparison baseline with Segmentation Head on nuScenes

Method	Modalities	Image Backbone	AP_{div}	AP_{ped}	AP_{bound}	mAP1
MapTR [13]	C	ResNet50	51.5	46.3	53.1	50.3
MapTR [13]	C	SwinT	45.2	52.7	52.3	50.1
PivotNet [8]	C	ResNet50	56.5	56.2	60.1	57.6
PivotNet [8]	C	SwinT	60.6	59.2	62.2	60.6
MapTRv2 [14]	C	ResNet50	62.4	59.8	62.4	61.5
MapQR [18]	C	ResNet50	68	63.4	67.3	66.3
MapQR [18]	C	SwinT	68.1	63.1	67.1	66.1
(ours) MapQR + Seg. Head	C	ResNet50	**68.8**	**64.9**	**67.6**	**66.7**
(ours) MapQR + Seg. Head	C	SwinT	68.1	**64.1**	**68**	**67.1**

We are comparing the effect of adding an Auxiliary Road Surface Segmentation Head to the MapQR [18] without any additional changes on the nuScenes validation split. Our baseline shows a slight superiority with the backbone ResNet50 [10], while with SwinT [17] we improve the quality by 1%, while the previous approaches had a slightly worse version with SwinT than the version with ResNet50 for all except PivotNet.

and each instance is modelled by 20 sequential points. (We set N = 100 and n = 20.) Our model is trained on 2 NVIDIA V100 GPUs with the total batch size of 32. The learning rate is set to 6e-4.

4.2. Comparisons with State-of-the-art Methods

Auxiliary Road Surface Segmentation Head To evaluate the effectiveness of incorporating scene context understanding via multi-task learning, we investigate the impact of adding an Auxiliary Road Surface Segmentation Head to our baseline MapQR model [18]. As discussed in Section 3.4, this auxiliary head is trained to predict a BEV mask representing the drivable road surface area, providing additional supervision to the shared BEV encoder without altering the main MapQR decoder or its vectorized map prediction task.

The results clearly demonstrate the benefit of the auxiliary segmentation task. With the ResNet50 backbone, adding the segmentation head yields a modest but consistent improvement, increasing the mAP1 from 66.3% for the baseline MapQR to 66.7% for MapQR + Seg. Head. This suggests that even with a standard backbone, explicitly supervising the model to understand the road plane helps refine the features used for vectorized map element prediction.

More notably, with the SwinT backbone, the improvement is significantly more pronounced. While the baseline MapQR with SwinT (66.1% mAP1) slightly underperformed its ResNet50 counterpart (66.3% mAP1), adding the segmentation head boosts the performance substantially to 67.1% mAP1. This represents a 1.0% absolute improvement over the SwinT baseline. Interestingly, this enhancement not only overcomes the initial deficit compared to ResNet50 but establishes the SwinT version of MapQR + Seg. Head as the top-performing configuration among those tested. This contrasts with the behavior observed for MapTR and MapTRv2, where SwinT generally did not provide an advantage over ResNet50 in the original papers, suggesting a potential synergy between the SwinT architecture and the auxiliary road segmentation task within our framework.

Overall, these comparisons validate our hypothesis: integrating an auxiliary task focused on understanding the broader spatial context, specifically the road surface area, serves as effective guidance for the BEV encoder. This leads to the generation of more accurate vectorized HD map elements, pushing the state-of-the-art performance for online map construction, particularly when combined with advanced backbones like SwinT.

5. Conclusion

...

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 2

[2] 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, 2020. 3

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

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 2

[6] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping, 2024. 1, 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[8] Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. PivotNet: Vectorized pivot learning for end-to-end HD map construction. In *ICCV*, pages 3672–3682, 2023. 1, 2, 3, 4

[9] Hao Dong, Xianjing Zhang, Jintao Xu, Rui Ai, Weihao Gu, Huimin Lu, Juho Kannala, and Xieyuanli Chen. SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation. *arXiv preprint arXiv:2211.15656*, 2022. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1, 3, 4

[11] Haotian Hu, Fanyi Wang, Yaonong Wang, Laifeng Hu, Jingwei Xu, and Zhiwang Zhang. Admap: Anti-disturbance framework for reconstructing online vectorized hd map, 2024. 1, 2

[12] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework, 2021. 1, 2

[13] 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. In *ICLR*, 2022. 1, 2, 3, 4

[14] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. MapTRv2: An end-to-end framework for online vectorized HD map construction. *arXiv preprint arXiv:2308.05736*, 2023. 2, 3, 4

[15] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. MGMap: Mask-guided learning for online vectorized HD map construction. In *CVPR*, pages 14812–14821, 2024.

[16] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. VectorMapNet: End-to-end vectorized HD map learning. In *ICML*, pages 22352–22369, 2023. 1, 2

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 4

[18] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *European Conference on Computer Vision*, 2024. 1, 2, 3, 4

[19] Nhan T. Luu. Clip unreasonable potential in single-shot face recognition, 2024. 2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2, 3

[21] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve, 2023. 2, 3

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 2

[23] Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 1435–1442. IEEE, 2024. 2

[24] Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. *arXiv preprint arXiv:2403.11761*, 2024. 1

[25] Anqi Shi, Yuze Cai, Xiangyu Chen, Jian Pu, Zeyu Fu, and Hong Lu. Globalmapnet: An online framework for vectorized global hd map construction. *arXiv preprint arXiv:2409.10063*, 2024. 2

[26] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. The effectiveness of mae pre-pretraining for billion-scale pretraining, 2024. 2

[27] Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5186–5195, 2023. 1

[28] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 2

[29] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025. 2

[30] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations, 2025. 2, 3

[31] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, 2020. 2

[32] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling, 2022. 2

[33] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction, 2023. 2

[34] Xiaoyu Zhang, Guangwei Liu, Zihao Liu, Ningyi Xu, Yunhui Liu, and Ji Zhao. Enhancing vectorized map perception with historical rasterized maps, 2024. 1, 2

[35] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, pages 13760–13769, 2022. 1