Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070111
ISSN 2516-5003 Vol. 7 Num. 1

Unified Prior Mask-Guided End-to-End Online Vectorized HD Map Construction

Yupeng Luo[1,a], Yuan Zhu[1,b], Ke Lu[1,c,*]

[1] *College of Automotive and Energy Engineering, Tongji University, No. 4800 Caoan Road, Shanghai, China*
[a]*ypluo@tongji.edu.cn*, [b]*yuan.zhu@tongji.edu.cn*, [c]*luke@tongji.edu.cn*
**Corresponding author*

Keywords: Vectorized High-Definition Map Construction, Autonomous Driving, Bird's-Eye View Perception, Selective State Space Model

Abstract: Online vectorized high-definition map construction has attracted increasing attention, which reduces the cost of manual annotation compared with traditional SLAM methods and provides critical static road information for downstream tasks such as localization and motion planning. However, while standard-definition maps and global maps from historical predictions are readily available priors, existing methods fail to uniformly and fully leverage this valuable information, leading to suboptimal mapping performance especially in long-range complex scenarios. To address this issue, we propose a prior-guided mapping method leveraging a unified prior mask, termed PGMapNet. Specifically, we design a unified mask-guided prior embedding generator, which fuses bird's-eye view (BEV) features with prior maps to generate a unified prior mask. The mask is further utilized to predict map instance information and generate prior embeddings, thereby providing positional and structural instance priors for the queries of the map decoder. Furthermore, a multi-point selective state space model (SSM) module is designed, which adaptively samples key region features of map instance points from BEV features, and performs interactive modeling on the sampled sequences via SSM, effectively enhancing the prediction accuracy of the map instance point set. Extensive experiments on the nuScenes dataset validate the effectiveness of the proposed method. Compared with the baseline model, our proposed model achieves an improvement of 2.5% mAP within the 30-meter range, and a further improvement of 2.9% mAP for the extended 50-meter range.

1. Introduction

High-definition (HD) maps serve as a critical component in autonomous driving systems, providing vehicles with essential information on static environmental features. They compensate for scenarios such as occlusion or sensor malfunctions, and furnish vital feature information for downstream tasks including localization, prediction and planning. Traditional HD maps are typically constructed via simultaneous localization and mapping (SLAM) approaches[1,2]. However, such methods rely on high-precision acquisition equipment and extensive manual post-annotation. Furthermore, they are insensitive to dynamic map changes and suffer from limited freshness, which

hinders their practical deployment in autonomous driving systems. Therefore, online HD map construction has attracted widespread attention, which leverages onboard sensors for real-time perception to predict key map elements. Vectorized HD maps usually comprise elements such as pedestrian crossings, lane dividers and road boundaries, which are represented as vectorized point sequences. Compared with raster maps, they can effectively reduce storage overhead. Some methods are implemented based on semantic segmentation, such as HDMapNet[3]. However, additional post-processing is required to obtain vectorized features, which impairs the real-time performance of the model. Subsequently, an increasing number of methods adopt an end-to-end paradigm to directly predict vectorized map elements based on Transformer architectures. VectorMapNet[4] proposes the first framework for end-to-end learning of vectorized HD maps, using polylines as the representation of map elements and transforming the mapping task into a sparse set detection problem. MapTR[5] proposes an equivalent representation method for map elements, and realizes unified map element decoding through consistent instance and point embeddings. MGMap[6] further boosts the model performance by predicting masks from BEV features to extract instance information, thereby guiding the decoder to address the sparse prediction problem. P-MapNet[7] leverages standard-definition (SD) maps as prior information to improve the mapping performance at long distances. To enhance the model's detection performance in occluded scenarios and long-distance ranges, StreamMapNet[8] further enhances detection accuracy through temporal fusion. On this basis, GlobalMapNet[9] enables on-vehicle global map construction, thus providing important global prior information for map prediction. Nevertheless, existing methods fail to effectively and uniformly fuse two types of readily available prior information, namely SD maps and global maps of historical predictions. SD maps typically provide road centerline information with low accuracy, and there exist significant semantic and geometric deviations between such maps and the vector elements predicted by onboard sensors.

To address this issue, this paper proposes a mapping method guided by a unified prior mask, which fully leverages the two types of prior maps to improve mapping accuracy. Specifically, a prior embedding generator based on a unified mask is designed. This generator fuses BEV features with the two prior maps to generate a unified mask for map instance prediction, and produces prior embeddings that are injected into map queries, thereby providing critical prior positional and structural information. Furthermore, we propose a multi-point selective state space model (SSM) module. By adaptively sampling information from key regions and performing sequence prediction with the SSM, the decoder effectively improves the detection accuracy of map elements. The contributions of this paper are summarized as follows:

(1) A vectorized HD map construction method guided by a unified prior mask is proposed, which fully exploits SD maps and global maps generated by historical predictions as prior information to effectively enhance mapping accuracy.

(2) A prior embedding generator based on a unified mask is designed. Targeting the element differences between the two prior maps, the generator fuses them with BEV features as the carrier and predicts prior instance embeddings, which furnish critical positional and structural prior information for map queries. Meanwhile, a multi-point SSM module is proposed, which improves the representational and learning capability for point sequences of map instances.

(3) Extensive experiments on the nuScenes dataset validate the effectiveness of our proposed method, which achieves 2.5% mAP gain within 30 m and 2.9% mAP gain over the extended 50 m range against the baseline.

2. Methodology

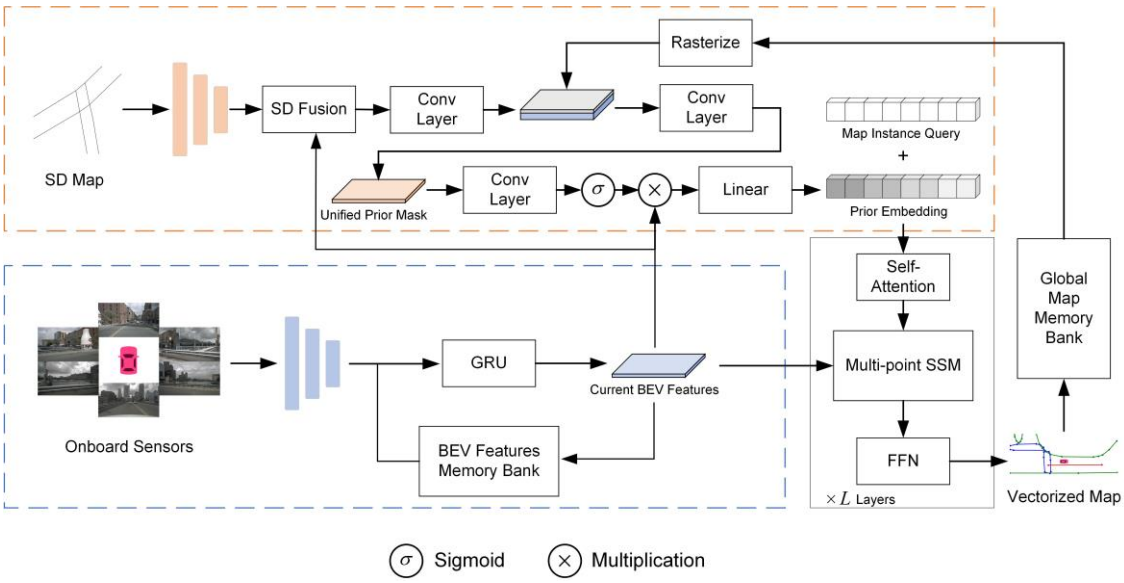


Figure 1: Overall architecture of our proposed PGMapNet.

Our model processes synchronized surround-view images to generate lightweight vectorized HD maps at both local and global scales. Map elements consist of categorical labels and polyline point sequences. The model is primarily composed of four components, including a BEV feature encoder, a unified mask-guided prior embedding generator, a DETR-like map element decoder, and a global map memory bank. First, the surround-view images are encoded by a BEV feature encoder to generate BEV features with spatiotemporal consistency. Next, the unified prior mask is generated and map instances are predicted by fusing the BEV features with the two prior maps, ultimately yielding prior embeddings that are injected into map queries. Subsequently, in the map element decoder, three types of map element instances are decoded through multi-layer fusion with the BEV features, resulting in the local map. Finally, following GlobalMapNet[9], the generated vectorized local map is saved via transformation to the global coordinate system, which is utilized to provide global priors for subsequent map prediction.

2.1. BEV Feature Encoder

First, a shared CNN image backbone and a Feature Pyramid Network (FPN) are employed to extract and aggregate 2D features from multi-view images. Subsequently, a BEV feature extractor samples the corresponding features from the BEV space to the 2D image space, yielding BEV features with spatial consistency. Then, following StreamMapNet[8], a temporal fusion module based on a Gated Recurrent Unit (GRU) is used to recurrently fuse historical BEV features, thereby constructing BEV features F_{BEV} with spatiotemporal consistency.

2.2. Unified Mask-Guided Prior Embedding Generator

To uniformly exploit the information from the two forms of prior maps, BEV features are used as a carrier to fuse with the two prior maps for generating a unified prior mask. This mask is further processed to generate instance information, which is combined with the current BEV features to predict prior embeddings that are injected into the queries of the map decoder.

Based on the vehicle's current GPS positioning information, the SD map data of the

corresponding region is extracted from OpenStreetMap[10] and then rasterized. However, due to positioning errors and the low accuracy of the SD map, a certain misalignment exists between the SD map and BEV features. Inspired by P-MapNet[7], a fusion method based on multi-head cross-attention is adopted to achieve adaptive alignment, thus enabling effective fusion. Specifically, to reduce the computational cost, the BEV features F_{BEV} and the SD map features F_{SD} are downsampled by a convolutional network. The BEV features F_{BEV} are integrated with sine positional embeddings, then fused with F_{SD} in multi-head attention layers, and finally upsampled to generate a prior mask M_{SD}.

When a historical predicted global map exists for the region the vehicle is traversing, the model extracts the corresponding regional data and fuses it into the prior mask. Specifically, based on the vehicle's current pose, the historical predicted global map of the corresponding region is extracted from the global map memory bank and projected into the ego-vehicle coordinate system via the pose matrix, achieving alignment with the BEV features. Next, the global map elements are rasterized, concatenated with the prior mask M_{SD} encoded by convolutional layers, and further fused through additional convolutional layers, as illustrated in Figure 1.

The updated unified prior mask M_U contains the instance information of the two prior maps, which is used to generate prior instance embeddings and provide critical positional and structural information for subsequent map queries. Specifically, the unified prior mask M_U is first processed by convolutional layers with sigmoid activation function to extract instance maps M_{ins}, which are then multiplied by the transposed BEV features F_{BEV} to generate the prior instance embeddings. Subsequently, the prior instance embeddings are processed by a linear layer and injected into the map queries, which are then fed into the DETR-like decoder to predict the point set sequences of map elements.

2.3. Multi-Point SSM Module

In 3D object detection, BEVFormer[11] assumes a single reference point as an anchor relative to the object center, allowing each query to sample features from the generated BEV features. However, a map element exhibits a highly irregular and elongated shape, resulting in non-locality in the BEV space. Therefore, StreamMapNet[8] proposes a multi-point attention mechanism, which replaces the single central anchor with multiple predicted points to adapt to the map construction task.

This design, however, neglects the relationships among sampling points across different queries. To fully model the intra-query sampling point dependencies and inter-query sampling point correlations, we design a multi-point map instance decoder based on the State Space Model (SSM). Specifically, similar to Deformable DETR[12], sampling offsets are first derived from queries through a linear layer, and then the reference points are updated to obtain sampling positions. Bilinear interpolation is then applied to extract key sampled features from the BEV features, which are further organized into sequences. Subsequently, we employ Mamba-2[13], which is an SSM model with efficient implementation based on matrix multiplication, to process the sampled feature sequences, followed by a two-layer convolutional network to generate sequences for each query. The relevant equations are as follows.

$$O_i = f_{offset}(Q_{i-1}) \tag{1}$$

$$x_{sample} = grid\_sample(F_{BEV}, P_{i-1} + O_i) \tag{2}$$

$$Q_i = Conv(SSM(x_{sample})) \tag{3}$$

Here, i denotes the i-th layer, and O_i denotes the sample offsets predicted by the previous query Q_{i-1} through a linear layer f_{offset}. P_{i-1} represents the previous coordinates which are added to O_i to generate new sample positions. x_{sample} denotes the sampled sequences, which are processed through $SSM(\cdot)$ and convolutional layers $Conv(\cdot)$ to generate the updated query Q_i.

Compared with traditional deformable attention, this method enables both intra-query and inter-query interactions among sampling points, which facilitates the effective differentiation of distinct map instances. Benefiting from the selective mechanism of SSM, the adaptive encoding of point sequences enhances their representational capacity and promotes a more comprehensive scene understanding.

2.4. Loss Function

The model is trained in an end-to-end manner, and standard bipartite matching is utilized to associate predicted map instances containing class and point sequence (c_{pred}, p_{pred}) with ground-truth instances (c_{gt}, p_{gt}). The matching cost for polyline is defined as Smooth L1 Loss, while the matching cost for category classification adopts Focal Loss, and the final matching cost is formulated as follows.

$$C_{match} = \alpha_1 L_{SmoothL1}(p_{pred}, p_{gt}) + \alpha_2 L_{Focal}(c_{pred}, c_{gt}) \tag{4}$$

For the overall loss function, Focal Loss is employed for the classification branch, and Smooth L1 Loss is adopted for the positional regression of point set sequences. In addition, to enhance the learning capability of the unified prior mask M_U, we supervise the instance maps M_{ins} generated from M_U. Meanwhile, we predict a semantic mask M_{seg} from M_U via a convolutional layer and apply auxiliary supervision. The combination of Cross-Entropy Loss and Dice Loss is adopted for both instance masks and the semantic mask. The overall loss is formulated as follows.

$$L = \beta_1 L_{SmoothL1}(p_{pred}, p_{gt}) + \beta_2 L_{Focal}(c_{pred}, c_{gt}) + \beta_3 L_{ins}(M_{ins}, M_{gt\_ins}) + \beta_4 L_{\text{seg}}(M_{seg}, M_{gt\_seg}) \tag{5}$$

3. Experiments

3.1. Dataset and Implementation Details

Extensive experiments are conducted to validate our model on the nuScenes[14] dataset. The nuScenes dataset offers abundant scenarios for autonomous driving, which consists of 1000 scenes providing six synchronized cameras with a resolution of 900×1600 and accurate ego-vehicle poses. The whole dataset is divided into training, validation and test sets with 700, 150 and 150 scenes respectively. Each scene lasts for 20 seconds, and key samples are annotated at a frequency of 2 Hz. However, as stated in StreamMapNet[8], an overlap of over 84% of locations exists in the training and validation sets under the official split, which leads to training data leakage into the validation set and is unfavorable for evaluating the generalization ability of the model. To address this issue, we adopt the split strategy proposed in StreamMapNet to ensure that the scenes in the validation set are totally unseen during training.

Our model is trained on 8 A800 GPUs with a total batch size of 32, and the AdamW optimizer is adopted with an initial learning rate of 5×10^{-4}. We employ ResNet50 as the image backbone and

utilize the BEV feature extractor inherited from BEVFormer[11]. We build the overall model based on GlobalMapNet[9], which is constructed upon StreamMapNet, and adopt consistent hyperparameter settings and data augmentation processing. For the SSM settings, the hidden state size is set to 64, while the feature expansion factor is set to 2. For mask supervision, β_3 and β_4 are set to 2 and 15, respectively. The model is trained for 24 epochs on the nuScenes dataset.

To be consistent with existing state-of-the-art works, we consider three types of map elements, namely pedestrian crossings, lane dividers and road boundaries. We conduct model evaluation under two distance ranges, including the conventional range of 30 m forward and backward with 15 m left and right, as well as an extended range of 50 m forward and backward with 25 m left and right. Average Precision (AP) is adopted as the evaluation metric, which is initially proposed by HDMapNet[3] and VectorMapNet[4]. Specifically, AP is calculated and evaluated under thresholds of $\{0.5\ m, 1.0\ m, 1.5\ m\}$ for the 30 m range and $\{1.0\ m, 1.5\ m, 2.0\ m\}$ for the 50 m range, respectively.

3.2. Main Experimental Results

Table 1: Quantitative comparison of models in the new split of nuScenes validation set.

Range	Method	Backbone	Image Size	AP_{ped}	AP_{div}	AP_{bound}	mAP
60 × 30 m	StreamMapNet	ResNet50	480 × 800	28.3	30.5	41.3	33.4
	GlobalMapNet	ResNet50	480 × 800	28.9	33.5	42.2	34.9
	PGMapNet (Ours)	ResNet50	480 × 800	**33.4**	**33.9**	**44.9**	**37.4**
100 × 50 m	StreamMapNet	ResNet50	480 × 800	25.4	21.4	26.7	24.5
	GlobalMapNet	ResNet50	480 × 800	28.0	20.3	26.3	24.9
	PGMapNet (Ours)	ResNet50	480 × 800	**28.9**	**24.4**	**30.0**	**27.8**

Table 1 presents the performance comparison of various models on the new split of nuScenes dataset, covering both 30 m and 50 m perception ranges, where the results are derived from our implementation. Our method demonstrates considerable improvement over the baselines. Specifically, compared with StreamMapNet, our model achieves gains of 4.0% mAP and 3.3% mAP within the 30 m and 50 m ranges, respectively. In contrast, compared with GlobalMapNet, we obtain improvements of 2.5% mAP and 2.9% mAP for the two ranges in turn. Notably, our model achieves a more substantial performance gain over GlobalMapNet within the 50 m range, which demonstrates that fusing the SD Map, generating unified prior embeddings and effective sequence modeling can effectively boost the model performance.

3.3. Ablation Studies

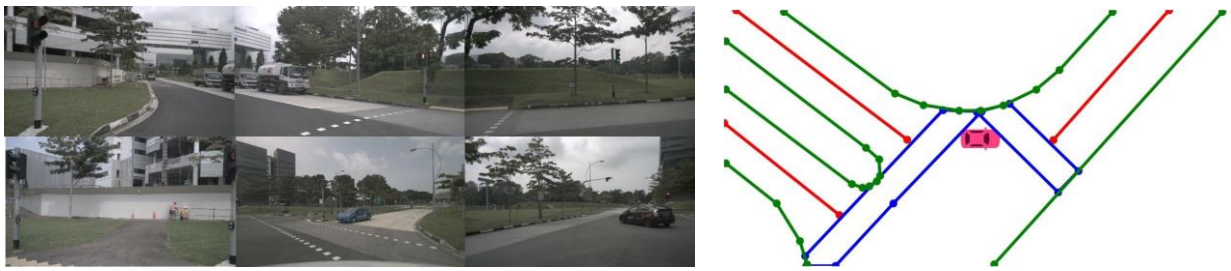
We verify the effectiveness of each proposed component via ablation experiments on the new split of nuScenes dataset within the 30 m perception range. The baseline method is built upon GlobalMapNet with the global map fusion module removed. As shown in Table 1 and Table 2, this baseline suffers a 0.1% mAP drop. The method equipped with the unified mask-guided prior embedding generator achieves remarkable performance improvements. After introducing the SD map, the mAP is boosted by 1.5%, proving that even though the SD map provides only low-

accuracy road centerline information, it can still effectively guide the model in predicting map instances after being fused via cross-attention. The performance is further enhanced when the historical predicted global map prior is additionally incorporated. All these three groups of experiments adopt the multi-point attention module proposed in StreamMapNet as the decoder, while the fourth variant further replaces it with our proposed multi-point SSM module, leading to an extra 1.0% mAP improvement. This demonstrates that the hidden state space in SSM is capable of effectively modeling the information of intra-query and inter-query sampling points, thus enabling distinct and accurate decoding of distinct map instances.

Table 2: Ablation studies of each component in PGMapNet.

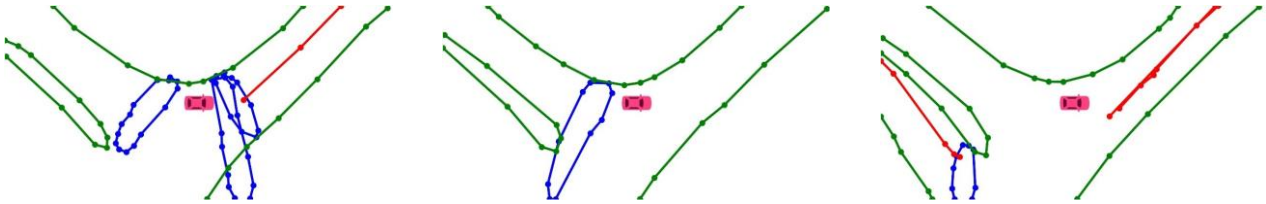
Index	Unified Mask-Guided Prior Embedding Generator		Multi-Point SSM Module	AP_{ped}	AP_{div}	AP_{bound}	mAP
	SD Map	Global Map					
1	-	-	-	29.2	32.6	42.5	34.8
2	√	-	-	29.7	34.7	44.5	36.3
3	√	√	-	31.0	**35.0**	43.3	36.4
4	√	√	√	**33.4**	33.9	**44.9**	**37.4**

3.4. Qualitative Analysis



(a) Multi-view images. (b) Ground-truth of the HD map.

Figure 2: Multi-view images and HD map ground-truth.



(a) StreamMapNet. (b) GlobalMapNet. (c) PGMapNet(Ours).

Figure 3: Mapping performance comparison of the three models under the same scenario.

As shown in Figure 2 and Figure 3, we present the comparison of mapping performance among different models in the same complex scenario within the 30 m perception range. Compared with the results of StreamMapNet in Figure 3(a) and GlobalMapNet in Figure 3(b), the proposed method in this paper achieves better detection performance for lane dividers of non-ego lanes and effectively avoids redundant detection of pedestrian crossing. In addition, our method can accurately detect the boundary of the left road, which cannot be realized by the above two methods. This result demonstrates that the fusion of multiple prior information can improve the model's detection accuracy for long-distance scene elements, especially for key elements such as road boundaries, which is of great significance for enhancing the safety of autonomous driving systems. Meanwhile, benefiting from the multi-point SSM module, the model's ability to detect and

distinguish three types of road elements is significantly improved.

4. Conclusions

In this paper, we target the problem that existing online vectorized HD map construction methods fail to uniformly and fully exploit two kinds of readily available prior maps, i.e., SD maps and historical predicted global vectorized maps. Accordingly, we propose a unified prior mask-guided end-to-end online vectorized HD map construction method, termed PGMapNet, which sufficiently fuses the two types of prior information and enhances the mapping capability. The proposed method takes BEV features as the carrier to effectively fuse the two prior maps and generate a unified mask, which is used to predict instance information and yield prior embeddings to be injected into map queries, providing key positional and structural prior information. In addition, a multi-point SSM module is designed to adaptively extract key features of point sequences and realize intra-query and inter-query interaction of point feature sequences within the SSM, thus enhancing the capability of sequence modeling. Extensive experimental results on the nuScenes dataset validate the effectiveness of our proposed method. In the future, we intend to explore the guiding effect of prior semantic information on map construction.

References

[1] Zhang, J., & Singh, S. (2014, July). LOAM: Lidar odometry and mapping in real-time. In Robotics: Science and systems (Vol. 2, No. 9, pp. 1-9).

[2] Shan, T., & Englot, B. (2018, October). Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 4758-4765). IEEE.

[3] Li, Q., Wang, Y., Wang, Y., & Zhao, H. (2022, May). Hdmapnet: An online hd map construction and evaluation framework. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 4628-4634). IEEE.

[4] Liu, Y., Yuan, T., Wang, Y., Wang, Y., & Zhao, H. (2023, July). Vectormapnet: End-to-end vectorized hd map learning. In International conference on machine learning (pp. 22352-22369). PMLR.

[5] Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., & Huang, C. (2022). Maptr: Structured modeling and learning for online vectorized hd map construction. arXiv preprint arXiv:2208.14437.

[6] Liu, X., Wang, S., Li, W., Yang, R., Chen, J., & Zhu, J. (2024). Mgmap: Mask-guided learning for online vectorized hd map construction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14812-14821).

[7] Jiang, Z., Zhu, Z., Li, P., Gao, H. A., Yuan, T., Shi, Y., ... & Zhao, H. (2024). P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. IEEE Robotics and Automation Letters, 9(10), 8539-8546.

[8] Yuan, T., Liu, Y., Wang, Y., Wang, Y., & Zhao, H. (2024). Streammapnet: Streaming mapping network for vectorized online hd map construction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 7356-7365).

[9] Shi, A., Cai, Y., Chen, X., Pu, J., Fu, Z., & Lu, H. (2024). Globalmapnet: An online framework for vectorized global hd map construction. arXiv preprint arXiv:2409.10063.

[10] Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. IEEE Pervasive computing, 7(4), 12-18.

[11] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2024). Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3), 2020-2036.

[12] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

[13] Dao, T., & Gu, A. (2024). Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060.

[14] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11621-11631).