Enhancing Vectorized Map Perception with Historical Rasterized Maps

Xiaoyu Zhang[1,*], Guangwei Liu[2,*], Zihao Liu[3], Ningyi Xu[3], Yunhui Liu[1,✉], and Ji Zhao[2,†]

[1] The Chinese University of Hong Kong
[2] Huixi Technology
[3] Shanghai Jiao Tong University
zhang.xy@link.cuhk.edu.hk xuningyi@sjtu.edu.cn
yhliu@mae.cuhk.edu.hk zhaoji84@gmail.com

Abstract. In autonomous driving, there is growing interest in end-to-end online vectorized map perception in bird's-eye-view (BEV) space, with an expectation that it could replace traditional high-cost offline high-definition (HD) maps. However, the accuracy and robustness of these methods can be easily compromised in challenging conditions, such as occlusion or adverse weather, when relying only on onboard sensors. In this paper, we propose **HRMapNet**, leveraging a low-cost **H**istorical **R**asterized **Map** to enhance online vectorized map perception. The historical rasterized map can be easily constructed from past predicted vectorized results and provides valuable complementary information. To fully exploit a historical map, we propose two novel modules to enhance BEV features and map element queries. For BEV features, we employ a feature aggregation module to encode features from both onboard images and the historical map. For map element queries, we design a query initialization module to endow queries with priors from the historical map. The two modules contribute to leveraging map information in online perception. Our HRMapNet can be integrated with most online vectorized map perception methods. We integrate it in two state-of-the-art methods, significantly improving their performance on both the nuScenes and Argoverse 2 datasets. The source code is released at `https://github.com/HXMap/HRMapNet`.

Keywords: Autonomous driving · Bird's-Eye-View · Vectorized map perception · Historical map

1 Introduction

High-definition (HD) maps comprise positions and structures of vectorized map elements (e.g., lane divider, pedestrian crossing, and road boundaries), playing a vital role in the navigation of self-driving vehicles. Traditionally, HD maps are constructed offline, utilizing SLAM-based methods [36,51] and complex pipelines

*: Equal contribution. †: Project lead. ✉: Corresponding author.

for annotation and vectorization. The high cost of constructing and maintaining an HD map severely impedes the development of autonomous driving. Consequently, researchers are turning to online map perception using onboard sensors.

The HD map used in autonomous driving is a type of *vectorized map*, which is a collection of point sets for each map element. Such vectorized representation is friendly for downstream tasks, including motion prediction [8] and planning. Some existing works treat map perception as a segmentation task [20, 52] and produce a *rasterized map*, which is a rectangular grid of pixels recording semantic labels for each position. However, a rasterized map lacks instance information and requires complex processing to be converted to the desired vectorized map.

To address the limitations above, recent work MapTR [25] defines HD map perception as a point set prediction task and utilizes DETR [4] to directly predict vectorized map elements in bird's-eye-view (BEV) space. From then, different methods [6, 32, 47, 50] are proposed to improve online vectorized map perception. This trend raises the expectation of potentially discarding offline HD maps in autonomous driving. However, relying solely on onboard sensors for online map perception poses challenges. Some challenging conditions, including adverse weather or occlusion, can significantly impact its accuracy and robustness.

In this paper, we want to underscore the crucial role of a historical map. But unlike traditional high-cost HD maps, we can alleviate the requirement of the map and maintain a low-cost one, thanks to the improving performance of online map perception. We propose HRMapNet, a novel framework designed to maintain and leverage a global historical rasterized map for vectorized map perception. Here, we choose a rasterized map to keep historical information for the following reasons. 1) Vectorized maps can be rasterized easily and efficiently [19,50]. 2) It is straightforward to merge/retrieve a local rasterized map to/from a global one. 3) A rasterized map provides clear priors of where to search for desirable map elements. 4) A rasterized map takes only a small memory footprint.

As the pipeline of HRMapNet in Fig. 1, vectorized maps from online perception are rasterized and then used to update a global historical map. For online map perception, a local rasterized map within the current perception range is retrieved and serves as complements to onboard sensors. The map updating and retrieving can be realized easily. Such pipeline can be integrated with most existing state-of-the-art (SOTA) online vectorized map perception methods.

The maintained global historical rasterized map can be set from empty and updated gradually from online perception results. When revisiting previous locations, retrieved local maps can enhance map perception by providing additional prior information. In practice, such historical maps can even be constructed and maintained collectively by a crowd of vehicles. Then, our method can be extended to facilitate crowdsourcing information for online map perception.

Existing vectorized map perception methods typically encode onboard images into BEV features and use learnable queries to decode desirable map elements. To take full advantage of historical rasterized maps within this well-established framework, we propose two novel modules to enhance both BEV features and map element queries. Specifically, we introduce a map feature aggregation mod-

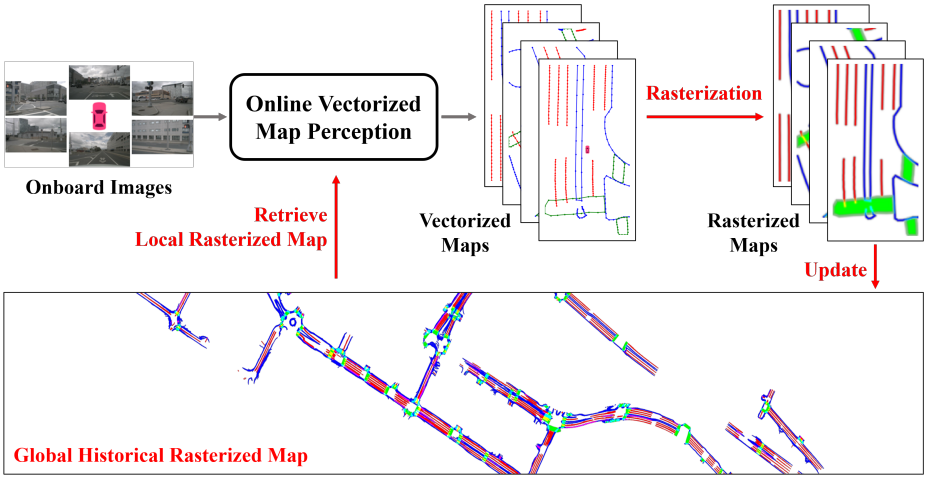

Fig. 1: Pipeline of the proposed HRMapNet. The words in red are what we design to maintain and leverage a historical map for online perception.

ule to encode features from both images and the retrieved rasterized map, which compensates for insufficient features from onboard images alone. Moreover, we encode the retrieved rasterized map into prior embeddings and design a query initialization module, in which base map element queries firstly interact with these map prior embeddings. Then, the initialized queries can search for desirable map elements more efficiently. As a result, HRMapNet utilizes both onboard images and a maintained historical rasterized map to achieve superior performance.

In summary, our main contributions are as follows:

- We propose HRMapNet, a framework leveraging historical rasterized maps for online vectorized map perception. Past predicted vectorized maps are rasterized to update a global historical rasterized map, which serves as complementary information to benefit subsequent online map perception.
- We design two modules to enhance both BEV features and learnable map element queries to take advantage of a historical map. For BEV features, we employ a BEV feature aggregation module to encode features from both images and the retrieved rasterized map. For map element queries, we design a query initialization module to search for desirable map elements efficiently. Both modules improve the online perception performance.
- We integrate HRMapNet with two SOTA methods (MapTRv2 [26] and StreamMapNet [48]), and remarkable improvements are demonstrated on both nuScenes [1] and Argoverse 2 [43] datasets under the same settings. We also provide extra results to demonstrate the robustness and potential usage for practical self-driving applications.

2 Related Work

Since map elements are commonly constructed in BEV space, online map perception benefits a lot from BEV feature learning [5, 15, 22, 23, 31], which transforms image features from surrounding cameras of a self-driving vehicle to BEV space. For example, [15, 31] lift image features to 3D space and utilize pooling to produce BEV features. [5, 22, 23] learn BEV representations in a transformer architecture. Map perception is generally compatible with all of these methods.

2.1 Map Perception with Single Frame

At early stages, map perception mainly focuses on lane detection [7, 16, 40], road topology reasoning [2, 3, 21, 24, 39] or map segmentation [10, 34, 52], which are mainly constructing rasterized maps and need post-processing to produce desired vectorized map elements for downstream applications. For example, predicted rasterized maps are clustered to acquire final vectorized maps in HDMapNet [20].

VectorMapNet [27] and MapTR [25] are pioneer works to predict vectorized map elements directly. VectorMapNet designs a map element detector and a polyline generator to produce final vectorized maps. MapTR proposes a unified permutation-equivalent modelling and utilizes the DETR [4] paradigm to predict vectorized map elements directly. Following these breakthroughs, vectorized map perception becomes popular in autonomous driving research, leading to the development of many methods for improving performance. The evolved version MapTRv2 [26] adds decoupled self-attention in the decoder and auxiliary losses, improving the accuracy largely. ScalableMap [47] exploits the structural property of map elements and designs a progressive decoder for long-range perception. MapVR [50] introduces differentiable rasterization and rending-based loss for superior sensitivity. Furthermore, instead of simple point set representation, BeMapNet [32] predicts Bézier control points and PivotNet [6] predicts pivot points instead of a fixed number of points for accurate results.

2.2 Map Perception with Complementary Information

The above methods predict map elements using a single frame, which limits further improvements. Recent advancements have expanded beyond single-frame perception, incorporating complementary information. For instance, extra standard-definition (SD) maps are explored to help HD map perception [18] and lane-topology understanding [29]. Satellite maps are used to augment onboard camera data for map perception in [9]. These methods require extra data for online perception and thus increase cost in practical applications.

Temporal information is a more accessible complement for online perception. It has been widely used in BEV feature learning [22] and object detection [41]. For vectorized map perception, StreamMapNet [48] leverages temporal information through query propagation and BEV feature fusion. SQD-MapNet [42] introduces stream query denoising to facilitate temporal consistency. In these methods, short-term previous frames in temporal are utilized for perception.

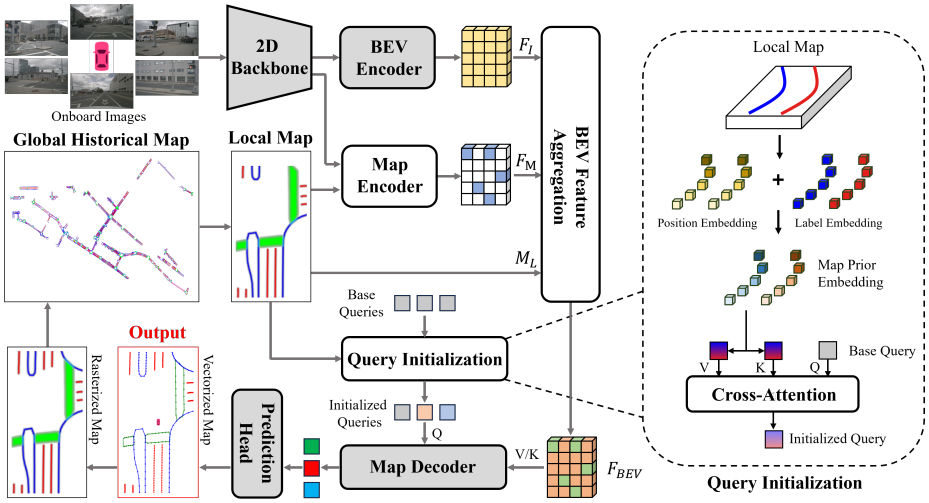

Fig. 2: Architecture of the proposed HRMapNet. The gray blocks are kept unchanged from SOTA online vectorized map perception methods.

One step further, if all temporal information is collected and utilized, a map can be constructed. In [46], a map of past LiDAR scans is utilized for object detection in autonomous driving. NMP [45] constructs a map of past BEV features for map segmentation. But BEV features take substantial memory footprint, which limits its practical usage. Take the Boston map in nuScenes dataset [1] as an example, BEV features require over 11 GB memory in NMP [45]. By contrast, we propose to maintain a low-cost historical rasterized map for vectorized map perception, which takes only 120 MB memory for the same Boston map.

3 Method

3.1 Overview

Our proposed HRMapNet is designed as a complement to SOTA online vectorized map perception methods. As illustrated in Fig. 2, HRMapNet maintains a global historical rasterized map (described in Sec. 3.2) to aid online perception. Given surrounding images as input, 2D features are extracted from a shared backbone and transformed to BEV space. We introduce a map encoder and feature aggregation module (described in Sec. 3.3) to obtain enhanced BEV features from both onboard cameras and retrieved local maps. Additionally, we also design a novel query initialization module (described in Sec. 3.4), working before the original map decoder. This module aims to endow base queries with prior information from local maps, enabling more efficient search for desirable map elements. Finally, vectorized map elements are predicted directly from the prediction head and can be rasterized to merge into the global map.

3.2 Global Rasterized Map

We firstly introduce the rasterized map used to save historical information. As illustrated in the bottom left corner of Fig. 2, vectorized maps are rasterized to update a global map. Here we adopt the rasterization method used in [19]. Briefly, the label of each pixel in the rasterized map is determined based on its distance to vectorized elements' boundary. From each online prediction at i-th timestamp, we can obtain a semantic mask, referred as local rasterized map $M_i^l \in \{0,1\}^{H\times W\times N}$, where H and W denote the spatial shape of the perception range in BEV space; N is the number of map element categories and $N = 3$ as in previous methods [25, 48], representing lane divider, pedestrian crossing and road boundary, respectively. Therefore, $M_i^l(p) \in \{0,1\}^{1\times N}$ indicates whether and what map elements exist at position $p = (x, y)$ for each category; the value 1 indicates existence and the value 0 indicates non-existence.

Such rasterized map we utilize is analogous to occupancy grid map [30], a well-established concept in robot navigation and mapping [11, 12]. Occupancy grid mapping [37] is extensively studied for updating a global map from local observations. We use a similar method to update the global rasterized map $M^g \in \mathbb{R}^{H^g\times W^g\times N}$, where H^g and W^g denote the spatial shape of the global map. For map updating, the local coordinate p of each pixel of M_i^l is firstly transformed to the global coordinate p^g based on the ego-pose $T_i = [R_i, t_i]$:

$$p^g = f_{g\leftarrow l}(p; T_i) \triangleq \text{round}(R_i p + t_i) \tag{1}$$

where R_i and t_i are relative rotation and translation, and round$(\cdot)$ denotes the rounding function which converts a continuous coordinate to a discrete coordinate in the rasterized map. Alternatively, given global coordinate p^g and pose $[R_i, t_i]$, its corresponding local coordinate p is

$$p = f_{l\leftarrow g}(p^g; T_i) \triangleq \text{round}(R_i^T(p^g - t_i)) \tag{2}$$

Then the global map M^g can be updated based on the local map M_i^l for each category. Take one category for example:

$$M^g(p^g) \leftarrow \begin{cases} M^g(p^g) + S^+ & \text{if } M_i^l(p) = 1 \\ M^g(p^g) - S^- & \text{if } M_i^l(p) = 0 \end{cases} \tag{3}$$

where p is determined by Eq. (2), M^g records the status for each map element category, S^+ and S^- are defined values to update status based on local prediction results. This simple method integrates local results into a global map efficiently, facilitating continuous refinement and updating. For online perception, a local rasterized map is retrieved from the global map based on the ego-pose T_i, and a threshold S_{th} is used to determine whether map elements exist for each category:

$$M_i^l(p) = \begin{cases} 1 & \text{if } M^g(p^g) > S_{th} \\ 0 & \text{if } M^g(p^g) \le S_{th} \end{cases} \tag{4}$$

where p^g is determined by Eq. (1).

In our implementation, the global map M^g is scaled to 8-bit unsigned int values to reduce memory consumption. As a result, a rasterized map consumes only a small memory footprint, about 1 MB per kilometer.

3.3 BEV Feature Aggregation

Various methods [15, 22, 23, 28, 31] have been proposed to transform features from perspective view to BEV space, serving as a fundamental module in map perception. For example, MapTRv2 [26] utilizes BEVPoolv2 [15] to acquire BEV features $F_I \in \mathbb{R}^{H \times W \times C}$, where C is the number of feature channels. We keep this module unchanged and introduce an aggregation module to enhance BEV features with prior information from local maps.

In HRMapNet, the retrieved local map M^l serves as priors indicating where deserves more attention for map element perception. Therefore, inspired by FB-BEV [23], we add additional BEV queries at locations where map elements exist in the local map (i.e., $M^l(p) \neq \mathbf{0}$). These additional BEV queries are projected onto images to extract relevant features through spatial cross-attention [22]. Then, additional BEV features are acquired where map elements exist in the local map. For the locations where no map elements exist (i.e., $M^l(p) = \mathbf{0}$), corresponding additional BEV features are set as zeros. These additional BEV features are formulated as $F_M \in \mathbb{R}^{H \times W \times C}$.

In the feature aggregation module, F_I, F_M and M^l are fused together:

$$F_{BEV} = \text{Conv}(\text{Concat}(F_I + F_M, M^l)) \tag{5}$$

Here, F_M is added to F_I to compensate for deficiencies. Additionally, M^l can be regarded as special BEV features with clear semantic information. Thus we concatenate it with BEV features and use convolution to acquire the enhanced BEV features $F_{BEV} \in \mathbb{R}^{H \times W \times C}$ for further processing.

3.4 Query Initialization

In addition to fusing the retrieved rasterized map into BEV features, a novel query initialization module is designed to facilitate efficient search for desirable map elements. Within a DETR [4] paradigm, a set of learnable queries will interact with extracted features to search for desirable elements. Without prior information, queries would search from random and prediction results are refined gradually through several decoder layers.

In HRMapNet, the retrieved rasterized map provides prior information about where map elements may exist and thus can facilitate map element queries search for desirable elements efficiently. As illustrated on the right of Fig. 2, the proposed query initialization works before the original map decoder. Base queries will firstly interact with prior features embedded from the local map.

In detail, for a valid location where map elements exist in the retrieved local map M^l, its position p is related to a learnable position embedding $PE(p) \in \mathbb{R}^{1 \times C}$; and the semantic vector $M^l(p)$ is projected to a label embedding $LE(p) \in$

$\mathbb{R}^{1\times C}$ using a linear projection. Then, we acquire a map prior embedding for each valid position:

$$ME(p) = PE(p) + LE(p) \tag{6}$$

Map prior embedding $ME(p) \in \mathbb{R}^{1\times C}$ encodes where map elements may exist based on the retrieved local map. To fuse priors, base queries interact with a set of map prior embeddings through cross-attention [38]. Then, initialized queries search desirable elements in BEV features through original decoder layers. Moreover, to improve efficiency and save memory consumption, the retrieved local rasterized map M^l is downsampled before extracting map prior embeddings.

3.5 Implementation Details

Training. The prediction head and training loss remain identical to SOTA vectorized map perception methods. Taking MapTRv2 [26] as an example, the prediction head predicts a class score and sequential 2D point positions for each element. Classification loss, point-to-point loss and edge direction loss are used for training; one-to-many loss [17], dense prediction loss and depth loss are used as auxiliary supervision. In each training epoch, the global map is set from empty and updated gradually from prediction results.

Testing. To ensure fair comparison, the global map is also initialized as **empty** by default and updated from prediction results. Since testing frames are typically evaluated in chronological order, most testing frames can still benefit from the updated global map from their preceding frames.

4 Experiments

To demonstrate the effectiveness, we integrate HRMapNet with two SOTA online vectorized map perception methods, MapTRv2 [26] and StreamMapNet [48].

4.1 Experimental Setup

Datasets. We evaluate HRMapNet on two commonly used self-driving datasets, nuScenes [1] and Argoverse 2 [43]. The nuScenes dataset provides 6 surrounding images captured across 1000 scenes in 4 locations, while Argoverse 2 provides 7 surrounding images captured across 1000 scenes in 6 cities. The two datasets comprise multiple traversals in both training and validation sets, providing diverse and comprehensive data for evaluation.

Metric. Following MapTRv2 and StreamMapNet, three map element categories (i.e., lane divider, pedestrian crossing and road boundary) are predicted. Chamfer distance is used to determine whether the prediction matches with the ground truth under three thresholds (0.5 m, 1.0 m, and 1.5 m). The mean average precision (mAP) is calculated for the three categories.

Table 1: Comparison on nuScenes [1]. In each block, the first row is the method as baseline and improved methods are labelled by "#". In the "Modality" column, - means using only a single frame; "SDMap" means adding an extra SDMap as input; "Temporal" means using temporal information; "HRMap" denotes using a historical rasterized map. The results of MapTR and P-MapNet are taken from P-MapNet; the results of StreamMapNet and SQD-MapNet are taken from SQD-MapNet, which are also consistent with the results reproduced by ourselves. Other results are taken from their papers. FPS is measured on a single NVIDIA A100 GPU with batch size of 1. The improvements introduced by our method are labelled in red.

Method	Modality	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP	FPS
VectorMapNet [27]	-	110	36.1	47.3	39.3	40.9	-
PivotNet [6]	-	24	56.2	56.5	60.1	57.6	-
BeMapNet [32]	-	30	57.7	62.3	59.4	59.8	-
MapTR [25]	-	24	41.2	49.5	51.1	47.3	-
# P-MapNet [18]	SDMap	24	43.7	50.9	53.5	49.4	-
StreamMapNet [48]	-	24	60.4	61.9	58.9	60.4	22.5
# SQD-MapNet [42]	Temporal	24	63.0	65.5	63.3	63.9	-
# HRMapNet	HRMap	24	**63.8**	**69.5**	**65.5**	**66.3**$^{(\uparrow 5.9)}$	21.1
MapTRv2 [26]	-	24	59.8	62.4	62.4	61.5	19.6
MapTRv2 [26]	-	110	68.1	68.3	69.7	68.7	19.6
# HRMapNet	HRMap	24	**65.8**	**67.4**	**68.5**	**67.2**$^{(\uparrow 5.7)}$	17.0
# HRMapNet	HRMap	110	**72.0**	**72.9**	**75.8**	**73.6**$^{(\uparrow 4.9)}$	17.0

Details. We keep all training and validation details the same as MapTRv2 and StreamMapNet. We use MapTRv2 as an example to elucidate the subsequent settings, more details can be found in their original papers.

Each map element is modelled as 20 sequential points. The perception range is set as [-30m, 30m] from rear to front and [-15m, 15m] from left to right, and the resolution of BEV features is 0.3m×0.3m. The shape of a local rasterized map is set the same as BEV features; the resolution of both local and global rasterized maps is also 0.3m×0.3m. We set S^+ as 30 and S^- as 1 to update the global map.

Our model is trained on 8 NVIDIA A100 GPUs with a batch size of 8×4, the learning rate is set to 6×10^{-4}. ResNet50 [14] is used as the backbone to extract image features.

4.2 Comparison with SOTA Methods

Comparison on nuScenes. As illustrated in Tab. 1, we compare HRMapNet with SOTA vectorized map perception methods. HRMapNet largely outperforms the methods using only single frame images (i.e., VectorMapNet, PivotNet, BeMapNet, MapTR, StreamMapNet and MapTRv2). Specifically, when trained with 24 epochs, HRMapNet boosts the performance of StreamMapNet and Map-

Table 2: Comparison on Argoverse 2 [43]. †: The results are re-evaluated using the official codes under the same setting with StreamMapNet.

Method	Modality	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP	FPS
StreamMapNet [48]	-	30	62.0	59.5	63.0	61.5	20.2
# SQD-MapNet [42]	Temporal	30	**64.9**	60.2	64.9	63.3	-
# HRMapNet	HRMap	30	63.8	**62.5**	**66.6**	**64.3**(↑2.8)	19.5
MapTRv2† [26]	-	30	60.0	68.7	64.2	64.3	18.1
# HRMapNet	HRMap	30	**65.1**	**71.4**	**68.6**	**68.3** (↑4.0)	16.2

Table 3: Comparisons on new split data sets proposed in StreamMapNet.

Dataset	Method	Modality	AP_{ped}	AP_{div}	AP_{bou}	mAP
nuScenes	StreamMapNet [48]	Temporal	29.6	30.1	41.9	33.9
	# HRMapNet	HRMap	**36.9**	**30.3**	**44.0**	**37.1**(↑3.2)
Argoverse 2	StreamMapNet [48]	Temporal	56.9	55.9	61.4	58.1
	# HRMapNet	HRMap	**60.1**	**58.3**	**66.0**	**61.5**(↑3.4)

TRv2 by +5.9 mAP and +5.7 mAP, respectively. When trained with 110 epochs, HRMapNet still outperforms MapTRv2 by +4.9 mAP under the same setting.

In comparison to methods introducing complementary information, HRMapNet also stands out for its comprehensive utilization of all historical information. Besides onboard images, P-MapNet adds extra SDMap from OpenStreetMap [13], but the improvement is lower than ours. SQD-MapNet is a concurrent work which leverages stream query denoising strategy to benefit from the results of previous frames. Since all predicted results are preserved in a global rasterized map, HRMapNet leverages not only temporal information but all past results for online perception, and thus achieves superior performance.

Moreover, HRMapNet maintains high efficiency in terms of inference speed when integrated with StreamMapNet and MapTRv2, running at 21.1 FPS and 17.0 FPS, respectively. This ensures HRMapNet not only enhances performance but also remains practical for real-time applications in autonomous driving.

Comparison on Argoverse 2. The results on Argoverse 2 dataset, as presented in Tab. 2, further demonstrate the effectiveness of HRMapNet. HRMapNet achieves significant enhancements across both StreamMapNet and MapTRv2, with an increase of +2.8 mAP for StreamMapNet, surpassing the performance of SQD-MapNet which leverages temporal information. Similarly, when compared to MapTRv2, HRMapNet exhibits superior performance with a gain of +4.0 mAP. These results underscore the effectiveness and versatility of our method across different methodologies and datasets.

Comparison on new split sets. The above experiments are conducted on the commonly used original dataset split, in which overlap of locations exist between

Table 4: Ablations on each component of HRMap. The improvement introduced by each component is labelled in red.

Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
MapTRv2 [26]	59.8	62.4	62.4	61.5
+ Feature Aggregation	62.6	64.5	66.8	64.6$^{(\uparrow 3.1)}$
+ Query Initialization	65.8	67.4	68.5	67.2$^{(\uparrow 2.6)}$

Table 5: Ablations on map resolution in query initialization. The default setting is highlighted in blue. Mem denotes the maximum GPU memory consumed in training.

Method	Resolution	AP_{ped}	AP_{div}	AP_{bou}	mAP	Mem. (GB)	FPS
MapTRv2	-	59.8	62.4	62.4	61.5	22.25	19.6
#HRMapNet	0.3 m	63.9	66.5	68.6	66.3	64.94	16.6
	0.6 m	**65.8**	**67.4**	68.5	**67.2**	31.41	17.0
	0.9 m	64.1	65.3	**69.6**	66.4	25.17	17.1
	1.2 m	63.2	67.2	69.4	66.6	25.22	17.1
	1.5 m	64.3	65.6	68.2	66.0	25.94	17.2

training and validation sets. StreamMapNet proposes new splitting methods for nuScenes and Argoverse 2, in which training and validation data are separated in locations. We also provide results on these new split sets in Tab. 3. On the new split data, StreamMapNet utilizes query propagation and BEV fusion to integrate temporal information. We do not use these two temporal fusion modules and only integrate utilizing a global rasterized map. HRMapNet still improves the base StreamMapNet by over +3.0 mAP on these two datasets, which reinforces the value of integrating a global rasterized map.

4.3 Ablation Study

In this subsection, we provide some ablation studies of our method, which are conducted on nuScenes with 24 epochs and use MapTRv2 [26] as the baseline.

Feature aggregation and query initialization. For online perception, our method leverages global map information in both BEV features and queries. We provide an ablation study about these two modules in Tab. 4. From MapTRv2, integrating global map information into BEV features brings an improvement of +3.1 mAP. Introducing query initialization further improves the performance by +2.6 mAP. Both components have significant positive effect for integrating global map information into online perception.

To further demonstrate that the proposed query initialization helps search for map elements more efficiently, an ablation study of decreasing decoder layers is provided in the supplementary material.

Map resolution in query initialization. In query initialization, all valid positions where map elements exist are embedded as map priors to endow in-

Table 6: Testing mAP with different localization errors. The same model is tested with varying levels of noise. The random noise is subject to a normal distribution. σ_t and σ_r are applied standard deviations for translation and rotation respectively.

		σ_t (m)			
		0	0.05	0.1	0.2
σ_r (rad)	0	**67.2**	67.1	66.9	66.2
	0.005	66.7	66.7	66.6	66.0
	0.01	66.2	66.0	66.0	65.4
	0.02	64.6	64.5	64.2	63.8

formation to base queries. However, sometimes too many prior embeddings are extracted, consuming large amounts of memory. To address this, local rasterized maps are downsampled to a coarser resolution before extracting map prior embeddings. In Tab. 5, we provide an ablation study of the resolution used to encode map prior embeddings. For the resolution of 0.3 m, downsampling is not used, resulting in a large number of embeddings and significant GPU memory consumption during training. As the resolution decreases, memory usage also decreases rapidly. The impact on inference speed is minimal. We set the resolution to 0.6 m as the default setting, because the memory consumption is acceptable and it gets the best performance. The results also indicate HRMapNet consumes about 9 GB extra GPU memory in training, compared to MapTRv2.

4.4 Extra Results for Practical Usage

Robustness to localization error. As described in Sec. 3.2, the global map is updated from local predicted rasterized maps based on ego-pose. In autonomous driving, ego-pose can be localized with high accuracy using GNSS modules or SLAM-based methods [36, 51]. To assess the robustness of HRMapNet to localization errors, we conduct additional experiments on the nuScenes dataset, as presented in Tab. 6. The model is based on MapTRv2 and trained with 24 epochs; all results are from the same model with varying levels of localization errors. We add random noises to both translation and rotation of ego-pose, and thus both map updating and retrieving would be affected.

The results clearly demonstrate the robustness of HRMapNet to localization errors, particularly in terms of translation. One contributing factor is the relatively small map resolution (0.3m) utilized in our method. Despite varying levels of localization errors, HRMapNet consistently achieves comparable results, experiencing only a 1 mAP drop in most cases. Even in the worst case with the largest noise, a historical map still brings benefits (63.8 mAP) compared to the baseline, MapTRv2 (61.5 mAP).

Considering localization with 0.1 m error for translation and 0.01 rad error for rotation is a common requirement in autonomous driving [33], these extra results indicate the effectiveness of HRMapNet in practical usage.

Table 7: Ablations on initial maps. The model used here is the same with only different initial maps. "Empty" is the default setting. "Validation Map" denotes using validation data to construct a global map for the first running. "Training Map" denotes the global map constructed during training.

Initial Map	AP_{ped}	AP_{div}	AP_{bou}	mAP
Empty	65.8	67.4	68.5	67.2
Validation Map	72.0	71.9	73.9	72.6
Training Map	81.8	85.9	83.4	83.7

Different initial maps. In the above experiments, HRMapNet is tested with an empty initial global map for a more fair comparison. The global map is updated gradually from perception results and benefits later prediction. For many frames, the online perception actually only benefits from short-term previous frames in temporal, which weakens the power of using a global historical map.

Here, we provide extra results with pre-built initial maps in Tab. 7. The model used here is the same as in Tab. 1, integrated with MapTRv2 and trained 24 epochs. Note that the model is **not re-trained or finetuned**, and we only test it with different initial maps. Here are two kinds of maps can be provided.

For the "validation map", the same model is tested with validation data twice. The first time is running with an empty initial map. The map is updated gradually as validation data comes in and the final global map is saved. This global map is loaded for the second validation. There is actually no extra data input, the model constructs a global map by itself and use it again for validation. With the help of this more complete map, the performance of the same model is further enhanced by +5.4 mAP.

Besides, the global map constructed during training is saved and loaded again for validation. As stated in Sec. 4.2, there are overlaps in location between training and validation data. Thus this training map can also benefit online perception for validation. Because this training map is more accurate, the performance is improved largely by +16.5 mAP.

We provide these extra results to show the potential of HRMapNet for practical usage in autonomous driving, including crowdsourcing online map perception. Provided with an easily maintained global map, which may even be constructed by other vehicles, the performance of online vectorized map perception can be improved largely.

4.5 Qualitative Results.

In Fig. 3, we show some qualitative comparisons in three challenging scenarios: severe occlusion, rainy day and poor lighting at night. The online map perception relying only on onboard sensors can be easily affected by these inevitable factors. Leveraging a historical rasterized map, HRMapNet helps to improve online map perception ability to handle such challenges. More visualized results and analysis are included in the supplementary material.

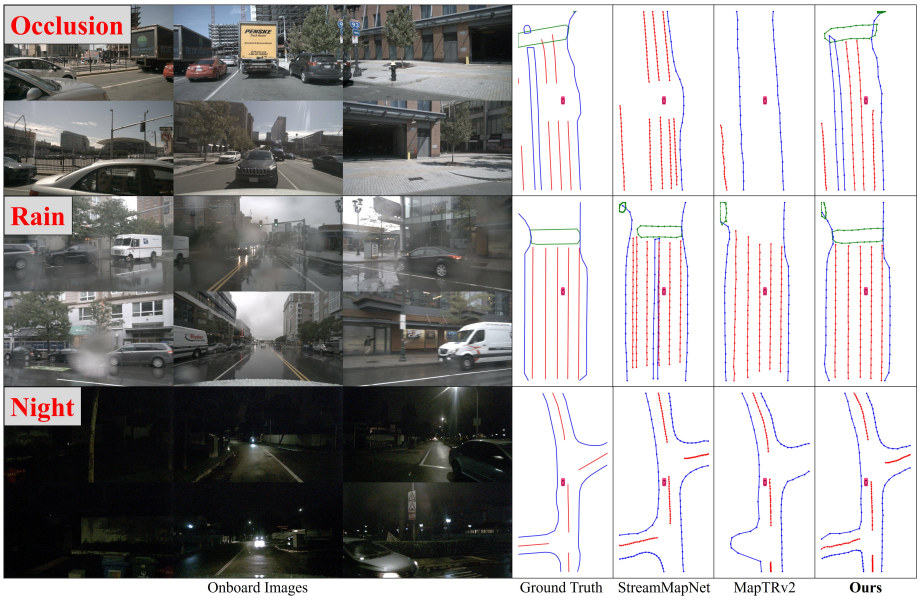

Fig. 3: Visualized results in three challenging scenarios: severe occlusion, rainy day and poor lighting at night. All three methods are trained with 24 epochs. Ours is based on MapTRv2. Lane divider, pedestrian crossing and road boundary are illustrated in red, green and blue respectively.

5 Discussion and Conclusion

In this paper, we propose to leverage historical information by maintaining a global rasterized map for improved online vectorized map perception. The global rasterized map can be constructed and maintained easily from past prediction results. We utilize such historical rasterized maps as complementary information for both BEV feature aggregation and query initialization. The proposed framework is compatible with most existing online vectorized map perception methods. It is demonstrated our proposed HRMapNet can boost two SOTA online vectorized map perception methods by a large margin. We expect HRMapNet can be a basis for crowdsourcing map perception: an accurate global rasterized map is maintained by a crowd of self-driving vehicles and serves as priors for accurate online vectorized map perception for each vehicle.

Limitations. Our proposed HRMapNet mainly focuses on how to leverage a historical rasterized map for online vectorized map perception. We do not design elaborate map maintaining methods and only use a simple yet effective one from robotic occupancy grid mapping to merge local predictions to a global map. In practice, more intelligent methods could be explored and utilized, such as [49] for collaborative semantic mapping; [35] utilizing recurrent neural networks; [44] produces consistent rasterized maps from multiple predictions.

Acknowledgements

This work is supported by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, the CUHK T Sone Robotics Institute, and the InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
2. Can, Y.B., Liniger, A., Paudel, D.P., Van Gool, L.: Structured bird's-eye-view traffic scene understanding from onboard images. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15661–15670 (2021)
3. Can, Y.B., Liniger, A., Paudel, D.P., Van Gool, L.: Topology preserving local road network estimation from single onboard camera image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17263–17272 (2022)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
5. Chen, S., Cheng, T., Wang, X., Meng, W., Zhang, Q., Liu, W.: Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. arXiv preprint arXiv:2206.04584 (2022)
6. Ding, W., Qiao, L., Qiu, X., Zhang, C.: PivotNet: Vectorized pivot learning for end-to-end HD map construction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3672–3682 (2023)
7. Feng, Z., Guo, S., Tan, X., Xu, K., Wang, M., Ma, L.: Rethinking efficient lane detection via curve modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17062–17070 (2022)
8. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
9. Gao, W., Fu, J., Jing, H., Zheng, N.: Complementing onboard sensors with satellite map: A new perspective for HD map construction. In: IEEE International Conference on Robotics and Automation (ICRA) (2024)
10. Gosala, N., Petek, K., Drews-Jr, P.L.J., Burgard, W., Valada, A.: Skyeye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14901–14910 (2023)
11. Grisetti, G., Stachniss, C., Burgard, W.: Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 2432–2437 (2005)
12. Grisetti, G., Stachniss, C., Burgard, W.: Improved techniques for grid mapping with rao-blackwellized particle filters. IEEE Transactions on Robotics **23**(1), 34–46 (2007)

13. Haklay, M., Weber, P.: Openstreetmap: User-generated street maps. IEEE Pervasive Computing **7**(4), 12–18 (2008)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Huang, J., Huang, G.: Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. arXiv preprint arXiv:2211.17111 (2022)
16. Huang, S., Shen, Z., Huang, Z., Ding, Z.h., Dai, J., Han, J., Wang, N., Liu, S.: Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17451–17460 (2023)
17. Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: DETRs with hybrid matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19702–19712 (2023)
18. Jiang, Z., Zhu, Z., Li, P., Gao, H.a., Yuan, T., Shi, Y., Zhao, H., Zhao, H.: P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. arXiv preprint arXiv:2403.10521 (2024)
19. Lazarow, J., Xu, W., Tu, Z.: Instance segmentation with mask-supervised polygonal boundary transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4382–4391 (2022)
20. Li, Q., Wang, Y., Wang, Y., Zhao, H.: HDMapNet: An online HD map construction and evaluation framework. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 4628–4634 (2022)
21. Li, T., Jia, P., Wang, B., Chen, L., JIANG, K., Yan, J., Li, H.: Lanesegnet: Map learning with lane segment perception for autonomous driving. In: International Conference on Learning Representations (ICLR) (2024)
22. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision (ECCV). pp. 1–18 (2022)
23. Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., Alvarez, J.M.: FB-BEV: BEV representation from forward-backward view transformations. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6919–6928 (2023)
24. Liao, B., Chen, S., Jiang, B., Cheng, T., Zhang, Q., Liu, W., Huang, C., Wang, X.: Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. arXiv preprint arXiv:2303.08815 (2023)
25. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: MapTR: Structured modeling and learning for online vectorized HD map construction. In: International Conference on Learning Representations (ICLR) (2023)
26. Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X.: Maptrv2: An end-to-end framework for online vectorized hd map construction. arXiv preprint arXiv:2308.05736 (2023)
27. Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Vectormapnet: End-to-end vectorized hd map learning. In: International Conference on Machine Learning (ICML). pp. 22352–22369. PMLR (2023)
28. Liu, Y., Wang, T., Zhang, X., Sun, J.: PETR: Position embedding transformation for multi-view 3D object detection. In: European Conference on Computer Vision (ECCV). pp. 531–548 (2022)
29. Luo, K.Z., Weng, X., Wang, Y., Wu, S., Li, J., Weinberger, K.Q., Wang, Y., Pavone, M.: Augmenting lane perception and topology understanding with standard definition navigation maps. arXiv preprint arXiv:2311.04079 (2023)

30. Moravec, H., Elfes, A.: High resolution maps from wide angle sonar. In: IEEE International Conference on Robotics and Automation (ICRA). vol. 2, pp. 116–121 (1985)
31. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: European Conference on Computer Vision (ECCV) (2020)
32. Qiao, Limeng and Ding, Wenjie and Qiu, Xi and Zhang, C.: End-to-end vectorized HD-map construction with piecewise bezier curve. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13218–13228 (2023)
33. Reid, T.G., Houts, S.E., Cammarata, R., Mills, G., Agarwal, S., Vora, A., Pandey, G.: Localization requirements for autonomous vehicles. arXiv preprint arXiv:1906.01061 (2019)
34. Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
35. Schreiber, M., Belagiannis, V., Gläser, C., Dietmayer, K.: Dynamic occupancy grid mapping with recurrent neural networks. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 6717–6724 (2021)
36. Shan, T., Englot, B.: LeGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4758–4765 (2018)
37. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. The MIT Press (2005)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
39. Wang, H., Li, T., Li, Y., Chen, L., Sima, C., Liu, Z., Wang, B., Jia, P., Wang, Y., Jiang, S., Others: Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2024)
40. Wang, R., Qin, J., Li, K., Li, Y., Cao, D., Xu, J.: BEV-LaneDet: An efficient 3D lane detection based on virtual camera via key-points. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1002–1011 (2023)
41. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3621–3631 (2023)
42. Wang, S., Jia, F., Liu, Y., Zhao, Y., Chen, Z., Wang, T., Zhang, C., Zhang, X., Zhao, F.: Stream query denoising for vectorized hd map construction. arXiv preprint arXiv:2401.09112 (2024)
43. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021)
44. Xie, Z., Pang, Z., Wang, Y.X.: Mv-map: Offboard hd-map generation with multi-view consistency. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8658–8668 (2023)
45. Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Neural map prior for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17535–17544 (2023)

46. You, Y., Luo, K.Z., Chen, X., Chen, J., Chao, W.L., Sun, W., Hariharan, B., Campbell, M., Weinberger, K.Q.: Hindsight is 20/20: Leveraging past traversals to aid 3D perception. In: International Conference on Learning Representations (ICLR) (2022)
47. Yu, J., Zhang, Z., Xia, S., Sang, J.: Scalablemap: Scalable map learning for online long-range vectorized hd map construction. In: The 7th Conference on Robot Learning (CoRL). pp. 2429–2443. PMLR (2023)
48. Yuan, T., Liu, Y., Wang, Y., Wang, Y., Zhao, H.: Streammapnet: Streaming mapping network for vectorized online hd map construction. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7356–7365 (2024)
49. Yue, Y., Zhao, C., Li, R., Yang, C., Zhang, J., Wen, M., Wang, Y., Wang, D.: A hierarchical framework for collaborative probabilistic semantic mapping. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 9659–9665 (2020)
50. Zhang, G., Lin, J., Wu, S., song, y., Luo, Z., Xue, Y., Lu, S., Wang, Z.: Online map vectorization for autonomous driving: A rasterization perspective. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 31865–31877 (2023)
51. Zhang, J., Singh, S.: LOAM: Lidar odometry and mapping in real-time. In: Robotics: Science and Systems (RSS). vol. 2, pp. 1–9 (2014)
52. Zhou, B., Krähenbühl, P.: Cross-view transformers for real-time map-view semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13760–13769 (2022)