Neural Map Prior for Autonomous Driving

Xuan Xiong[1]   Yicheng Liu[1]   Tianyuan Yuan[2]   Yue Wang[3]   Yilun Wang[2]   Hang Zhao[2,1*]

[1]Shanghai Qi Zhi Institute   [2]IIIS, Tsinghua University   [3]MIT

Abstract

*High-definition (HD) semantic maps are crucial in enabling autonomous vehicles to navigate urban environments. The traditional method of creating offline HD maps involves labor-intensive manual annotation processes, which are not only costly but also insufficient for timely updates. Recent studies have proposed an alternative approach that generates local maps using online sensor observations. However, this approach is limited by the sensor's perception range and its susceptibility to occlusions. In this study, we propose **Neural Map Prior (NMP)**, a neural representation of global maps. This representation automatically updates itself and improves the performance of local map inference. Specifically, we utilize two approaches to achieve this. Firstly, to integrate a strong map prior into local map inference, we apply cross-attention, a mechanism that dynamically identifies correlations between current and prior features. Secondly, to update the global neural map prior, we utilize a learning-based fusion module that guides the network in fusing features from previous traversals. Our experimental results, based on the nuScenes dataset, demonstrate that our framework is highly compatible with various map segmentation and detection architectures. It significantly improves map prediction performance, even in challenging weather conditions and situations with a longer perception range. To the best of our knowledge, this is the first learning-based system for creating a global map prior.*

1. Introduction

Autonomous vehicles require high-definition (HD) semantic maps to accurately predict the future trajectories of other agents and to navigate urban environments safely. However, the majority of these vehicles rely on costly and labor-intensive pre-annotated offline HD maps. These maps are constructed through a complex pipeline involving multiple LiDAR scanning trips with survey vehicles, global point cloud alignment, and manual annotation of map elements.

---

*Corresponding at: hangzhao@mail.tsinghua.edu.cn

Figure 1. **Comparison of semantic map construction methods.** Traditional offline semantic mapping pipelines (the first row) involve a complex manual annotation pipeline and do not support timely map updates. Online HD semantic map learning methods (the second row) rely entirely on onboard sensor observations and are susceptible to occlusions. We propose the Neural Map Prior (NMP, the third row), an innovative neural representation of global maps designed to aid onboard map prediction. NMP is incrementally updated as it continuously integrates new observations from a fleet of autonomous vehicles.

Despite the high precision of these offline mapping solutions, their scalability is constrained, and they do not support timely updates in response to changing road conditions. As a result, autonomous vehicles may operate based on outdated maps, which could compromise driving safety.

Recent research has explored alternative methods for constructing HD semantic maps using onboard sensor observations, such as camera images and LiDAR point clouds [11,13,15]. These methods typically use deep learning techniques to infer map elements in real-time, thus ad-

dressing the issue of map updates associated with offline maps. Nevertheless, the quality of these inferred maps is generally inferior when compared to pre-constructed global maps. And this quality can degrade further under unfavorable weather conditions or in occluded scenarios. The comparison of different semantic map construction methods is provided in Figure 1.

In this study, we present Neural Map Prior (NMP), a novel hybrid mapping solution that combines the best of both worlds. NMP leverages neural representations to build and update a global map prior, thereby enhancing local map inference performance. The NMP methodology consists of two important stages: **global map prior update** and **local map inference**. The global map prior is automatically developed by aggregating sensor data from a fleet of self-driving cars. Onboard sensor data and the global map prior are then integrated into the local map inference process, which subsequently refines the map prior. These procedures are interconnected in a feedback loop that grows stronger as more data are collected from vehicles traversing the roads daily. One example is shown in Figure 2.

Technically, the global NMP is defined as sparse map tiles, where each tile corresponds to a specific real-world location and starts in an empty state. For each online observation from an autonomous vehicle, a neural network encoder first extracts local bird's-eye view (BEV) features. These features are then refined using the corresponding NMP prior features, derived from the global NMP's map tile. The refined BEV feature enables us to better infer the local semantic map and update the global NMP. As the autonomous vehicles traverse through various scenes, the local map inference phase and the global map prior update step mutually reinforce each other. This iterative process results in improved quality of the predicted local semantic map and maintains a more complete and up-to-date global NMP.

We demonstrate that our NMP can be easily applied to various state-of-the-art HD semantic map learning methods, effectively enhancing their accuracy. Through experiments conducted on the nuScenes dataset, our pipeline showcases remarkable performance improvements, including a **+4.32** mIoU for HDMapNet, **+5.02** mIoU for LSS, **+5.50** mIoU for BEVFormer, and **+3.90** mAP for VectorMapNet.

To summarize, our contributions are as follows:

1. We propose a novel mapping paradigm, Neural Map Prior, which integrates the maintenance of offline global maps and the inference of online local maps. Notably, the computational and memory resources required by our approach's local map inference are comparable to previous methods.

2. We propose current-to-prior attention and Gated Recurrent Unit modules. These are adaptable to mainstream HD semantic map learning methods and effec-

tively enhance their map prediction performance.

3. We conduct a comprehensive evaluation of our method on the nuScenes dataset, considering different map elements and four map segmentation/detection architectures. The results demonstrate consistent and significant improvements. Moreover, our approach demonstrates substantial progress in challenging scenarios, such as bad weather conditions and longer perception ranges.

2. Related Works

LiDAR SLAM-Based Mapping. Autonomous driving system requires an understanding of road map elements, including lanes, pedestrian crossing, and traffic signs, to navigate the world. Such map elements are typically provided by pre-annotated High-Definition (HD) semantic maps in existing pipelines [26]. Most current HD semantic maps are manually or semi-automatically annotated on LiDAR point clouds of the environment, merged from LiDAR scans collected from survey vehicles equipped with high-end GPS and IMU. SLAM algorithms are the most commonly used algorithms to fuse LiDAR scans into a highly accurate and consistent point cloud. First, to match LiDAR data at two nearby timestamps, pairwise alignment algorithms such as ICP [1], NDT [2], and their variants [29] are employed, using either semantic [39] or geometry information [24]. Second, accurately estimating the poses of the ego vehicles is critical for building a globally consistent map and is formulated as either a non-linear least-square problem [10] or a factor graph [7]. Yang et al. [35] presented a method for reconstructing city-scale maps based on pose graph optimization under the constraint of pairwise alignment factors. To reduce the cost of manual annotation of semantic maps, Jian et al. [9] proposed several machine-learning techniques for extracting static elements from fused LiDAR point clouds and cameras. However, maintaining an HD semantic map remains a laborious and costly process due to the requirement for high precision and timely updates. In this paper, we propose using neural map priors as a novel mapping paradigm to replace human-curated HD maps, supporting timely updates to the global map prior and enhancing local map learning, potentially making it a more scalable solution for autonomous driving.

Semantic Map Learning. Semantic map learning constitutes a fundamental challenge in real-world map construction and has been formulated as a semantic segmentation problem in [18]. Various approaches have been employed to address this issue, including aerial images in [19], LiDAR point clouds in [34], and HD panoramas in [31]. To enhance fine-grained segmentation performance, crowdsourcing tags have been proposed in [32]. Recent studies have concentrated on deciphering BEV semantics from

Figure 2. **Demonstration of NMP for autonomous driving in adverse weather conditions.** Ground reflections during rainy days make online HD map predictions harder, posing safety issues for an autonomous driving system. NMP helps to make better predictions, as it incorporates prior information from other vehicles that have passed through the same area on sunny days.

onboard camera images [17, 36] and videos [4]. Relying solely on onboard sensors for model input poses a challenge, as the inputs and target map belong to different coordinate systems. Cross-view learning methodologies, such as those found in [5,11,21,23,25,27,33,40], exploit scene geometric structures to bridge the gap between sensor inputs and BEV representations. Our proposed method capitalizes on the inherent spatial properties of BEV features as a neural map prior, making it compatible with a majority of BEV semantic map learning techniques. Consequently, this approach holds the potential to enhance online map prediction capabilities.

Neural Representations. Recently, advances have been made in neural representations [8, 14, 20, 22, 28, 37]. NeuralRecon [30] presents an approach for implicit neural 3D reconstruction that integrates reconstruction and fusion processes. Unlike traditional methods that first estimate depths and subsequently perform fusion offline. Similarly, our work learns neural representation by employing the encoded image features to predict the map prior through a neural network.

3. Neural Map Prior

The aim of this work is to improve local map estimation performance by leveraging a global neural map prior. To achieve this, we propose a pipeline, depicted in Figure 3, which is specifically designed to concurrently train both the global map prior update and local map learning with integrating a fusion component. Moreover, we address the memory-intensive challenge associated with storing features of urban streets by introducing a sparse tile format for the global neural map prior, as detailed in Section 4.8.

Problem Setup. Our model operates on typical autonomous driving systems equipped with an array of onboard sensors, such as surround-view cameras and GPS/IMU, for precise localization. We assume a single-frame setting, similar to [11], which adopts a BEV encoder-decoder model for inferring local semantic maps. The BEV encoder is denoted as F_{enc}, and the decoder is denoted as F_{dec}. Additionally, we create and maintain a global neural map prior $p^g \in \mathbb{R}^{H_G \times W_G \times C}$, where H_G and W_G represent the height and width of the city, respectively. Each observation consists of input from the surrounding cameras $\mathbf{I}$ and the ego vehicle's position in the global coordinate system $\mathbf{G}_{\text{ego}} \in \mathbb{R}^{4 \times 4}$. We can transform the local coordinate of each pixel of the BEV, denoted as $l_{\text{ego}} \in \mathbb{R}^{H \times W \times 2}$ (where H and W denote the size of the BEV features), to a fixed global coordinate system using $\mathbf{G}_{\text{ego}}$. This transformation results in $p_{\text{ego}} \in \mathbb{R}^{H \times W \times 2}$. Initially, we acquire the online BEV features $o = F_{\text{enc}}(\mathbf{I}) \in \mathbb{R}^{H \times W \times C}$, where C represents the network's hidden embedding size. We then query the global prior p^g using the ego position p_{ego} to obtain the local prior BEV features $p_{t-1}^l \in \mathbb{R}^{H \times W \times C}$. A fusion function is subsequently applied to the online BEV features and the local prior BEV features to yield refined BEV features:

$$f_{\text{refine}} = F_{\text{fuse}}(o, p_{t-1}^l), f_{\text{refine}} \in \mathbb{R}^{H \times W \times C}. \quad (1)$$

Finally, the refined BEV features are decoded into the final map outputs by the decoder F_{dec}. Simultaneously, the global map prior p^g is updated using f_{refine}. The global neural network prior acts as an external memory, capable of incrementally integrating new information and simultaneously offering knowledge output. This dual functionality ultimately leads to improved local map estimation performance.

Figure 3. **The model architecture of NMP.** The top yellow box illustrates the online HD map learning process, which takes images as input and processes them through a BEV encoder and decoder to generate map segmentation results. Within the green box, customized fusion modules—comprising C2P attention and GRU—are designed to effectively integrate prior map features between the encoder and decoder, subsequently decoded to produce the final map predictions. In the bottom blue box, the model queries map tiles that overlap with the current BEV feature from storage. After the update, the neural map is returned to the previously extracted map tiles.

3.1. Local Map Learning

In order to accommodate the dynamic nature of road networks in the real world, advanced online map learning algorithms have recently been developed. These methods generate semantic map predictions based solely on data collected by onboard sensors. In contrast to earlier approaches, our proposed method incorporates neural priors to bolster accuracy. As road structures on maps are subject to change, it is imperative that recent observations take precedence over older ones. To emphasize the importance of current features, we introduce an asymmetric fusion strategy that combines current-to-prior attention and gated recurrent units.

Current-to-Prior (C2P) Cross Attention. We introduce the current-to-prior cross-attention mechanism, which employs a standard cross-attention approach [16] to operate between current and prior BEV features. Concretely, We divide each BEV feature into patches and add them with a set of learnable positional embeddings, which will be described subsequently. Current features produce queries, while prior features produce keys and values. A standard cross-attention is then applied, succeeded by a fully connected layer. Ultimately, we assemble the output queries to derive the refined BEV features, which maintain the same dimensions as the input current features. The resulting refined BEV features are expected to exhibit superior quality compared to both prior and current features.

Positional Embedding. It has been observed that the accuracy of predicted maps declines as the distance from the ego vehicle increases. To address this issue, we propose the integration of position embeddings, a set of grid-shaped learnable parameters, into the fusion module. The aim is to augment the spatial awareness of the fusion module regarding the feature positions, empowering it to learn to trust the current features closer to the ego vehicle and rely more on the prior features for distant locations. Specifically, two position embeddings are introduced: $PE_p \in \mathbb{R}^{H \times W \times C}$ for the prior features and $PE_c \in \mathbb{R}^{H \times W \times C}$ for the current features, respectively, before the fusion module F_{fuse}. Here, H and W represent the height and width of the BEV features. These embeddings provide spatial awareness to the fusion module, effectively allowing it to assimilate information from varying feature distances and locations.

3.2. Global Map Prior Update

To update the global map prior with the refined features generated by the C2P attention module, an auxiliary module is introduced, devised to attain a balanced ratio between the current and prior features. This process is illustrated in Figure 3. Intuitively, the module regulates the updating rate of the global map prior. A high update rate may lead to corruption of the global map prior due to suboptimal

local observations, while a low update rate may result in the global map prior's inability to promptly capture changes in road conditions. Therefore, we introduce a 2D convolutional variant of the Gated Recurrent Unit [6] module into NMP, serving to balance the updating and forgetting ratio. Local map prior features p_{t-1}^l, updated at $t-1$, are extracted from the global neural map prior p_{t-1}^g. The refined features generated by the C2P attention module are denoted as o'. Integrating o' with the local prior features p_{t-1}^l, the GRU yields the new prior features p_t^l at time t. Subsequently, these features are passed through the decoder to predict the local semantic map and the global neural map prior p_t^g is updated at the corresponding location by directly replacing them with p_t^l. Let z_t denote the update gate, r_t the reset gate, σ the sigmoid function, w_* the weight for 2D convolution, and $\odot$ the Hadamard product. Via the following operations, the GRU fuses o' with the prior feature p_{t-1}^l:

$$
\begin{aligned}
z_t &= \sigma(\text{Conv2D}([p_{t-1}^l, o'], w_z)) \\
r_t &= \sigma(\text{Conv2D}([p_{t-1}^l, o'], w_r)) \\
\tilde{p}_t^l &= \tanh(\text{Conv2D}([r_t \odot p_{t-1}^l, o'], w_h)) \\
p_t^l &= (1 - z_t) \odot p_{t-1}^l + z_t \odot \tilde{p}_t^l
\end{aligned}
\tag{2}
$$

Within the GRU, the update gate z_t and reset gate r_t are instrumental in determining the fusion of information from the previous traversal (*i.e.*, prior feature p_{t-1}^l) with the current BEV feature o'. Furthermore, they govern the incorporation of information from the current BEV feature o' into the global map prior feature p_t^l. GRU enables the model to better adapt to various road conditions and mapping scenarios more effectively.

4. Experiments

Datasets. We validate our NMP on the nuScenes dataset [3], a large-scale autonomous driving benchmark that includes multiple traversals with precise localization and annotated HD map semantic labels. The NuScenes dataset contains 700 scenes in the *train*, 150 in the *val*, and 150 in the *test*. Data were collected using a 32-beam Li-DAR operating at 20 Hz and six cameras offering a 360-degree field of view at 12 Hz. Annotations for keyframes are provided at 2 Hz. Each scene has a duration of 20 seconds, resulting in 28,130 and 6,019 frames for the training and validation sets, respectively.

Metric. We assess the quality of HD semantic learning using two metrics: Mean Intersection over Union (mIoU) and Mean Average Precision (mAP), as presented in HDMapNet [11]. In accordance with the methodology detailed in HDMapNet, we evaluate three static map ele-

Table 1. **Quantitative analysis of map segmentation.** The performance of online map segmentation methods and their NMP versions on the nuScenes validation set. By adding prior knowledge, NMP consistently improves these methods. (* HDMapNet remains the same as in the original work, while LSS uses the same backbone as BEVFormer.)

Model	mIoU			
	Divider	Crossing	Boundary	All
HDMapNet	41.04	16.23	40.93	32.73
HDMapNet + NMP	**44.15**	**20.95**	**46.07**	**37.05**
△ mIoU	+3.11	+4.72	+5.14	+4.32
LSS*	45.19	26.90	47.27	39.78
LSS* + NMP	**50.20**	**30.66**	**53.56**	**44.80**
△ mIoU	+5.01	+3.76	+6.29	+5.02
BEVFormer*	49.51	28.85	50.67	43.01
BEVFormer* + NMP	**55.01**	**34.09**	**56.52**	**48.54**
△ mIoU	+5.50	+5.24	+5.95	+5.53

ments:road boundary, lane divider, and pedestrian crossing on the nuScenes dataset.

4.1. Implementation Details

Base models. We primarily perform our experiments using the BEVFormer model [12] (the version excluding the temporal aspect), selected for its strength in BEV feature extraction abilities and its exceptional performance in map semantic segmentation. To validate the broad applicability of our methods, as shown in Table 1 and 2, we incorporate our NMP paradigm into four recently proposed camera-based map semantic segmentation and detection methods, which serve as our baseline models: HDMapNet [11], LSS [23], BEVFormer [12], and VectorMapNet [15]. Each of these methods implements distinct 2D–3D feature-lifting strategies: MLP-based unprojection is adopted by HDMapNet, depth-based unprojection by LSS, geometry-aware transformer-like models by BEVFormer, and homography-based unprojection by VectorMapNet. For the comparisons presented in Table 4 and Table 7, we only use the GRU fusion module.

C2P Attention. For all linear layers within the current-to-prior attention module, we set the dimension of the features to 256. For patching, we use a patch size of 10×10, corresponding to a 3m $\times$ 3m area in BEV. This setting preserves local spatial information while conserving parameters.

Global Map Resolution. We use a default map resolution of 0.3m for the rasterized neural map priors for all experiments and conduct an ablation study on the resolution in Table 7.

Table 2. **Quantitative analysis of vectorized map detection.** By adding prior knowledge, the NMP boosts the performance of VectorMapNet.

| Model | Average Precision | | | |
	$AP_{Divider}$	$AP_{Crossing}$	$AP_{Boundary}$	mAP
VectorMapNet	47.3	36.1	39.3	40.9
VectorMapNet + NMP	**49.6**	**42.9**	**41.9**	**44.8**
△ AP	+2.3	+6.8	+2.6	+3.9

Table 3. **Comparison of model performance at different BEV ranges.** As the perception range increases, the online method performance declines.; NMP significantly improves the results.

| BEV Range | + NMP | mIoU | | | |
		Divider	Crossing	Boundary	All
$60m \times 30m$	X	49.51	28.85	50.67	43.01
	✓	**55.01**	**34.09**	**56.52**	**48.54**
△ mIoU		+5.50	+5.24	+5.95	+5.53
$100m \times 100m$	X	43.41	29.07	56.57	43.01
	✓	**49.51**	**32.67**	**59.94**	**47.37**
△ mIoU		+6.10	+3.60	+3.60	+4.36
$160m \times 100m$	X	41.21	26.42	51.74	39.79
	✓	**46.85**	**29.25**	**57.22**	**44.44**
△ mIoU		+5.64	+2.83	+5.48	+4.65

Table 4. **Comparison of intra-trip fusion and inter-trip fusion.**

| Intra or Inter Trips | mIoU | | | |
	Divider	Crossing	Boundary	All
Baseline	49.51	28.85	50.67	43.01
Intra-trip fusion	51.87	30.34	53.74	45.31(+2.30)
Inter-trip fusion	**53.41**	**31.92**	**55.15**	**46.82(+3.81)**

Table 5. **Performance in adverse weather conditions.**

| Weather | + NMP | mIoU | | | |
		Divider	Crossing	Boundary	All
Rain	X	50.25	26.90	44.54	40.56
	✓	**54.64**	**30.62**	**54.19**	**46.48**
△ mIoU		+4.39	+3.72	+9.65	+5.92
Night	X	51.02	21.17	48.99	40.39
	✓	**54.66**	**33.78**	**55.92**	**48.12**
△ mIoU		+3.64	+12.61	+6.93	+7.73
NightRain	X	55.76	00.00	47.60	34.45
	✓	**61.22**	00.00	**50.84**	**37.35**
△ mIoU		+5.46	+00.00	+3.24	+2.90
Normal	X	49.27	29.49	52.11	43.62
	✓	**53.46**	**35.27**	**57.75**	**48.82**
△ mIoU		+4.19	+5.78	+5.64	+5.20

4.2. Neural Map Prior Helps Online Map Inference

In this section, we show that the effectiveness of NMP is agnostic to various model architectures and evaluation metrics. To illustrate this, we integrate NMP into the aforementioned four base models: HDMapNet, LSS, BEVFormer, and VectorMapNet. We use the same hyperparameter settings as in their original designs. During training, we freeze all the modules before the BEV features and only train the C2P attention module, the GRU, the local PE, and the decoder. For testing, all samples are arranged chronologically. As evidenced in Table 1 and Table 2, NMP consistently improves map segmentation and detection performance compared to the baseline models. Qualitative results are shown in Figure 4. These findings suggest that NMP is a generic approach that can potentially be applied to other mapping frameworks.

4.3. Neural Map Prior Helps to See Further

Conventional maps used in autonomous driving systems provide crucial information about roads extending beyond the line of sight, aiding in navigation, planning, and informed decision-making. However, the recent adoption of onboard cameras for online map prediction as an alternative approach has introduced a limitation in the prediction range. This limitation arises due to the low resolution of distant areas in the captured images. To overcome this limitation, our proposed NMP enables an extended reach for online map prediction. Specifically, the NMP leverages prior history information generated by other trips, encapsulating rich contextual details about the scenes and significantly augmenting the capabilities of online map prediction. This enhancement is demonstrated in Table 3, which consistently shows improved segmentation results compared to the baseline methods across various BEV ranges, including $60m \times 30m$, $100m \times 100m$, and $160m \times 100m$.

4.4. Inter-trip Fusion is better than Intra-trip Fusion

In Table 4, we show the effectiveness of intra-trip fusion versus inter-trip fusion for map construction. Intra-trip refers to the scenario where the fusion is limited to a single traversal, while the inter-trip fusion model uses map priors generated from multiple traversals at the same location. The findings indicate that the integration of multi-traversal prior information is more helpful for accurate map construction, highlighting the significance of using multiple traversals.

4.5. Neural Map Prior is more helpful under Adverse Weather Conditions

Autonomous vehicles face challenges when driving in bad weather conditions or low light conditions, such as rain or night driving, which may impede accurate road information identification. However, our method, NMP, captures and retains the road's appearance under optimal weather and lighting conditions, thereby equipping the vehicle with enhanced and reliable information for precise road perception during current trips. As shown in Table 5, the applica-

Table 6. **Ablation on the fusion components.** MA stands for Moving Average. Local PE stands for the positional embedding proposed in § 3.1. CA stands for the C2P Attention proposed in § 3.1 and GRU stands for gated recurrent units proposed in § 3.2.

	Component				mIoU			
Name	MA	GRU	Local PE	CA	Divider	Crossing	Boundary	All
A					49.51	28.85	50.67	43.01
B	✓				52.19(+2.68)	33.70(+4.85)	55.34(+4.67)	47.07(+4.06)
C		✓			53.22(+3.71)	31.46(+2.61)	55.93(+5.26)	46.87(+3.86)
D				✓	53.25(+3.74)	33.13(+4.28)	55.15(+4.48)	47.17(+4.16)
E		✓	✓		52.96(+3.45)	**34.13(+5.28)**	56.14(+5.47)	47.74(+4.73)
F		✓		✓	**55.05(+3.74)**	31.37(+2.52)	56.19(+5.52)	47.53(+4.52)
G		✓	✓	✓	55.01(+5.50)	34.09(+5.24)	**56.52(+5.85)**	**48.54(+5.53)**

Table 7. **Ablation on the global map resolution.** 0.3m × 0.3m is a good design choice that balances storage size and accuracy.

NMP Grid Resolution	mIoU			
	Divider	Crossing	Boundary	All
Baseline	49.51	28.85	50.67	43.01
0.3m × 0.3m	**53.22**	31.46	**55.93**	**46.87(+3.86)**
0.6m × 0.6m	52.42	**31.63**	54.74	46.26(+3.25)
1.2m × 1.2m	51.36	30.24	52.78	44.79(+1.78)

Table 8. **Performance on Boston split.** The original split contains unbalanced historical trips for the training and validation sets; Boston split is more balanced.

Data Split	+ NMP	mIoU			
		Divider	Crossing	Boundary	All
Boston Split	✗	26.35	15.32	25.06	22.24
	✓	**33.04**	**21.72**	**32.63**	**29.13**
△ mIoU		+6.69	+6.40	+7.57	+6.89
Original Split	✗	49.51	28.85	50.67	43.01
	✓	**55.01**	**34.09**	**56.52**	**48.54**
△ mIoU		+5.50	+5.24	+5.95	+5.53

tion of NMP in challenging conditions, including rain and night-time driving, leads to more substantial improvements compared to normal weather scenarios. This indicates that our perception model effectively leverages the necessary information from the NMP to handle bad weather situations. However, given the smaller sample size and the limited prior trip data available for this sample, the improvements were less significant under night-rain conditions.

4.6. Ablation Studies on Fusion Components

GRU, C2P Attention and Local Position Embedding. In this section, we evaluate the effectiveness of the components proposed in Section 3. For the sake of comparison, we introduce a simple fusion baseline, termed Moving Average (MA). In this context, the C2P Attention and GRU are replaced with a moving average fusion function. The corresponding update rule can be represented as follows:

$$p_t^l = \alpha o + (1 - \alpha) o_{t-1}^l, \tag{3}$$

where α denotes a manually searched ratio, and other notations are defined in Section 3.2. Although both GRU and MA showcase comparable performance enhancements as updating modules, GRU is preferred due to the elimination of manual parameter searches required in MA. Both GRU and CA act as effective feature fusion modules, resulting in substantial performance enhancements. The slight edge of C2P attention over GRU indicates that the transformer architecture holds a minor advantage in fusing prior feature contexts. Comparing C to E and F to G in Table 6, we observe that local PE increases the IoU of the crossing by 2.67 and 2.72, respectively. This suggests that local PE has improved feature fusion, particularly in the challenging category of pedestrian crossings. Local PE enables the model to extract additional information from the map prior, thereby complementing current observations. In comparisons of C to F and E to G in Table 6, C2P Attention increases the IoU of the lane divider by 1.83 and 2.05, respectively, highlighting its effectiveness in handling lane structures. The attention mechanism extracts relevant features based on the spatial context, leading to a more accurate understanding of dividers and boundary structures. Overall, the ablation study confirms the effectiveness of all three proposed components for feature fusion and updating.

Map Resolution. We investigate the impact of different resolutions of global neural priors on the effectiveness of online map learning in Table 7. High resolutions are preferred to preserve details on the map. However, there is a trade-off between storage and resolution. Our experiments achieved good performance with an appropriate resolution of 0.3m.

4.7. Dataset Re-split

In the original split of the nuScenes dataset, some samples lack historical traversals. We adopt an approach similar to the one presented in Hindsight [38], to re-split the trips in Boston, and name it as *Boston split*. The Boston split ensures that each sample includes at least one historical trip, while the training and test samples are geographi-

Figure 4. **Qualitative results.** From the first to the fifth row: Ground truth, HDMapNet, BEVFormer, BEVFormer with Neural Map Prior and GRU weights. We also visualize z_t, the attention map of the last step of the GRU fusion process. The model learns to selectively combine current and prior map features: specifically, when the prediction quality of the current frame is good, the network tends to learn a larger z_t, assigning more weight to the current feature; when the prediction quality of the current frame is poor, usually at intersections or locations farther away from the ego-vehicle, the network tends to learn a smaller z_t for the prior feature.

cally disjoint. To estimate the proximity of two samples, we calculate the areal overlap, specifically IoU in the bird's-eye view, between the field of view of the two traversals. This approach results in 7354 training samples and 6504 test samples. The comparison of model performance on the original split and Boston split is shown in Table 8. The improvement of NMP observed on the Boston split is greater compared to the original split.

4.8. Map Tiles

We use map tile as the storage format for our global neural map prior. In urban environments, buildings generally occupy a substantial portion of the area, whereas road-related regions account for a smaller part. To prevent the map's storage size from expanding excessively in proportion to the physical scale of the city, we design a storage structure that divides the city into sparse pieces indexed by their physical coordinates. It consumes 70% less memory space than dense tiles. Furthermore, each vehicle does not need to store the entire city map; instead, it can download map tiles on demand. The trained model remains fixed, but

these map tiles are updated, integrated, and uploaded to the cloud asynchronously. As more trip data is collected over time, the map prior becomes broader and of better quality.

5. Conclusion

In this paper, we introduce a novel system, Neural Map Prior, which is designed to enhance online learning of HD semantic maps. NMP involves the joint execution of global map prior updates and local map inference for each frame in an incremental manner. A comprehensive analysis on the nuScenes dataset demonstrates that NMP improves online map inference performance, especially in challenging weather conditions and extended prediction horizons. Future work includes learning more semantic map elements and 3D maps.

Acknowledgments

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 2

[2] Peter Biber and Wolfgang Straßer. The normal distributions transform: A new approach to laser scan matching. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 3, pages 2743–2748. IEEE, 2003. 2

[3] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5

[4] Yigit Baran Can, Alexander Liniger, Ozan Unal, Danda Paudel, and Luc Van Gool. Understanding bird's-eye view semantic hd-maps using an onboard monocular camera. *arXiv preprint arXiv:2012.03040*, 2020. 3

[5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022. 3

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5

[7] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012. 2

[8] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *CVPR*, 2020. 3

[9] Jialin Jiao. Machine learning assisted high-definition map creation. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 367–373. IEEE, 2018. 2

[10] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995. 2

[11] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: A local semantic map learning and evaluation framework. *arXiv preprint arXiv:2107.06307*, 2021. 1, 3, 5

[12] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 5

[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. *arXiv preprint arXiv:2208.14437*, 2022. 1

[14] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *NeurIPS*, 2020. 3

[15] Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. *arXiv preprint arXiv:2206.08920*, 2022. 1, 5

[16] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[17] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters*, 4(2):445–452, 2019. 3

[18] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE international conference on computer vision*, pages 1689–1697, 2015. 2

[19] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3611–3619, 2016. 2

[20] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 3

[21] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 3

[22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *CVPR*, 2019. 3

[23] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 3, 5

[24] François Pomerleau, Francis Colas, Roland Siegwart, et al. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends® in Robotics*, 4(1):1–104, 2015. 2

[25] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147, 2020. 3

[26] Guodong Rong, Byung Hyun Shin, Hadi Tabatabaee, Qiang Lu, Steve Lemke, Mārtiņš Možeiko, Eric Boise, Geehoon Uhm, Mark Gerow, Shalin Mehta, et al. Lgsvl simulator: A high fidelity simulator for autonomous driving. *arXiv preprint arXiv:2005.03778*, 2020. 2

[27] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9200–9206. IEEE, 2022. 3

[28] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned

Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*, 2019. 3

[29] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009. 2

[30] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 3

[31] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016. 2

[32] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Holistic 3d scene understanding from a single geo-tagged image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3972, 2015. 2

[33] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 3

[34] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, pages 146–155. PMLR, 2018. 2

[35] Sheng Yang, Xiaoling Zhu, Xing Nian, Lu Feng, Xiaozhi Qu, and Teng Ma. A robust pose graph approach for city scale lidar mapping. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1175–1182. IEEE, 2018. 2

[36] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15536–15545, 2021. 3

[37] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 3

[38] Yurong You, Katie Z Luo, Xiangyu Chen, Junan Chen, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Hindsight is 20/20: Leveraging past traversals to aid 3d perception. *arXiv preprint arXiv:2203.11405*, 2022. 7

[39] Fisher Yu, Jianxiong Xiao, and Thomas Funkhouser. Semantic alignment of lidar data at city scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1722–1731, 2015. 2

[40] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. *arXiv preprint arXiv:2205.02833*, 2022. 3