2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Complementing Onboard Sensors with Satellite Maps: A New Perspective for HD Map Construction

Wenjie Gao[1], Jiawei Fu[2], Yanqing Shen[1], Haodong Jing[1], Shitao Chen[1†], Nanning Zheng[1]

Abstract— **High-definition (HD) maps play a crucial role in autonomous driving systems. Recent methods have attempted to construct HD maps in real-time using vehicle onboard sensors. Due to the inherent limitations of onboard sensors, which include sensitivity to detection range and susceptibility to occlusion by nearby vehicles, the performance of these methods significantly declines in complex scenarios and long-range detection tasks. In this paper, we explore a new perspective that boosts HD map construction through the use of satellite maps to complement onboard sensors. We initially generate the satellite map tiles for each sample in nuScenes and release a complementary dataset for further research. To enable better integration of satellite maps with existing methods, we propose a hierarchical fusion module, which includes feature-level fusion and BEV-level fusion. The feature-level fusion, composed of a mask generator and a masked cross-attention mechanism, is used to refine the features from onboard sensors. The BEV-level fusion mitigates the coordinate differences between features obtained from onboard sensors and satellite maps through an alignment module. The experimental results on the augmented nuScenes showcase the seamless integration of our module into three existing HD map construction methods. The satellite maps and our proposed module notably enhance their performance in both HD map semantic segmentation and instance detection tasks. Our code will be available at *https://github.com/xjtu-cs-gao/SatforHDMap*.**

I. Introduction

As an essential component in autonomous driving systems, high-definition (HD) maps contain precise geographic information and rich semantic details of map elements such as pedestrian crossings, lane dividers, and road boundaries. This information within HD maps enables ego-vehicle to locate itself within the road network, as well as provides route and navigation information for downstream prediction and motion planning modules. Conventional manual-based offline annotation approaches confront widespread implementation challenges, primarily due to their high labor costs. Recent research [1]–[3] attempted to construct HD maps online using vehicle onboard sensor data and achieved good performance. However, we observe that these methods are affected by vehicle's surrounding environment and detection range due to the inherent limitations of onboard sensors, including weak detection for long-range objects and vulnerability to occlusion by neighboring vehicles.

Our primary insight underscores the potential augmentation of HD map construction by complementing onboard sensors with cloud-based satellite maps. Rich cloud-based information can be easily accessed for vehicles during the driving process, including roadside data, previous bird's-eye view (BEV) information [4], and satellite maps of driving area. Among this information, satellite maps provide distinctive advantages. Firstly, satellite maps have the ability to cover the driving area, thus providing long-range information. Secondly, the top-down perspective provided by satellite maps is less prone to obstruction by other vehicles and can complement the perspective view of onboard sensors. Finally, given that most existing methods convert the perspective view into BEV before calculation, the satellite map information can be seamlessly integrated with these methods due to its top-down perspective.

In this study, we explore the seamless and efficient integration of satellite maps into existing HD map construction methods, as illustrated in Fig. 1. Our initial work involves generating satellite map tiles corresponding to each sample in nuScenes [5], complementing the nuScenes dataset. Considering that nuScenes utilizes a custom coordinate system, we establish a coarse coordinate transformation equation and generate satellite map tile on each sample's pose in nuScenes. The next challenge is **how to make full use of the satellite maps**. There are coordinate deviations after coarse coordinate transformation and obstruction of satellite maps by trees in minor scenarios, hindering the integration of satellite maps. To address the issues, we propose a hierarchical fusion module comprising feature-level fusion and BEV-level fusion. The former utilizes a masked cross-attention mechanism to refine features from onboard sensors using satellite maps, in which the mask is designed to avoid interference from irrelevant information in satellite maps and reduce unnecessary interactions. The latter uses an alignment module to alleviate the impact of coordinate deviations.

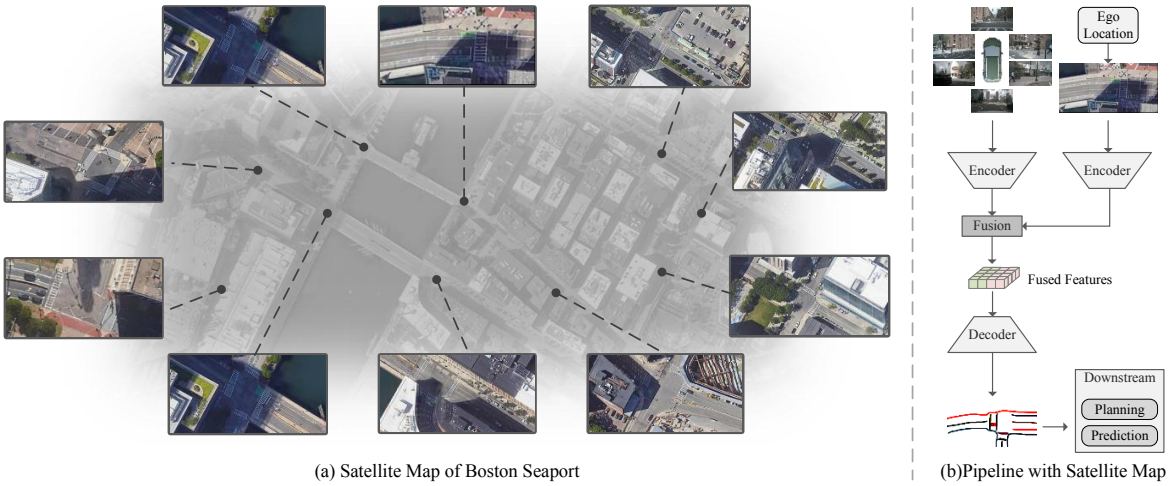
We perform comprehensive experiments to assess the performance of various fusion methods and validate the efficacy of satellite maps. The results show that even the simplest fusion method, such as concatenation directly, can significantly enhance the performance of the baseline model, indicating that satellite maps are an effective complement to onboard sensors. We further conduct experiments to integrate our proposed hierarchical fusion module into three HD

This work is supported by the National Key R&D Program of China (Grant No. 2022YFB2502900) and the National Natural Science Foundation of China (Grant No. 62088102).

[1]W. Gao, Y. Shen, H. Jing, S. Chen and N. Zheng are with National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Shaanxi 710049, P.R. China. {gaowenjie999, qing1159364090, jinghd}@stu.xjtu.edu.cn; {chenshitao, nnzheng}@mail.xjtu.edu.cn

[2]J. Fu is with The Chinese University of Hong Kong, Shatin, Hong Kong. jwfu@cse.cuhk.edu.hk

S. Chen† is with the corresponding author.



(a) Satellite Map of Boston Seaport (b)Pipeline with Satellite Map

Fig. 1. (a) Satellite maps provide comprehensive insights into the surrounding region. (b) The satellite map tile of ego location can be integrated into the current HD map construction pipeline to complement onboard sensors.

map construction methods. The results showcase remarkable performance improvements of +20.8 mIoU for HDMapNet [1], +7.9 mAP for VectorMapNet [2], and +2.3 mAP for MapTR [3] in both map semantic segmentation and instance detection tasks.

To summarize, our contributions include the following:

- We proactively explore the significance of satellite maps in HD map construction and release a complementary satellite map dataset for nuScenes, providing a new perspective for future HD map construction research.
- We propose a hierarchical fusion module to facilitate better fusion between satellite map and onboard sensor information. The feature-level fusion utilizes relevant information from satellite maps to enhance features from onboard sensors, while the BEV-level fusion mitigates the impact of coordinate offsets before concatenation.
- We integrate our module into three existing HD map construction methods and demonstrate significant improvement, particularly in long-range map construction.

II. Related Works

A. HD Map Construction

HD maps are an indispensable component in autonomous driving systems. Conventional HD maps are constructed by SLAM-based methods [6]–[9]. Recently, some methods [10]–[14] perform segmentation in a rasterized BEV space to obtain drivable areas or map elements. This representation, however, lacks instance-specific information and is proved incompatible with downstream modules [15]. Instead, recent works [1]–[4], [16]–[19] represent map elements as instances composed of individual points. To construct vectorized HD maps, HDMapNet [1] groups pixel-wise segmentation results with heuristic post-processing, which requires a significant amount of computations. VectorMapNet [2], InstaGraM [17] and MapTR [3] achieve end-to-end map element detection. The aforementioned methods rely exclusively on data from onboard sensors as inputs. A similar study NMP [4] proposes a new paradigm by acquiring historical BEV information of the ego-vehicle or other vehicles from cloud or local storage as prior knowledge. In contrast, satellite maps provide heightened accessibility and our proposed fusion module demonstrates superior performance in fusion.

B. Multi-Sensor Fusion

Multi-sensor fusion has been a prominent research area in the field of autonomous driving. Presently, prevalent sensor fusion methods can be categorized into Transformer-based methods [13], [20]–[25] and concatenation-based methods [13], [26]–[28]. The fundamental paradigm of Transformer-based methods involves mapping two types of features into a shared feature space, constructing queries, keys, and values, and subsequently utilizing attention mechanisms to facilitate fusion. Transfuser [22] concatenates LiDAR features and camera features, leveraging standard self-attention modules for fusion. AutoAlignV2 [23] introduces a multi-layer deformable cross-attention network to aggregate features from distinct modalities. The primary challenge encountered by concatenation-based methods pertains to the alignment of data originating from disparate sensors. Recent BEVFusion works [26], [27] use fully convolution layers with few residual blocks to compensate for such localized misalignments. Nevertheless, applying the above-mentioned methods to integrate satellite map into existing approaches directly yields suboptimal results. Therefore, we combine the unique attributes of satellite map and drew inspiration from these methods to design a hierarchical fusion module.

III. Complementary Dataset for nuScenes

The nuScenes [5] dataset is widely used in the field of autonomous driving, and most existing methods of HD map construction have been validated on it. It encompasses the entire suite of sensors for autonomous vehicles, including 6 cameras, 5 radars, 1 LiDAR, and GPS & IMU, and is composed of 1000 scenes lasting 20s duration, collected across four districts. However, nuScenes only incorporates

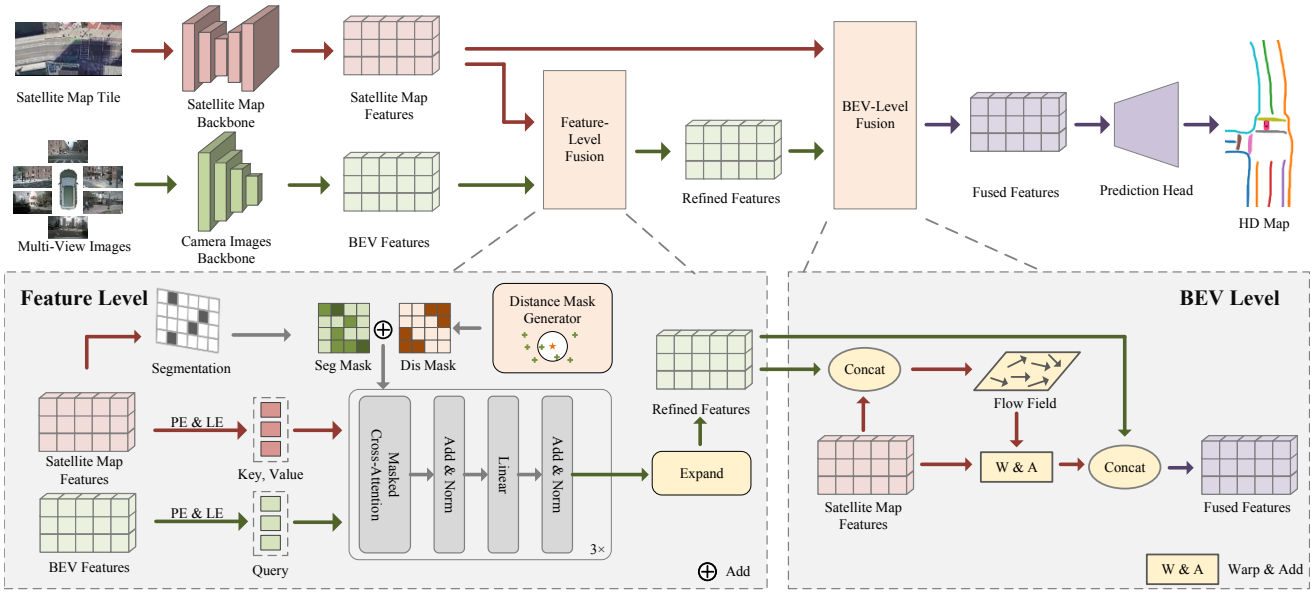


Fig. 2. Framework overview. PE stands for patch embedding and position embedding, LE stands for linear embedding. The green arrows represent the information flow from onboard sensors. The red arrows represent the information flow from satellite maps. Our framework utilizes two branches to extract features from multi-view images and satellite map tiles, respectively. A hierarchical fusion module, comprising feature-level fusion and BEV-level fusion, is designed to fuse the two features. The final task head is used to generate the HD maps from the fused features.

information from onboard sensors. Therefore, we obtain satellite maps of the four districts as a complementary dataset. The details are shown in Table I.

TABLE I

DETAILS OF COMPLEMENT DATASET FOR FOUR DISTRICTS. RESOLUTION STANDS FOR THE RESOLUTION OF SATELLITE MAP TILES.

District	Area Size	Samples	Resolution
Boston Seaport	$(2200m, 3500m)$	22266	(137, 273)
Singapore's One North	$(1700m, 2000m)$	8143	(102, 202)
Queenstown	$(3100m, 3600m)$	6604	(102, 202)
Holland Village	$(2600m, 2600m)$	3548	(102, 202)

Here we describe the process of generating satellite map tiles for each sample. Initially, the highest resolution satellite maps of four districts in nuScenes are downloaded from Google Maps. Next, we utilize a keypoint alignment method that selects the coordinates of five landmarks for alignment, establishing the coarse transformation equation between the nuScenes coordinate system and the satellite map coordinate system. After coarse alignment, the coordinate deviation is within $2m$. Following this, the corresponding satellite map region is acquired based on the position and orientation of each sample. The satellite maps are then sliced into the size of $(30m, 60m)$, which matches the configuration of most HD map construction methods. Ultimately, we retrieve the tiles of these regions using the bilinear interpolation method.

We release the satellite map tiles as a complementary dataset for nuScenes, which is available at *https://www.kaggle.com/datasets/wjgao0101/satfornuscenes*.

IV. SATELLITE MAP FUSION FRAMEWORK

In Section III, we present a complementary dataset of satellite maps for nuScenes. In this section, we introduce a framework based on a hierarchical fusion module for integrating the satellite maps into existing HD map construction methods, as shown in Fig. 2.

A. Overall Architecture

The general paradigm for HD map construction methods follows an encoder-decoder architecture, in which the encoder transforms data from onboard sensors, such as camera images, into BEV features and the decoder constructs HD maps from the extracted BEV features.

Within our framework, we utilize two distinct branches to extract features from camera images and satellite maps, respectively. The camera images branch follows HDMapNet, VectorMapNet, and MapTR [1]–[3] to take the camera images input $\mathcal{I}_{img} \in \mathbb{R}^{6\times H_i \times W_i \times 3}$ from onboard sensors. Various methods, including MLP, Inverse Perspective Mapping (IPM) [29] can be utilized to transform the images from perspective view to BEV view to obtain the BEV features $\mathcal{F}_{bev} \in \mathbb{R}^{H\times W\times C}$, where $H = 100$, $W = 200$, and $C = 64$. For the satellite map tile input $\mathcal{I}_{sat}$ with resized shape of $H \times W \times 3$, the satellite map branch adopts U-Net [30] architecture with a ResNet18 [31] backbone to extract features $\mathcal{F}_{sat} \in \mathbb{R}^{H\times W\times C}$. The architecture is designed to maintain the structural integrity of the image and avoid potential distortions. Furthermore, we design a hierarchical fusion module to integrate $\mathcal{F}_{sat}$ and $\mathcal{F}_{bev}$. In the feature-level fusion stage, we decode a BEV-satellite attention mask derived from distance mask and segmentation mask, which is generated from $\mathcal{F}_{sat}$. Subsequently, a masked cross-attention is used to refine $\mathcal{F}_{bev}$ and produce refined features $\mathcal{F}_{ref}$. The BEV-level fusion introduces a BEV alignment module to ensure the alignment of $\mathcal{F}_{ref}$ and $\mathcal{F}_{sat}$ before concatenation and obtain the final fused features $\mathcal{F}_{fus}$. Lastly, a task-specific head is utilized for different tasks, including map semantic segmentation and instance detection.

B. Feature-Level Fusion

The feature-level fusion is designed to leverage information from satellite maps to enhance the features extracted from onboard sensors pixel by pixel. Given that the satellite maps cannot be updated in real-time, it is essential to follow the rule of prioritizing BEV features during the feature-level fusion. Therefore, we design position embeddings to modify the fusion weights of different positions and use masks to avoid interference from irrelevant information in satellite maps.

Concretely, the inputs of the fusion module contain $\mathscr{F}_{bev}$ from onboard sensors and $\mathscr{F}_{sat}$ from satellite maps. We perform patch embedding on the features using patches with the size of (5, 5) and enhance them with learnable positional embeddings. Linear projection is used to learn BEV features as queries $Q \in \mathbb{R}^{N \times C_h}$ and satellite map features as keys and values $K, V \in \mathbb{R}^{M \times C_h}$, where $N = 800$, $M = 800$, and $C_h = 256$. We decode an attention mask $\mathscr{M}$, which is generated based on the segmentation of the satellite maps and distance information. Subsequently, the BEV features from onboard sensors are refined with the satellite map features by a masked cross-attention mechanism. Finally, the output is expanded and transformed to the BEV feature space to obtain the refined BEV features $\mathscr{F}_{ref}$.

We detail the position embeddings, mask generator, and masked cross-attention as follows.

Position Embedding. Vehicle onboard sensors are influenced by detection distances, whereas satellite maps are less susceptible to such influences. To achieve a balanced fusion process, we introduce position embeddings that emphasize the role of onboard sensors at close distances and satellite maps at longer distances. Specifically, we introduce two learnable parameters $PE_{bev} \in \mathbb{R}^{H_{grid} \times W_{grid} \times C}$ for BEV features from onboard sensor and $PE_{sat} \in \mathbb{R}^{H_{grid} \times W_{grid} \times C}$ for features from satellite maps, respectively. Here, H_{grid} and W_{grid} represent the height and width of patched BEV features.

Mask Generator. Satellite maps may introduce irrelevant or erroneous information when they are obstructed or not updated in real-time, which can potentially interfere with BEV features from onboard sensors. Additionally, interactions from long distances have proved ineffective [21]. To avoid unnecessary interactions between feature pairs, we introduce a BEV-Satellite attention mask $\mathscr{M} \in \mathbb{R}^{N \times M}$ that comprises distance mask and segmentation mask.

The distance mask is defined as,

$$\mathscr{M}_{dis}(x,y) = \begin{cases} -inf & \text{if Eul}(x,y) > D \\ 0 & \text{else} \end{cases}, \tag{1}$$

where $\text{Eul}(*)$ represents the Euclidean distance function, and $D = 5$ represents the threshold of distance.

Regarding the segmentation mask, we utilize a full convolution layer to generate the semantic segmentation information $\mathscr{S} \in \mathbb{R}^{H \times W}$ from $\mathscr{F}_{sat}$. Following this, we use patch embedding and linear embedding similar to BEV features to transform it into $\mathscr{E}_{seg} \in \mathbb{R}^{M \times 1}$ and expanded its dimensions to obtain the mask $\mathscr{M}_{seg} \in \mathbb{R}^{M \times N}$. The resultant mask $\mathscr{M}$ is defined as,

$$\mathscr{M} = \mathscr{M}_{seg}^{T} + \mathscr{M}_{dis}, \tag{2}$$

Masked Cross-Attention. Three masked cross-attention modules are cascaded to learn the associations between the BEV features from onboard sensors and features from satellite maps. We use the BEV features after linear embedding as queries Q and the satellite map features as keys K and values V, where $Q \in \mathbb{R}^{N \times C_h}$ and $K, V \in \mathbb{R}^{M \times C_h}$. Following [32], the masked attention modulates the attention matrix via,

$$X = \text{softmax}(\mathscr{M} + QK^{T})V + Q, \tag{3}$$

Finally, a feed-forward network (FFN) is used to calculate the refined features Q_{out}, which are of the same shape as the initial queries Q,

$$Q_{out} = \text{FFN}(X) + X, \tag{4}$$

C. BEV-Level Fusion

Although the coordinate of satellite map has been coarsely calibrated during dataset generation, there may still be some discrepancies between the satellite maps and the actually generated BEV features considering the localization errors. Therefore, it is not appropriate to concaten the refined BEV feature $\mathscr{F}_{ref}$ from vehicle sensors and features $\mathscr{F}_{sat}$ from satellite maps directly in BEV-level. To better fuse the two types of features, we perform an alignment operation before concatenation like the flow model [33], [34]. Specifically, $\mathscr{F}_{ref}$ and $\mathscr{F}_{sat}$ are concatenated together and passed through several convolutional layers to predict the coordinate offsets $\Delta \in \mathbb{R}^{H \times W \times 2}$ in each position. The warp operation is utilized to obtain aligned satellite map features $\tilde{\mathscr{F}}_{sat}$ by the bilinear interpolation kernel as,

$$\begin{aligned} \tilde{\mathscr{F}}_{sat}(h,w) = \sum_{h'}^{H} \sum_{w'}^{W} \mathscr{F}_{h'w'} \cdot \max(0, 1 - \left|h + \Delta_{hw1} - h'\right|) \\ \cdot \max(0, 1 - \left|w + \Delta_{hw2} - w'\right|), \end{aligned} \tag{5}$$

where Δ_{hw1}, Δ_{hw2} represent the learned 2D transformation offsets for position (h,w). The addition operation involves adding $\mathscr{F}_{sat}$ to $\tilde{\mathscr{F}}_{sat}$. We then concatenate $\tilde{\mathscr{F}}_{sat}$ and $\mathscr{F}_{ref}$ to obtain the final fused features $\mathscr{F}_{fus}$. Finally, we utilize the task-specific head to construct the HD maps.

V. Experiments

A. Dataset & Metrics

We evaluate our module on nuScenes and satellite map complementary dataset. We focus on two sub-tasks in HD map construction: map semantic segmentation and instance detection. We utilize HDMapNet as a baseline for the map semantic segmentation task and evaluate the performance using Mean Intersection over Union (mIoU). For the map instance detection task, we use MapTR and VectorMapNet as the baselines and evaluate the performance using Mean Average Precision (mAP).

B. Baseline Models

To validate the broad applicability of our methods, we incorporate our fusion module into three recently proposed camera-based HD map construction methods, which serve as our baseline methods as follows:

HDMapNet [1] introduces the map learning problem. It utilizes the MLP-based projection method to extract BEV features and involves a post-processing step to generate vectorized map elements. Apart from comparative experiments, we perform long-range construction experiments and ablation studies using HDMapNet as a baseline.

VectorMapNet [2] is the pioneering work in end-to-end map instance detection. It adopts two Transformers to predict key points and generate map elements, respectively.

MapTR [3] proposes a novel modeling approach for map elements and achieves the current state-of-the-art performance.

C. Comparison with Baselines

Effectiveness of satellite maps. We integrate the satellite map fusion module into three distinct baselines, each possessing varying network architectures, perspective-to-BEV projection methods, and construction tasks. Table II and Table III show the comparisons. In the map segmentation task, the satellite map fusion module achieves 20.8 higher IoU compared to HDMapNet. In the map instance detection task, it also achieves 7.9 higher mAP compared to VectorMapNet. Since MapTR implicitly represents BEV features, we perform patch embedding on the satellite map features and then utilize deformable attention for feature-level fusion. Surprisingly, this approach still achieves 2.3 higher mAP. These findings indicate that our proposed satellite map fusion module is a general approach that can potentially be applied to other HD map construction frameworks.

Influence of detection range. The detection range of HD maps greatly influences downstream planning and decision-making modules. However, due to the inherent limitations of onboard sensors, the performance of models declines significantly as the detection range increases. Compared to onboard sensors, satellite maps can provide information from long-range detection. We set various BEV ranges for comparison, including $60m \times 30m$, $60m \times 60m$, and $120m \times 60m$. The enhancement is shown in Table IV. After integrating the satellite map fusion module, we observe that as the distance increases, the rate of performance degradation of the model significantly decreases.

D. Ablation Studies

Feature-level fusion module. In Table V, we conduct ablation studies on different feature fusion methods, which include concatenation directly, standard attention [35], deformable attention [36], shift window attention [37], and masked attention(ours). We observed that direct concatenation leads to a performance improvement, demonstrating that satellite maps can effectively complement onboard sensors in HD map construction. On the other hand, attention-based methods show further enhancement in fusion performance. Notably, our proposed distance and segmentation mask strategies can achieve 1.6 higher mIoU compared to standard attention. This improvement can be attributed to the mask's ability to reduce unnecessary interactions, particularly in minor scenarios where satellite maps contain errors or irrelevant information.

BEV-level fusion module. Ablations on the BEV-level fusion are presented in Table VI. The experimental results indicate that solely performing coarse alignment during satellite map slicing leads to a decrease in the model's performance due to coordinate deviations. By introducing an alignment module before concatenation, the impact of coordinate deviations can be effectively alleviated, resulting in an improvement of 1.0 mIoU.

TABLE II

IOU SCORES (%) OF MAP SEMANTIC SEGMENTATION ON THE NUSCENES VALIDATION SET. SFM STANDS FOR THE SATELLITE MAP FUSION MODULE. BY ADDING SATELLITE MAP INFORMATION, THE FUSION MODULE CAN IMPROVE THE PERFORMANCE OF HDMAPNET(HDMAPNET REMAINS THE SAME AS IN THE ORIGINAL WORK).

Baseline	+ SFM	IoU score (%)			
		Divider	Crossing	Boundary	All
HDMapNet	-	40.6	18.7	39.5	32.9
	✓	54.9	53.4	52.9	53.7
ΔmIoU		**+14.3**	**+34.7**	**+13.4**	**+20.8**

TABLE III

MAP OF MAP INSTANCE DETECTION ON THE NUSCENES VALIDATION SET. BY ADDING SATELLITE MAP INFORMATION, THE FUSION MODULE BOOSTS THE PERFORMANCE OF BOTH VECTORMAPNET AND MAPTR(VECTORMAPNET AND MAPTR REMAIN THE SAME AS IN THE ORIGINAL WORK).

Baseline	+ SFM	mAP (%)			
		Divider	Crossing	Boundary	All
VectorMapNet	-	47.3	36.1	39.3	40.9
	✓	51.9	50.2	44.2	48.8
ΔmAP		**+4.6**	**+14.1**	**+4.9**	**+7.9**
MapTR	-	51.5	46.3	53.1	50.3
	✓	55.3	47.2	55.3	52.6
ΔmAP		**+3.8**	**+0.9**	**+2.2**	**+2.3**

TABLE IV

COMPARISON OF MODEL PERFORMANCE USING HDMAPNET AS THE BASELINE AT DIFFERENT BEV RANGES.

BEV Range	+ SFM	IoU score (%)			
		Divider	Crossing	Boundary	All
$60m \times 30m$	-	40.6	18.7	39.5	32.9
	✓	54.9	53.4	52.9	53.7
ΔmIoU		**+14.3**	**+34.7**	**+13.4**	**+20.8**
$60m \times 60m$	-	33.6	15.8	32.2	27.2
	✓	51.6	52.1	49.1	50.9
ΔmIoU		**+18.0**	**+36.3**	**+16.9**	**+23.7**
$120m \times 60m$	-	26.9	12.9	25.7	21.8
	✓	51.0	53.0	45.2	49.7
ΔmIoU		**+24.1**	**+40.1**	**+19.5**	**+27.9**

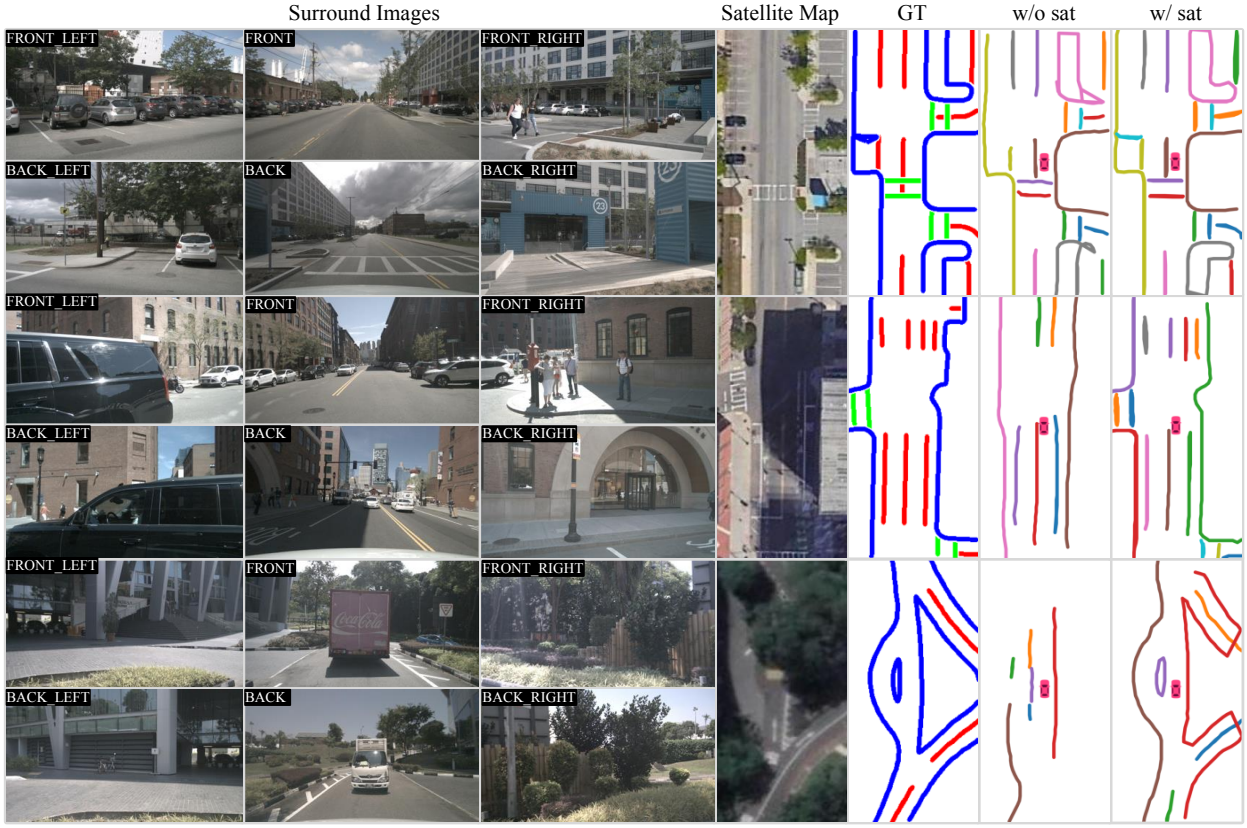


Fig. 3. Qualitative results of our method, where sat stands for satellite maps and GT stands for ground truth. After incorporating satellite maps, the model's performance is significantly improved in both complex scenarios and situations of occlusion by other vehicles. Moreover, the model exhibits stable enhancement in areas not covered by satellite maps.

TABLE V

ABLATION ON THE FEATURE-LEVEL FUSION MODULE. CONCAT STANDS FOR CONCATENATION. SA STANDS FOR STANDARD ATTENTION [35]. DA STANDS FOR DEFORMABLE ATTENTION [36]. SWA STANDS FOR SHIFT WINDOW ATTENTION [37]. MA(D) STANDS FOR DISTANCE MASK ATTENTION. MA(D+S) STANDS FOR DISTANCE AND SEGMENTATION MASK ATTENTION.

Method	IoU score (%)			
	Divider	Crossing	Boundary	All
baseline	40.6	18.7	39.5	32.9
concat	53.6(+13.0)	45.2(+26.5)	46.0(+6.5)	48.3(+15.4)
SA	51.5(+10.9)	**51.5(+32.8)**	50.2(+10.7)	51.1(+18.2)
DA	53.4(+12.8)	50.3(+31.6)	51.5(+12.0)	51.7(+18.8)
SWA	54.0(+13.4)	49.6(+30.9)	52.0(+12.5)	51.9(+19.0)
MA(d)	53.4(+12.8)	50.2(+31.5)	51.0(+11.5)	51.5(+18.6)
MA(d+s)	**54.5(+13.9)**	**51.5(+32.8)**	**52.2(+12.7)**	**52.7(+19.8)**

E. Qualitative Visualization

We visualize the HD map construction in Fig. 3. The first scenario illustrates that the satellite map can provide additional information and enhance the model capability in complex scenes, such as intersections. In the second scenario, the vehicle's left-side field of view is obstructed by another vehicle, resulting in the omission of an intersection. The satellite map capitalizes on its long-range data provision capability to aid in intersection detection. In the final scenario, where the satellite map is obstructed, our proposed method still consistently improves the model's performance, and this can be attributed to the mask's ability to avoid interference from irrelevant information.

TABLE VI

ABLATION ON THE BEV-LEVEL FUSION. WITHOUT ALIGNMENT STANDS FOR CONCATENATION DIRECTLY IN BEV LEVEL.

Method	IoU score (%)			
	Divider	Crossing	Boundary	All
baseline	40.6	18.7	39.5	32.9
w/o align	54.5(+13.9)	51.5(+32.8)	52.2(+12.7)	52.7(+19.8)
alignment	**54.9(+14.3)**	**53.4(+34.7)**	**52.9(+13.4)**	**53.7(+20.8)**

VI. CONCLUSION

In this paper, we explore the use of cloud-based satellite maps to complement onboard sensors for boosting HD map construction. We generate the corresponding satellite map tiles for each sample in nuScenes and release them as a complementary dataset. To integrate the satellite maps into existing methods, we propose a hierarchical fusion module that enhances features obtained from onboard sensors at the feature level and aligns features at the BEV level. Comprehensive experiments demonstrate that satellite map information and our proposed method can enhance the performance of existing models, particularly in long-range map construction. We believe that combining cloud-based data such as satellite maps with onboard sensor data will offer a novel perspective to future HD map construction tasks.

REFERENCES

[1] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.

[2] Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," *arXiv preprint arXiv:2206.08920*, 2022.

[3] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction," *arXiv preprint arXiv:2208.14437*, 2022.

[4] X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Neural map prior for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 535–17 544.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 621–11 631.

[6] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2. Berkeley, CA, 2014, pp. 1–9.

[7] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5135–5142.

[8] T. Shan and B. Englot, "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)*. IEEE, 2018, pp. 4758–4765.

[9] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[10] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 5935–5943.

[11] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[12] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[13] H. Dong, X. Zhang, X. Jiang, J. Zhang, J. Xu, R. Ai, W. Gu, H. Lu, J. Kannala, and X. Chen, "SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation and Prediction," *arXiv preprint arXiv:2211.15656*, 2022.

[14] S. Wang, W. Li, W. Liu, X. Liu, and J. Zhu, "LiDAR2Map: In defense of LiDAR-Based semantic map construction using online camera distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5186–5195.

[15] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.

[16] L. Qiao, W. Ding, X. Qiu, and C. Zhang, "End-to-end vectorized HD-Map construction with piecewise bezier curve," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 218–13 228.

[17] J. Shin, F. Rameau, H. Jeong, and D. Kum, "InstaGraM: Instance-level Graph Modeling for Vectorized HD Map Learning," *arXiv preprint arXiv:2301.04470*, 2023.

[18] Z. Xie, Z. Pang, and Y. Wang, "MV-Map: Offboard HD-Map generation with multi-view consistency," *arXiv preprint arXiv:2305.08851*, 2023.

[19] G. Zhang, J. Lin, S. Wu, Y. Song, Z. Luo, Y. Xue, S. Lu, and Z. Wang, "Online map vectorization for autonomous driving: A rasterization perspective," *arXiv preprint arXiv:2306.10502*, 2023.

[20] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[21] S. Pang, D. Morris, and H. Radha, "TransCAR: Transformer-based camera-and-radar fusion for 3D object detection," *arXiv preprint arXiv:2305.00397*, 2023.

[22] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Deformable feature aggregation for dynamic multi-modal 3D object detection," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 628–644.

[24] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 908–917.

[25] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[26] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[27] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

[28] G. P. Meyer, J. Charland, D. Hegde, A. Laddha, and C. Vallespi-Gonzalez, "Sensor fusion for joint 3d object detection and semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[29] H. A. Mallot, H. H. Bülthoff, JJ. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[32] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," 2022.

[33] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, and H. Shi, "AlignSeg: Feature-Aligned Segmentation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 550–557, 2022.

[34] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, and Y. Tong, "Semantic flow for fast and accurate scene parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 775–793.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021.