Mask2Map: Vectorized HD Map Construction Using Bird's Eye View Segmentation Masks

Sehwan Choi[1*], Jungho Kim[1*], Hongjae Shin[1], and Jun Won Choi[2**]

[1] Hanyang University, Korea
{sehwanchoi, junghokim, hjshin}@spa.hanyang.ac.kr
[2] Seoul National University, Korea
junwchoi@snu.ac.kr

Abstract. In this paper, we introduce Mask2Map, a novel end-to-end online HD map construction method designed for autonomous driving applications. Our approach focuses on predicting the class and ordered point set of map instances within a scene, represented in the bird's eye view (BEV). Mask2Map consists of two primary components: the *Instance-Level Mask Prediction Network* (IMPNet) and the *Mask-Driven Map Prediction Network* (MMPNet). IMPNet generates Mask-Aware Queries and BEV Segmentation Masks to capture comprehensive semantic information globally. Subsequently, MMPNet enhances these query features using local contextual information through two submodules: the *Positional Query Generator* (PQG) and the *Geometric Feature Extractor* (GFE). PQG extracts instance-level positional queries by embedding BEV positional information into Mask-Aware Queries, while GFE utilizes BEV Segmentation Masks to generate point-level geometric features. However, we observed limited performance in Mask2Map due to inter-network inconsistency stemming from different predictions to Ground Truth (GT) matching between IMPNet and MMPNet. To tackle this challenge, we propose the *Inter-network Denoising Training* method, which guides the model to denoise the output affected by both noisy GT queries and perturbed GT Segmentation Masks. Our evaluation conducted on nuScenes and Argoverse2 benchmarks demonstrates that Mask2Map achieves remarkable performance improvements over previous state-of-the-art methods, with gains of 10.1% mAP and 4.1% mAP, respectively. Our code can be found at https://github.com/SehwanChoi0307/Mask2Map

Keywords: Online HD Map Construction · Instance Segmentation · Vectorized Representation · Denoising Training · Autonomous Driving

1 Introduction

High-definition (HD) maps are considered pivotal elements in ensuring safe and effective navigation for autonomous vehicles [1,8,10,11,19]. They facilitate pre-

[*] Equal contribution.
[**] Corresponding author.

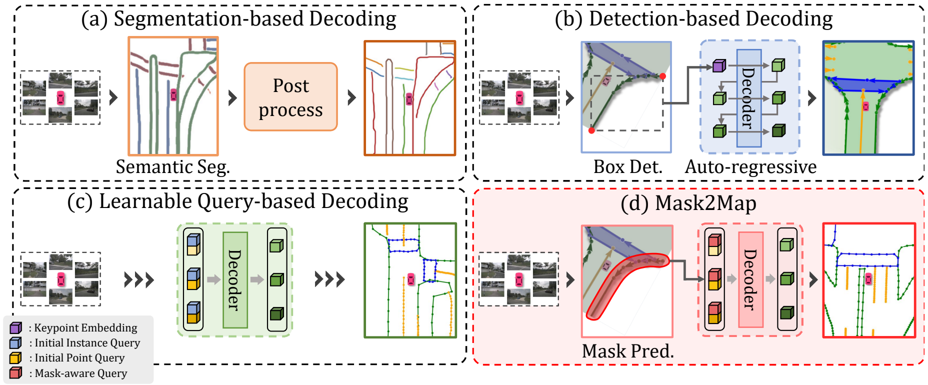

Fig. 1: Comparison of several online HD map construction methods: (a) Segmentation-based decoding, (b) detection-based decoding, (c) learnable query-based decoding, (d) proposed Mask2Map. Our Mask2Map utilizes Mask-Aware Queries to capture global-scale semantic information about a scene, enabling the generation of vectorized HD map components through subsequent query decoding.

cise planning and obstacle avoidance by providing detailed positional and semantic information about map instances. HD map has traditionally been constructed offline utilizing Simultaneous Localization and Mapping (SLAM)-based methods [33, 34, 42], involving complex processes that require significant labor intensity and economic costs. In addition, this approach is limited in its ability to provide timely updates in response to changing road conditions. Recent research has moved towards learning-based online HD map construction using onboard sensors, focusing on the generation of local maps around an autonomous vehicle. This approach eliminates costly management of HD maps, allowing immediate updates to reflect current road conditions and expansion to new locations.

Early works viewed the map construction as a semantic segmentation challenge [18,25,27,28,31,32,37,43,44] based on bird's-eye view (BEV) representation obtained from various sensors. They predicted the class label for each pixel in a raster format, avoiding the complexity of generating precise vector contours. While this method provides semantic map information, delineating map components of various classes, it falls short in capturing the precise key locations and their structural relations. Hence, its outputs are not suitably formatted for direct application to downstream tasks, such as motion forecasting [7, 13] and planning [3]. To overcome this limitation, online generation of vectorized HD map have been studied in [9, 17, 20, 21, 24, 35, 38, 40]. These methods are capable of directly producing vectorized map entities, a common feature in HD maps.

To date, various online HD map construction methods have been proposed. As depicted in Fig. 1 (a), the segmentation-based decoding method [17] was initially proposed, which involved semantic segmentation followed by the generation of vectorized maps using heuristic post-processing algorithms. However, this approach required significant processing time. The detection-based decoding method [24] identified key points corresponding to various instances and then generated vectorized map components sequentially, as shown in Fig. 1 (b). Nonetheless, relying solely on key points may not adequately capture the diverse shapes of instances, thus hindering the generation of accurate HD maps. Recently, various learnable query-based decoding methods were proposed [9, 20, 21, 38, 40], which directly predicted vectorized map components by decoding learnable queries from the BEV features in parallel, as illustrated in Fig. 1 (c). Since initial learnable queries are unrelated to a given scene, they restrict the ability to simultaneously capture the semantic and geometric information of map instances in complex scenes.

In this study, we introduce a novel end-to-end HD map construction framework, referred to as Mask2Map. As illustrated in Fig. 1 (d), Mask2Map distinguishes itself from existing approaches by leveraging segmentation masks designed to differentiate between different classes of instances in the BEV domain. The proposed Mask2Map architecture comprises two networks: an *Instance-Level Mask Prediction Network* (IMPNet) and a *Mask-Driven Map Prediction Network* (MMPNet). Initially, IMPNet constructs Multi-scale BEV Features from sensor data and generates Mask-Aware Queries to capture the semantic features of instances from a global perspective. Following the framework of the instance segmentation model, Mask2Former [5], we devise Mask-Aware Queries capable of generating BEV Segmentation Masks associated with instances of different classes in the BEV domain. Subsequently, based on the Mask-Aware Queries provided by IMPNet, MMPNet dynamically predicts the ordered point set of map instances from a local perspective in the BEV domain. In a nutshell, MMPNet focuses on the generation of coherent and refined map components by leveraging comprehensive semantic scene information obtained from IMPNet.

We introduce several innovative approaches to enhance accuracy in predicting HD maps. First, we devise the *Positional Query Generator* (PQG), which generates instance-level positional queries capturing comprehensive location information to enhance Mask-Aware Queries. Second, while most existing methods construct the HD map without considering the point-level information of each map instance, we introduce the *Geometric Feature Extractor* (GFE) to capture the geometric structure for each instance. GFE processes the BEV Segmentation Masks to extract point-level geometrical features for map instances from BEV features. Third, we observe limited performance in Mask2Map due to inter-network inconsistency when the queries from IMPNet and those from MMPNet are associated with GTs from different instances. To address this problem, we propose an *Inter-network Denoising Training* strategy [15, 41]. This approach utilizes noisy GT queries and perturbed GT Segmentation Masks as input to IMPNet and guides the model to counteract the noise, thereby ensuring inter-network consistency and enhancing the performance of HD map construction.

We evaluate the proposed Mask2Map on multiple challenging benchmarks, including nuScenes [2] and Argoverse2 [36]. Our Mask2Map achieves remarkable performance gains over the previous state-of-the-art (SOTA) methods on both benchmarks. In particular, on nuScenes benchmark, Mask2Map achieves

71.6% mAP, outperforming previous best camera-based method MapTRv2 [21] by 10.1% mAP. In the rasterization-based evaluation metric [40], Mask2Map achieves 54.7% mAP SOTA performance, more than 18.0% mAP higher than MapTRv2. On Argoverse2 benchmark, Mask2Map outperforms MapTRv2 by 4.1% mAP with the same backbone ResNet50 [12].

The contributions of this study are summarized as follows:

- We present Mask2Map, a new framework for online HD map construction. Our model captures semantic information at the instance-level from the scene and uses it to generate fine-grained map components subsequently. We integrate a key element from Mask2Former [5]: the Mask-Aware Query, redesigned to extract semantic masks in the BEV domain.
- We design a mask-guided hierarchical feature extraction architecture to efficiently encode instance-level positional information and point-level geometric information of map instances.
- We present an Inter-network Denoising Training strategy that uses noisy GT queries and perturbed GT Segmentation Masks to ensure inter-network consistency and boost the performance of HD map construction.

2 Related Works

BEV Segmentation Methods. The BEV segmentation task refers to the task of gathering information about the static environment surrounding a vehicle using sensor data. Recently, many BEV segmentation methods have adopted learning-based approaches, utilizing robust deep learning backbone models developed for 3D perception [18, 25, 27, 28, 31, 32, 37, 43, 44].

These methods typically extract BEV features from sensor data and perform semantic segmentation on the BEV domain using rasterized images of static scenes as GT. Lift-Splat-Shoot (LSS) [28] transformed features extracted from multi-view cameras into 3D features using predicted depth information and then generated BEV representation by pooling these features. CVT [44] used cross-view attention to learn geometric transformations from perspective view to BEV domain using camera-aware positional embedding. BEVFormer [18] modeled BEV representations unified by interacting with spatial and temporal information through predefined grid-shaped BEV queries. BEVSegFormer [27] conducted BEV semantic segmentation by employing a deformable cross-attention module, which generated dense semantic queries from multi-view camera features without relying on camera intrinsic and extrinsic parameters.

Vectorized HD Map Construction Methods. Recently, online HD map construction methods have received much attention for their potential to replace HD maps in autonomous driving and provide useful information for robot planning and localization. These methods predicted detailed map instances surrounding an ego vehicle using sensor data in real-time [9, 17, 20, 21, 24, 30, 35, 38, 40].

HDMapNet [17] produced vectorized HD maps using a semantic segmentation model with BEV features and a post-processing method to refine the result. However, this approach demands significant computation time. To enhance processing efficiency, query-based methods have been introduced, which leveraged Transformer attention to decode scenes and directly predict sequences of ordered points for map instances. VectorMapNet [24] introduced a two-stage framework that first detects bounding boxes of map instances and then sequentially predicts the points of each instance with an auto-regressive decoder. MapTR [20] leveraged the architecture of DETR [45] to represent map instances as ordered point sets and encode them using hierarchical queries for a Transformer decoder. MapTRv2 [21] further extended its capability by using depth supervision to learn 3D geometric information and conducting semantic segmentation on both perspective views and BEV. MapVR [40] generated a vectorized map for each map instance and subsequently transformed it into a rasterized map using a differentiable rasterizer, providing instance-level segmentation supervision. PivotNet [9] predicted map instances by generating an ordered list of pivotal points that are crucial for capturing the overall shapes of map components.

Denoising Training Strategy. Perception models based on the DETR architecture [4–6, 23, 26, 45] have adopted query-based prediction using Transformer architecture, assigning GT labels to predictions via bipartite matching to ensure proper supervision. However, such assignments can occasionally result in inconsistencies in matching across epochs or layers [15, 41]. For instance, different GT labels may be assigned to the same query over different epochs, consequently resulting in slower convergence and decreased performance.

To address this challenge, DN-DETR [15] introduced a denoising training strategy. This strategy integrates queries derived from noisy ground truth (GT) bounding boxes into the existing queries of the DETR decoder, assigning the task of predicting GT bounding boxes from these GT queries. This approach has proven effective in stabilizing bipartite matching across training epochs. MP-Former [41] addressed the issue of inconsistent mask predictions occurring between consecutive decoder layers. MP-Former employed a mask-piloted training approach that utilized both GT queries and GT masks intentionally corrupted with noise to alleviate the negative impact of inaccurate mask predictions. Mask DINO [16] introduced a unified denoising training framework that enhanced the stability of multi-task learning for object detection and segmentation tasks.

3 Method

In this section, we present the details of the proposed HD map construction method, Mask2Map.

3.1 Overview

An overall architecture of the Mask2Map is depicted in Fig. 2. The Mask2Map architecture comprises two networks: IMPNet and MMPNet. First, IMPNet generates Mask-Aware Queries capturing holistic semantic information from a global

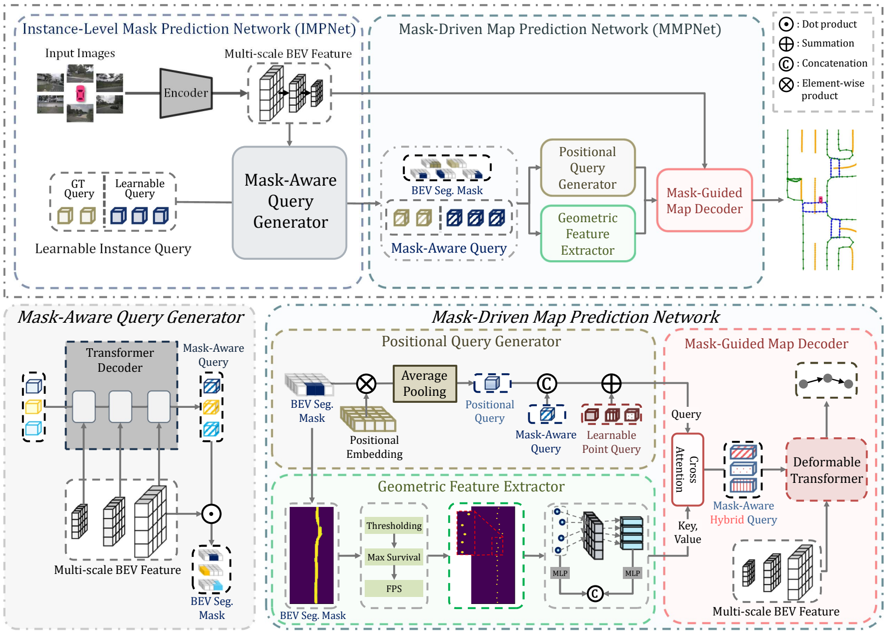

Fig. 2: Overall structure of Mask2Map. The Mask2Map system consists of IMPNet and MMPNet. IMPNet generates Mask-Aware Queries and BEV Segmentation Masks using Multi-scale BEV Features extracted from sensor data. Then, MMPNet predicts the class and ordered point set of map instances using PQG, GFE, and Mask-Guided Map Decoder. Both PQG and GFE generate semantic geometrical features on the map instances, and the Mask-Guided Map Decoder constructs vectorized maps based on these features.

perspective. Subsequently, MMPNet constructs a more detailed vectorized map from a local perspective using geometric information acquired through PQG and GFE.

3.2 Instance-Level Mask Prediction Network (IMPNet)

IMPNet consists of BEV Encoder and Mask-Aware Query Generator. BEV Encoder extracts Multi-scale BEV Features from the sensor data and Mask-Aware Query Generator produces Mask-Aware Queries, which are subsequently used to generate BEV Segmentation Masks.

BEV Encoder. IMPNet generates BEV features by processing multi-view camera images, LiDAR point clouds, or a fusion of both modalities. Multi-view camera images are transformed into BEV representation utilizing the LSS operation [28]. LiDAR point clouds are converted into BEV representation through voxel encoding [39]. When integrating both camera and LiDAR sensors for fusion, BEV features extracted from both modalities are concatenated and passed through additional convolutional layers.

Next, BEV Encoder produces BEV features of multiple scales through downsampling layers. These multi-scale features are then jointly encoded using the Deformable Transformer Encoder [45] to encode relations between Multi-scale BEV Features. This process yields Multi-scale BEV Features $\mathbf{F}^{\mathrm{BEV}} = \{F_l^{\mathrm{BEV}}\}_{l=1}^{S}$, where l denotes the scale index and S represents the total number of scales. The scale index of $l = 1$ represents the smallest scale, whereas $l = S$ signifies the largest scale. We denote H and W as the height and width of the BEV feature F_S^{BEV} at the largest scale.

Mask-Aware Query Generator. Mask-Aware Query Generator extracts Mask-Aware Queries from Multi-scale BEV Features using the Mask Transformer proposed in Mask2Former [5]. The Mask-Aware Queries are initialized with learnable vectors and are decoded through M layers of the Transformer decoder. Given Multi-scale BEV Features $\mathbf{F}^{\mathrm{BEV}}$ and the BEV Segmentation Masks $\mathbf{M}_{m-1} = \{M_{i,m-1}\}_{i=1}^{N_I}$ obtained at the $(m-1)$-th decoding layer, the Mask-Aware Queries $\mathbf{q}_{m-1}^{\mathrm{mask}} = \{q_{i,m-1}^{\mathrm{mask}}\}_{i=1}^{N_I}$ are updated as

$$\hat{\mathbf{M}}_{m-1} = \begin{cases} 0 & \text{if } \mathbf{M}_{m-1} > \tau_M \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

$$Q_m = \mathbf{q}_{m-1}^{\mathrm{mask}} W^Q, \quad K_m = F_m^{\mathrm{BEV}} W^K, \quad V_m = F_m^{\mathrm{BEV}} W^V \tag{2}$$

$$\mathbf{q}_m^{\mathrm{mask}} = \mathrm{softmax}(\hat{\mathbf{M}}_{m-1} + Q_m K_m^T) V_m + \mathbf{q}_{m-1}^{\mathrm{mask}}, \tag{3}$$

where τ_M denotes a threshold, N_I denotes the number of the Mask-Aware Queries, and W^Q, W^V, and W^K are learnable weight matrices. Finally, the BEV Segmentation Masks $\mathbf{M}_m$ are obtained by applying dot product between BEV feature F_S^{BEV} of the largest scale and the Mask-Aware Queries $\mathbf{q}_m^{\mathrm{mask}}$ along the channel axis. Then the sigmoid function is applied to normalize the BEV Segmentation Masks. These BEV Segmentation Masks are then fed back into the next decoding layer for further refinement. After M decoding layers, IMPNet ends up with the final Mask-Aware Queries $\mathbf{q}^{\mathrm{mask}} = \mathbf{q}_M^{\mathrm{mask}}$ and the BEV Segmentation Masks $\mathbf{M}^{\mathrm{BEV}} = \mathbf{M}_M$, which are delivered to the subsequent MMPNet.

3.3 Mask-Driven Map Prediction Network (MMPNet)

MMPNet comprises three main components: the Positional Query Generator, the Geometric Feature Extractor, and the Mask-Guided Map Decoder. The Positional Query Generator injects positional information to enhance Mask-Aware Queries, while the Geometric Feature Extractor processes the BEV Segmentation Masks to extract geometrical features from the BEV feature. Finally, the Mask-Guided Map Decoder predicts both the class and coordinates of ordered points for map instances using the features provided by the Positional Query Generator and Geometric Feature Extractor.

Positional Query Generator. While Mask-Aware Queries carry semantic information about map instances, they lack positional information. To enable MMPNet to generate coordinates of points for map instances, it is essential to integrate positional information in the BEV domain into the Mask-Aware Queries. PQG initially derives the sparsified BEV mask from the BEV Segmentation Mask $\mathbf{M}^{\text{BEV}}$,

$$\hat{\mathbf{M}}^{\text{PQG}} = \begin{cases} \mathbf{M}^{\text{BEV}} & \text{if } \mathbf{M}^{\text{BEV}} > \tau_P \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

PQG injects the 2D positional embedding PE to the sparsified BEV mask $\hat{\mathbf{M}}_i^{\text{PQG}}$, where PE is generated by sinusoidal functions [4]. Then, the positional queries f_i^{pos} is obtained by applying average pooling in both x and y domains, i.e.,

$$f_i^{\text{pos}} = \frac{1}{N_{\text{nz}}^i} \sum_{x=1}^{H} \sum_{y=1}^{W} (\hat{\mathbf{M}}_i^{\text{PQG}}(x,y) \otimes PE(x,y)), \tag{5}$$

where $i \in [1, N_I]$, N_{nz}^i denotes the number of non-zero pixels in $\hat{\mathbf{M}}_i^{\text{PQG}}$ and $\otimes$ is the element-wise product. The positional queries $\mathbf{f}^{\text{pos}}$ are concatenated with the Mask-Aware Queries $\mathbf{q}^{\text{mask}}$ to generate the Combined Positional Queries $\hat{\mathbf{f}}^{\text{pos}}$.

Next, the Combined Positional Query $\hat{f}_i^{\text{pos}}$ is used to produce N_P point-level features for the i-th map instance. Towards this goal, PQG replicates $\hat{f}_i^{\text{pos}}$ N_P times and adds them with N_P Learnable Point Queries q_1, ..., q_{N_P}, generating the PQG Query Features $\mathbf{q}_i^{\text{PQG}} = \{q_{i,j}^{\text{PQG}}\}_{j=1}^{N_P}$,

$$q_{i,j}^{\text{PQG}} = \hat{f}_i^{\text{pos}} + q_j, \tag{6}$$

where $i \in [1, N_I]$ and $j \in [1, N_P]$. Note that the learnable queries q_1 through q_{N_P} give the Mask-Aware Queries an idea of the sequential order of points generated for the i-th map instance. The resulting PQG Query Features $\mathbf{q}_i^{\text{PQG}}$ are delivered to the Mask-Guided Map Decoder.

Geometric Feature Extractor. GFE generates point-wise features that capture the geometric structure of map instances. Initially, using the threshold τ_G, GFE produces the sparsified BEV mask $\hat{\mathbf{M}}^{\text{GFE}}$ from the BEV Segmentation Mask $\mathbf{M}^{\text{BEV}}$. To generate point-wise geometric features, GFE samples N_S key pixels from the sparsified BEV mask $\hat{M}_i^{\text{GFE}}$. First, we employ the *Max Survival* method, which selects the strongest pixel from the non-overlapping window of size $G \times G$ sliding on $\hat{M}_i^{\text{GFE}}$ while setting the remaining pixels to zero. Next, we apply the Farthest Point Sampling (FPS) method [29] to iteratively select the output of the Max Survival method and identify N_S key points. Finally, based on the positions of the N_S key points, N_S features are pooled from the BEV features F_S^{BEV} of the largest scale. Concurrently, the (x, y) coordinates of these N_S key points are encoded using a Multi-Layer Perceptron (MLP). These two features are concatenated, resulting in the GFE Features denoted as $\mathbf{f}_i^{\text{GFE}} = \{f_{i,j}^{\text{GFE}}\}_{j=1}^{N_S}$.

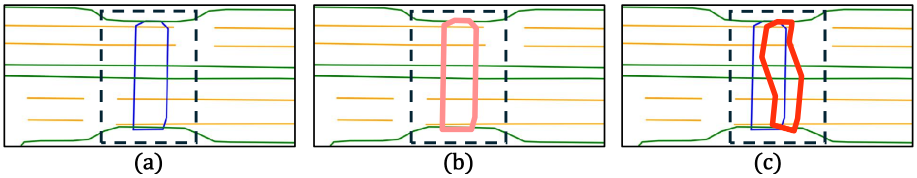

Fig. 3: Illustration of proposed Map Noise method. (a) The blue polygon denotes a vectorized GT of a pedestrian crossing. (b) The pink polygon represents a GT Segmentation Mask without noise. (c) The red polygon represents the result of adding Map Noise to the GT Segmentation Mask.

Mask-Guided Map Decoder. The Mask-Guided Map Decoder predicts the class and sequence of ordered points for vectorized map components based on PQG Query Features $\mathbf{q}_i^{\mathrm{PQG}}$ and GFE Features $\mathbf{f}_i^{\mathrm{GFE}}$. By using $\mathbf{q}_i^{\mathrm{PQG}}$ as queries and $\mathbf{f}_i^{\mathrm{GFE}}$ as keys and values, the cross-attention module produces Mask-Aware Hybrid Queries $\mathbf{q}_i^{\mathrm{Hybrid}} = \{q_{i,j}^{\mathrm{Hybrid}}\}_{j=1}^{N_P}$. These queries are subsequently decoded by a Deformable Transformer [45], utilizing the Multi-scale BEV Features $\mathbf{F}^{\mathrm{BEV}}$ as values. Finally, the prediction heads predict instance classification scores and normalized BEV coordinates for each map instance through the classification and regression heads, respectively.

3.4 Inter-network Denoising Training

Mask2Map passes Mask-Aware Queries from IMPNet to MMPNet for hierarchical refinement of instance features. To ensure efficient training, we assign instance segmentation loss for IMPNet and map construction loss for MMPNet. Following the training strategy suggested in [4], queries used by IMPNet and those by MMPNet should be matched to their respective GT through bipartite matching. However, inconsistencies in this matching process can occur when the queries used in both IMPNet and MMPNet are matched with GTs associated with different instances. We observe that this inter-network inconsistency tends to cause slower convergence and diminished performance.

To address this issue, we adopt a denoising training strategy [41]. The key idea is to merge noisy GT Queries, which are derived from each GT instance, into the learnable queries within IMPNet. (see Fig.2). Our model is trained to denoise these queries by directly matching them with their corresponding GTs. This is contrasted with the learnable queries, which are matched to the GTs through bipartite matching. Thus, this strategy is called Inter-network Denoising Training. This process guides the model to establish a correspondence between the queries used in IMPNet and MMPNet, effectively mitigating inter-network inconsistency. Additionally, alongside GT Queries, we also generate GT Segmentation Masks, which replace the BEV Segmentation Masks for IMPNet.

We generate GT Queries by assigning one of the C class embedding vectors corresponding to a GT class of each instance, where C denotes the number of classes. We add a flipping noise by randomly replacing the class embedding vector with one of other classes at a probability of λ. Simultaneously, we also generate perturbed GT Segmentation Masks by adding *Map Noises* to a sequence of ordered points of each instance and rasterize them, as shown in Fig. 3 (c).

The combination of the noisy GT Queries and learnable queries is referred to as Learnable Instance Queries. Instead of utilizing BEV Segmentation Masks, we exclusively employ the perturbed GT Segmentation Masks for noisy GT Queries. The noisy GT Queries pass through both IMPNet and MMPNet and their prediction results are matched with the corresponding GTs without the bipartite matching.

3.5 Training Loss

The total loss L used to train Mask2Map is given by

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{map}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{dn}}, \tag{7}$$

where $\mathcal{L}_{\text{seg}}$ is the loss term for training IMPNet on a BEV segmentation task, $\mathcal{L}_{\text{map}}$ is the loss term for training MMPNet on a map construction task, $\mathcal{L}_{\text{aux}}$ is the auxiliary loss term, $\mathcal{L}_{\text{dn}}$ is the loss term for Inter-network Denoising Training.

We use bipartite matching with Hungarian solver to assign the queries used by IMPNet and those by MMPNet to their respective GTs. Based on the assignment, we calculate the $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{map}}$. We adopt [5] to obtain the loss term $\mathcal{L}_{\text{seg}}$. The loss term $\mathcal{L}_{\text{map}}$ consists of L1 loss for regression of vectorized map instances, focal loss [22] for the classification of instances, and cosine similarity loss between the direction calculated from adjacent points of GT and the one from the predictions. The auxiliary loss term $\mathcal{L}_{\text{aux}}$ calculates the error for depth estimation and 2D map semantic segmentation tasks conducted on the camera perspective-view features [21]. The loss term $\mathcal{L}_{\text{dn}}$ is the summation of two terms, each corresponding to the loss used in IMPNet and MMPNet without bipartite matching between GT and predictions for noisy GT Queries.

4 Experiments

4.1 Experimental Settings

Datasets. Both nuScenes [2] and Argoverse2 [36] datasets are real-world driving datasets that provide HD map labels annotated by hand. The nuScenes dataset consists of $28K$ frames for training set and $6K$ frames for validation set. Each frame is annotated at 2Hz from six surrounding RGB cameras and a 32-beam LiDAR covering 360° field of view. The Argoverse2 dataset contains training set of $110K$ frames and validation set of $24K$ frames. Annotations for keyframes are presented at 10hz with 7 ring cameras and two 32-beam LiDARs.
Evaluation Metrics. We define the perception range $[-15.0m, 15.0m]$ for the lateral direction and $[-30.0m, 30.0m]$ for the longitudinal direction. Following prior works [9, 20, 21, 24, 38, 40], we categorize map instances into three types for HD map construction: *Pedestrian Crossing*, *Lane Divider*, and *Road Boundary.*

We adopt two evaluation metrics: Average Precision (AP) based on Chamfer distance proposed in [17] and AP based on rasterization proposed in [40]. We primarily utilize the Chamfer distance metric, using thresholds of 0.5, 1.0, and 1.5 meters for mean AP (mAP). For rasterization-based mean AP ($mAP^{\dagger}$), we measure intersection over union for each map instance, with thresholds set $\{0.50, 0.55, ..., 0.75\}$ for pedestrian crossings and $\{0.25, 0.30, ..., 0.50\}$ for line-shaped elements. To further evaluate the matching consistency ratio of inter-network, we use the Query Utilization ($Util$) metric proposed by [41], which calculates the consistency ratio of GT matches in MMPNet's first decoder layer with the matches in IMPNet's last layer.

Implementation Details. We adopted ResNet50 [12] for image backbone networks. For nuScenes, images with dimensions of 1600×900 were resized by a 0.5 ratio. In the case of Argoverse2, seven images with dimensions of 1550×2048 for the front view, and others with dimensions of 2048×1550 were padded to 2048×2048 before being resized by a 0.3 ratio. The LiDAR point clouds were voxelized with a voxel size of 0.1, 0.1, and 0.2. The voxel features were extracted by SECOND [39]. We employed six BEV encoder layers and three mask Transformer layers in IMPNet. We employed six Transformer decoder layers in MMPNet. The thresholds for BEV Segmentation Mask, τ_M, τ_P, and τ_G were set to 0.5, 0.3, and 0.8, respectively. We configured the number of instance queries to 50 and the number of point queries to 20. In GFE, we set the window size (G) for the Max Survival method to 4 and the number of sampling points (N_S) to 20. The flipping noise probability λ was set to 0.2. For optimization, we employed AdamW with a weight decay of 0.01 and utilized cosine annealing as a scheduler. The initial learning rate was set to 6e-4. Our model was trained on 4 RTX3090 GPUs with batch size 4 per GPU.

4.2 Performance Comparison

Results on nuScenes. Table 1 presents the comprehensive performance analysis of Mask2Map on the nuScenes validation set [2], utilizing the Chamfer distance metric. Mask2Map establishes a new state-of-the-art performance, exhibiting substantial improvements over existing methods [9,17,20,21,24,30,40]. When using camera input only, Mask2Map achieves noteworthy results of 71.6% mAP at 24 epochs and 74.6% mAP at 110 epochs, outperforming the previous state-of-the-art model, MapTRv2 [21], by 10.1% mAP and 5.9% mAP, respectively. When using camera-LiDAR fusion, Mask2Map achieves the performance gain of 9.4% mAP over MapTRv2. Table 2 evaluates the performance of Mask2Map based on a rasterization-based metric. Notably, our Mask2Map method achieves a remarkable performance gain of 18.0 mAP over MapTRv2.

Results on Argoverse2. Table 3 presents the performance evaluation of several HD map construction methods [20, 21, 24, 40] on the Argoverse2 validation set [36]. The proposed Mask2Map demonstrates significant performance improvements compared to existing models. Mask2Map surpasses the current leading method, MapTRv2, by 4.1% mAP, demonstrating that our model achieves the consistent performance across different scenarios.

Table 1: Comparison with SOTA methods on the nuScenes validation set. FPSs are measured on the same machine equipped with RTX3090. The "-" denotes that the associated results are not available. "C" and "L" respectively denote camera and LiDAR. The "R50", "PP" and "Sec" respectively correspond to ResNet50 [12], PointPillars [14], and SECOND [39].

Method	Modality	Backbone	Epoch	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP	FPS
MapTR [20]	C	R50	24	46.3	51.5	53.1	50.3	15.1
MapVR [40]	C	R50	24	47.7	54.4	51.4	51.2	15.1
PivotNet [9]	C	R50	24	56.2	56.5	60.1	57.6	-
BeMapNet [30]	C	R50	30	57.7	62.3	59.4	59.8	-
MapTRv2 [21]	C	R50	24	59.8	62.4	62.4	61.5	14.1
Ours	C	R50	24	**70.6**	**71.3**	**72.9**	**71.6**	10.1
VectorMapNet [24]	C	R50	110	36.1	47.3	39.3	40.9	2.2
MapTR [20]	C	R50	110	56.2	59.8	60.1	58.7	15.1
MapVR [40]	C	R50	110	55.0	61.8	59.4	58.8	15.1
MapTRv2 [21]	C	R50	110	68.1	68.3	69.7	68.7	14.1
Ours	C	R50	110	**73.6**	**73.1**	**77.3**	**74.6**	10.1
VectorMapNet [24]	C+L	R50 & PP	110	37.6	50.5	47.5	45.2	-
MapTR [20]	C+L	R50 & Sec	24	55.9	62.3	69.3	62.5	6.0
MapVR [40]	C+L	R50 & Sec	24	60.4	62.7	67.2	63.5	6.0
MapTRv2 [21]	C+L	R50 & Sec	24	65.6	66.5	74.8	69.0	5.8
Ours	C+L	R50 & Sec	24	**76.5**	**76.6**	**82.1**	**78.4**	4.1

Table 2: Comparison of SOTA methods on nuScenes validation set with rasterization-based metric. All models use camera modality for input and ResNet50 [12] as backbones. The "∗" indicates results reproduced using public codes.

Method	Epoch	$AP^{\dagger}_{ped}$	$AP^{\dagger}_{divider}$	$AP^{\dagger}_{boundary}$	$mAP^{\dagger}$
MapTR [20]	24	32.4	23.5	17.1	24.3
MapVR [40]	24	37.5	33.1	23.0	31.2
MapTRv2* [21]	24	49.9	34.7	25.7	36.7
Ours	24	**62.9**	**52.3**	**48.9**	**54.7**

Table 3: Comparison with state-of-the-art methods on the Argoverse2 validation set. All of the presented results derive from models trained using camera data as the input.

Method	Backbone	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet [17]	EB0	13.1	5.7	37.6	18.8
VectorMapNet [24]	R50	38.3	36.1	39.2	37.9
MapTR [20]	R50	54.7	58.1	56.7	56.5
MapVR [40]	R50	54.6	60.0	58.0	57.5
MapTRv2 [21]	R50	62.9	72.1	67.1	67.4
Ours	R50	**68.1**	**72.7**	**73.7**	**71.5**

4.3 Ablation Study

We conducted an ablation study to evaluate the contributions of the core ideas of Mask2Map. Camera-only input and ResNet50 [12] backbone were used in these experiments. Training was conducted on 1/4 of the nuScenes training dataset for 24 epochs. Evaluation was performed on the entire validation set.

Contributions of Main Components. Table 4 demonstrates the impact of each component of Mask2Map. We evaluated performance by adding each component one by one. The first row represents a baseline model using an LSS-based BEV encoder for extracting BEV features and Deformable attention for predicting vectorized map instances [21]. When adding IMPNet into the baseline model, we notice a substantial 5.9% increase in mAP, indicating that the inclusion of Mask-Aware Queries, capable of generating instance segmentation results, sig-

Table 4: Ablation study of main components of Mask2Map

	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
Baseline	30.1	41.5	46.6	39.4
+IMPNet	42.7	45.3	48.1	45.3
+MMPNet	45.7	47.0	54.4	49.1
+Denoising	**52.9**	**55.5**	**58.5**	**55.6**

Table 5: Ablation study for evaluating the contributions of MMPNet's submodules

PQG	GFE	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
		47.6	51.1	53.8	50.8
	✓	49.7	54.2	57.8	53.9
✓		49.9	**55.6**	57.2	54.2
✓	✓	**52.9**	55.5	**58.5**	**55.6**

Table 6: Ablation study for evaluating the effect of Inter-network Denoising Training

Denoising	$Util$	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
	24.7	45.7	47.0	54.4	49.1
✓	**74.7**	**52.9**	**55.5**	**58.5**	**55.6**

Table 7: Ablation study for Map Noise applied to GT Segmentation Masks

Map Noise	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
	49.7	**56.1**	**58.6**	54.8
✓	**52.9**	55.5	58.5	**55.6**

nificantly boosts the performance of HD map construction. Furthermore, the addition of MMPNet results in a further improvement of 3.8% mAP, underscoring the significant contribution of injecting positional and geometric information of map instances through BEV Segmentation Masks. Lastly, our Inter-network Denoising Training offers a 6.5% additional increase in mAP, emphasizing its effectiveness in enhancing performance.

Contributions of MMPNet's Submodules. In our investigation detailed in Table 5, we explored the contributions of PQG and GFE. GFE alone contributes to a notable 3.1% increase in mAP over the baseline, while PQG alone yields a 3.4% improvement in mAP. The combination of PQG and GFE further enhances performance by 4.8% mAP, demonstrating their complementary effects.

Effect of Inter-network Denoising Training on Matching Consistency. We further investigate the impact of Inter-network Denoising Training. As depicted in Table 6, Inter-network Denoising Training drastically increases the matching ratio $Util$ from 24.7% to 74.7%, which is translated into a substantial 6.5% increase in overall mAP performance. This demonstrates that our Inter-network Denoising Training effectively mitigates the inconsistency in query-to-GT matching between IMPNet and MMPNet, as intended.

Impact of Noise in Inter-network Denoising Training. In Table 7, we explore the effect of Map Noise used in the Inter-network Denoising Training. We compare our method with a baseline using GT Segmentation Masks without Map Noise. Our findings show that adding Map Noise to GT results in a 0.8% improvement in mAP over the baseline, indicating the benefit of this approach.

4.4 Qualitative Analysis

Qualitative Results. Fig. 4 presents the qualitative results produced by the proposed Mask2Map. We compare our method with the current SOTA, MapTRv2. Note that Mask2Map yields notably better map construction results than MapTRv2. More qualitative results are provided in Supplementary Material.

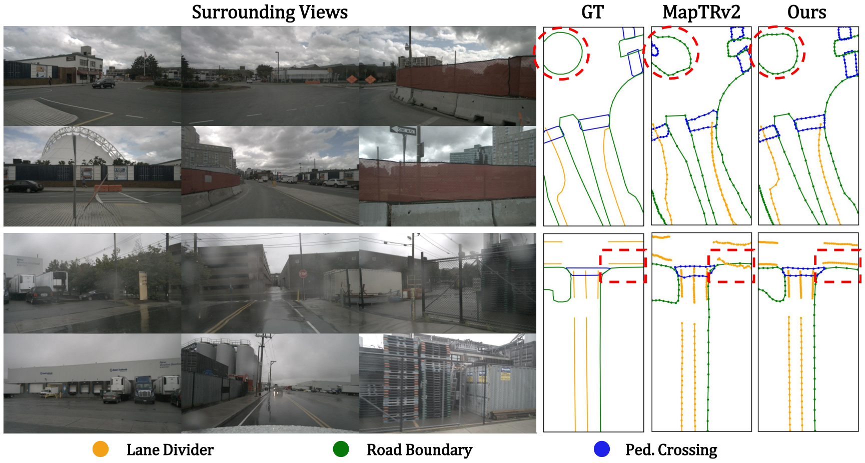

Fig. 4: Qualitative results on the nuScenes validation set. We compared our method with MapTRv2. The regions marked by a red ellipse and rectangle emphasize the superior results generated by our proposed model.

5 Discussion and Conclusion

Limitations and Future Work. As for future work, we consider improving Mask2Map in two aspects. (i) Temporal information is known to improve the reliability of results in autonomous driving perception tasks. However, our model currently relies solely on input from the current frame, which may lead to performance degradation in scenes occluded by objects. Temporal fusion methods through queries or BEV features of previous frames may provide promising paths toward addressing this limitation. (ii) Our experiments showed that Mask2Map's FPS decreased compared to the current SOTA, MapTRv2 [21], in exchange for substantial performance gains. To meet real-time requirements, we consider employing model compression and optimization methods. These techniques will be a promising avenue to improve the FPS without sacrificing performance.

Conclusion. In this paper, we introduced an end-to-end online HD map construction method called Mask2Map. Mask2Map utilizes IMPNet to generate both Mask-Aware Queries and BEV Segmentation Masks, capturing semantic scene context from a global perspective. Subsequently, MMPNet enhances Mask-Aware Queries by incorporating semantic and geometrical information through PQG and GFE. Finally, Mask-Guided Map Decoder predicts the class and ordered point set of map instances. Additionally, we proposed Inter-network Denoising Training to mitigate inter-network inconsistency arising from differing bipartite matching results between IMPNet and MMPNet. Our evaluation on nuScenes and Argoverse2 benchmarks demonstrated that the proposed ideas yielded significant performance improvements over the baseline, surpassing existing HD map construction methods by considerable margins.

Acknowledgements. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No.2020R1A2C2012146], and the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129).

References

1. Antonello, M., Carraro, M., Pierobon, M., Menegatti, E.: Fast and robust detection of fallen people from a mobile robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4159–4166. IEEE (2017)
2. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11621–11631 (2020)
3. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
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. Springer (2020)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1290–1299 (2022)
6. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation pp. 17864–17875 (2021)
7. Choi, S., Kim, J., Yun, J., Choi, J.W.: R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8525–8535 (2023)
8. Da, F., Zhang, Y.: Path-aware graph attention for hd maps in motion prediction. In: International Conference on Robotics and Automation (ICRA). pp. 6430–6436. IEEE (2022)
9. Ding, W., Qiao, L., Qiu, X., Zhang, C.: Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3672–3682 (2023)
10. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11525–11533 (2020)
11. Hasan, A.M., Samsudin, K., Ramli, A.R., Azmir, R., Ismaeel, S.: A review of navigation systems (integration and algorithms). Australian Journal of Basic and Applied Sciences (AJBAS) **3**(2), 943–959 (2009)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

13. Kim, B., Park, S.H., Lee, S., Khoshimjonov, E., Kum, D., Kim, J., Kim, J.S., Choi, J.W.: Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14636–14645 (2021)
14. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12697–12705 (2019)
15. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13619–13627 (2022)
16. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3041–3050 (2023)
17. Li, Q., Wang, Y., Wang, Y., Zhao, H.: Hdmapnet: An online hd map construction and evaluation framework. In: International Conference on Robotics and Automation (ICRA). pp. 4628–4634. IEEE (2022)
18. 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. Springer (2022)
19. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., Urtasun, R.: Learning lane graph representations for motion forecasting. In: European Conference on Computer Vision (ECCV). pp. 541–556 (2020)
20. 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: The Eleventh International Conference on Learning Representations (ICLR) (2023)
21. 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)
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
23. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. International Conference on Learning Representations (ICLR) (2022)
24. 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)
25. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3262–3272 (2023)
26. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3651–3660 (2021)
27. Peng, L., Chen, Z., Fu, Z., Liang, P., Cheng, E.: 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). pp. 5935–5943 (2023)

28. 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). pp. 194–210 (2020)
29. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++ deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). pp. 5105–5114 (2017)
30. Qiao, L., Ding, W., Qiu, X., Zhang, C.: End-to-end vectorized hd-map construction with piecewise bezier curve. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13218–13228 (2023)
31. Reiher, L., Lampe, B., Eckstein, L.: A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–7. IEEE (2020)
32. Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11138–11147 (2020)
33. 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. IEEE (2018)
34. Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D.: Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5135–5142. IEEE (2020)
35. Shin, J., Rameau, F., Jeong, H., Kum, D.: Instagram: Instance-level graph modeling for vectorized hd map learning. arXiv preprint arXiv:2301.04470 (2023)
36. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS) (2021)
37. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In: Conference on Robot Learning (CoRL). pp. 989–1000. PMLR (2023)
38. Xu, Z., Wong, K.K., Zhao, H.: Insightmapper: A closer look at inner-instance information for vectorized high-definition mapping. arXiv preprint arXiv:2308.08543 (2023)
39. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
40. Zhang, G., Lin, J., Wu, S., Luo, Z., Xue, Y., Lu, S., Wang, Z., et al.: Online map vectorization for autonomous driving: A rasterization perspective pp. 31865–31877 (2023)
41. Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L.M., Zhang, L.: Mp-former: Mask-piloted transformer for image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18074–18083 (2023)
42. Zhang, J., Singh, S.: Loam: Lidar odometry and mapping in real-time. In: Robotics: Science and Systems (RSS). vol. 2, pp. 1–9. Berkeley, CA (2014)
43. Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., Lu, J.: Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv:2205.09743 (2022)

44. Zhou, B., Krähenbühl, P.: Cross-view transformers for real-time map-view semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13760–13769 (2022)
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2021)