# MGMAPNET: MULTI-GRANULARITY REPRESENTATION LEARNING FOR END-TO-END VECTORIZED HD MAP CONSTRUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The construction of Vectorized High-Definition (HD) map typically requires capturing both category and geometry information of map elements. Current state-of-the-art methods often adopt solely either point-level or instance-level representation, overlooking the strong intrinsic relationships between points and instances. In this work, we propose a simple yet efficient framework named MGMapNet (Multi-Granularity Map Network) to model map element with a multi-granularity representation, integrating both coarse-grained instance-level and fine-grained point-level queries. Specifically, these two granularities of queries are generated from the multi-scale bird's eye view (BEV) features using a proposed Multi-Granularity Aggregator. In this module, instance-level query aggregates features over the entire scope covered by an instance, and the point-level query aggregates features locally. Furthermore, a Point Instance Interaction module is designed to encourage information exchange between instance-level and point-level queries. Experimental results demonstrate that the proposed MGMapNet achieves state-of-the-art performance, surpassing MapTRv2 by 5.3 mAP on nuScenes and 4.4 mAP on Argoverse2 respectively.

## 1 INTRODUCTION

Perceiving and understanding road map elements are crucial for ensuring the safety in autonomous driving applications (Xiao et al., 2020; Xu et al., 2023; Prakash et al., 2021). High-Definition (HD) maps provide category and geometry information about road elements, enabling autonomous vehicles to maintain lane position, anticipate intersections, and plan optimal routes to mitigate potential risks. However, constructing HD map requires significant human effort for annotating and updating, which limits scalability over large areas. Recent research, such as (Li et al., 2022a; Liao et al., 2022; 2023; Ding et al., 2023b; Yuan et al., 2024; Hu et al., 2021), focuses on learning-based methods as alternatives to construct HD map from onboard sensors.

These methods can be mainly divided into two categories based on the representation in use: rasterized map based representation (Li et al., 2022a;b; Liu et al., 2023b; Xiong et al., 2023) and vectorized map based representation (Ding et al., 2023b; Li et al., 2023; Liao et al., 2023).

Rasterized map based methods often require complex post-processing to meet the need of downstream modules, such as planning. Consequently, this process may result in suboptimal results that are not entirely end-to-end optimized. Therefore, there has been increasing attention paid to end-to-end map construction methods (Shin et al., 2023; Qiao et al., 2023b; Zhang et al., 2024) using vectorized representations.

Vectorized map based methods commonly employ Bird's Eye View (BEV) (Fadadu et al., 2022; Chen et al., 2017; Liang et al., 2019; You et al., 2019; Liang et al., 2018) space for end-to-end perception, effectively integrating various sensor information such as surround-view cameras and Lidar. State-of-the-art (SOTA) methods typically adopt DETR-like architectures, comprising encoder and decoder components. The encoder initially extracts multi-sensor information into BEV representation, while the decoder subsequently decodes the category and geometry information of each road element through queries. These methods achieve an end-to-end vectorized representation of output

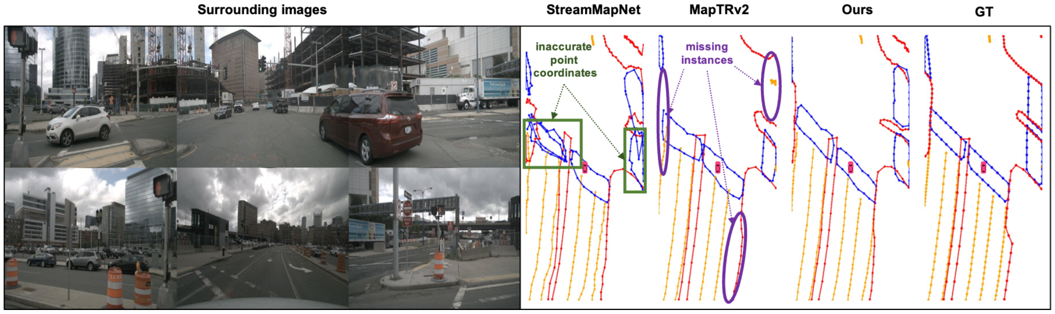


Figure 1: **Comparison of visualization results.** Visual comparison among StreamMapNet, MapTRv2, and MGMapNet, with vehicle-centric views featuring red boundaries, orange dividers, and blue pedestrian crossings. The green boxes denote the phenomenon of inaccurate point coordinates in instance-level queries, while the purple ellipses indicate the phenomenon of missing instances in point-level queries. StreamMapNet employs MPA for single-frame results. Best viewd in color.

map elements, eliminating the need for the complex post-processing steps involved in rasterized maps representation.

SOTA methods use either point-level queries or instance-level queries to generate map elements. Point-level queries are good at describing the geometric position of road elements. For instance, in MapTR (Liao et al., 2022) and MapTRv2 (Liao et al., 2023), a permutation-equivalent point expression accurately represents the location information of map elements, ensuring stable training processes. However, these methods may lack an overall description of map elements, leading to deficiencies in representing lane relationships. For example, MapTRv2 may miss lane lines in distant and merging scenarios, as the region illustrated in the purple ellipses of Fig. 1.

While instance-level queries excel at capturing the overall category information of a road element, they may struggle to accurately represent geometric details, especially for irregular or elongated map elements. For example, in StreamMapNet (Yuan et al., 2024), the Multi-Point Attention is proposed to capture the overall information of road elements, allowing for longer attention ranges while maintaining computational efficiency. However, this method may encounter difficulties in accurately perceiving the geometry of irregular or elongated elements, leading to local disturbances. The green boxes in Fig. 1 highlight the issue of inaccurate point coordinates obtained from instance-level queries, where the map elements are detected though, their positional accuracy is compromised.

The primary challenge lies in balancing detailed and comprehensive representations. The balance between detail and overview remains a major challenge in current research, and existing methods do not adequately address this issue. To integrate both fine-grained local positions and coarse-grained global classification information, we propose MGMapNet (Multi-Granularity Map Network), a framework that represents map elements using multi-granularity queries. Within each decoder layer, point-level queries and instance-level queries are simultaneously computed by querying the multi-scale BEV features using Multi-Granularity Aggregator. Subsequently, Point Instance Interaction, including point-to-point attention and point-to-instance attention, is designed to enhance the intrinsic relationships. Ultimately, point-granularity queries are utilized for localizing point coordinates, while instance-granularity queries are employed for determining the categories of map elements.

Our main contributions can be summarized as follows:

- We propose a robust multi-granularity representation, enabling the end-to-end construction of vectorized HD map by employing coarse-grained instance-level and fine-grained point-level queries in one framework.
- The Multi-Granularity Aggregator, combined with Point Instance Interaction, facilitates an efficient interaction between point-level and instance-level queries, effectively exchanging category and geometry information.
- We incorporated several strategy optimizations into the training, enabling our proposed MGMapNet to achieve state-of-the-art (SOTA) single-frame performance on both the nuScenes and Argoverse2 datasets.

## 2 RELATED WORK

**Online HD Map Construction.** In recent years, researchers have increasingly utilized onboard sensors in autonomous driving to construct HD map. Previous work (Huang et al., 2023)(Chen et al., 2022) has focused on projecting and lifting map elements detected on the Perspective View (PV) plane into 3D space for map reconstruction. With the aim of better integrating multiple sensors such as panoramic cameras and LiDAR, construction methods for online HD map are gradually transitioning to BEV representation. Currently, HD map construction can be broadly categorized into two types: rasterized map-based and vectorized map-based methods. Rasterized methods, such as HDMapNet (Li et al., 2022a), utilize BEV features for semantic segmentation, followed by a post-processing step to obtain vectorized map instances. Similarly, BEV-LaneDet (Wang et al., 2023) outputs confidence scores, embeddings for clustering, y-axis offsets, and average heights for each grid. While rasterized maps can provide detailed road information, the requirement for post-processing limits their applications. With the emergence of vectorized DETR-like (Carion et al., 2020) end-to-end methods, the need for post-processing is eliminated. VectorMapNet (Liu et al., 2023a) is the first end-to-end map reconstruction model that utilizes transformers. MapTR and MapTRv2 (Liao et al., 2022; 2023) introduce a novel and unified modeling method for map elements, addressing ambiguity and ensuring stable learning processes. PivotNet (Ding et al., 2023b) employs unified, pivot-based representations for map elements and is formulated as a direct set prediction paradigm.

However, these methods often exclusively use either point-level queries or instance-level queries, missing out on the mutual advantages of both granularities. To address this limitation, this paper introduces a multi-granularity mechanism for representing map elements. This mechanism adaptively derives features at both fine-grained point granularity and coarse-grained instance granularity, thus preserving local details as well as global map information.

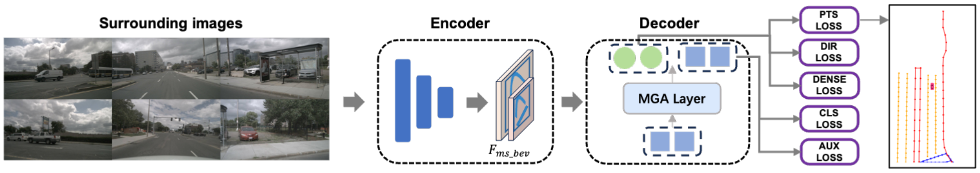
**Lane Detection.** Lane detection can be regarded as a subtask of high-definition map construction, focusing on the detection of lane elements within road scenes. Current methods (Li et al., 2019; Zheng et al., 2022; Tabelini et al., 2021b) predominantly engage in lane detection from a single perspective view (PV) image, and the majority of lane detection datasets provide annotations only from a single perspective. LaneATT (Tabelini et al., 2021a) proposes a novel anchor-based attention mechanism that aggregates global information. Unlike lane detection, vectorized HD map construction involves more complex map elements within the vehicle's perception range, including lane markings, curbs, and sidewalks.

## 3 METHOD

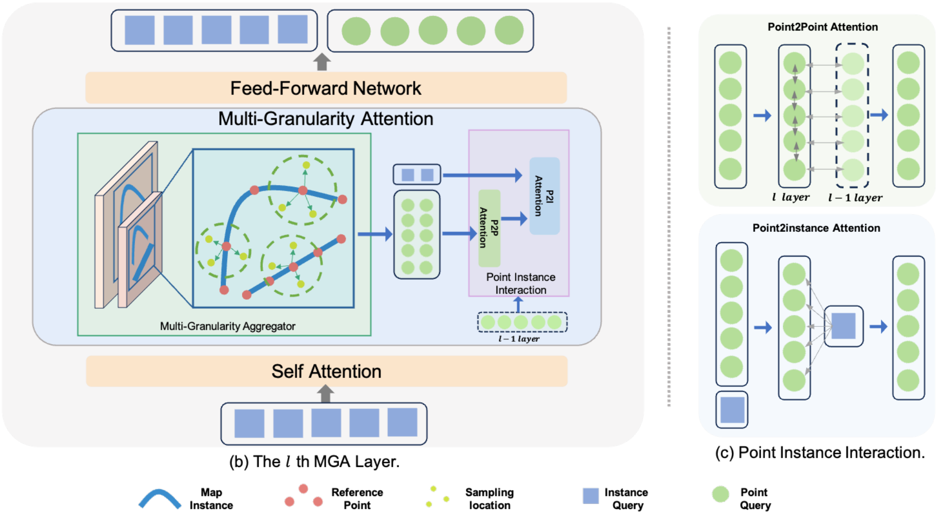
### 3.1 OVERALL ARCHITECTURE

The overall network architecture of MGMapNet is depicted in Fig. 2 (a). Similar to other DETR-like end-to-end HD map construction models, MGMapNet comprises a BEV Feature Encoder, responsible for extracting multi-scale BEV features from perspective view images, and a Transformer Decoder, which stacks multiple layers of Multi-Granularity Attention to generate predictions for map elements. The prediction from each layer encapsulates both category and geometry information within the perception range.

**BEV Feature Encoder** The model takes surrounding-view RGB images as inputs, expressing them as unified perceptual BEV feature representation for subsequent transformer decoder. The unified BEV feature is denoted as $\mathbf{F}_{bev} \in \mathbb{R}^{C\times H\times W}$ , where $C, H, W$ represent the feature channels, height, and width of the BEV feature, respectively. Given the diverse lengths of map elements, relying solely on a single-scale BEV feature fails to meet the requirements for detecting all elements with different lengths. Therefore, we employ downsample modules to reduce the spatial resolution of BEV features $\mathbf{F}_{bev}$ by half, generating $\mathbf{F}^{'}_{bev} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$. More scales might be benificial, but we found two scales are already good enough. Both $\mathbf{F}_{bev}$ and $\mathbf{F}^{'}_{bev}$ are utilized in the decoder afterwards. $\mathbf{F}_{ms\_bev} \in \mathbb{R}^{C\times(\frac{H}{2}\times\frac{W}{2}+H\times W)}$ represents multi-scale BEV features, which are obtained by concatenating the flattened tensors of $\mathbf{F}_{bev}$ and $\mathbf{F}^{'}_{bev}$. As will be shown in the experiments in the following section, the multi-scale BEV features greatly improve the overall performance.



(a) The overall architecture of MGMapNet.



(b) The $l$ th MGA Layer.

(c) Point Instance Interaction.

Figure 2: Overview of the MGMapNet. (a) The overall architecture of MGMapNet starts with multi-view image inputs, which are processed through an encoder and decoder to output vectorized map representations. (b) A schematic diagram of the $l$ th MGA layer. The figure depicts the interactions within the Multi-Granularity Attention, including Multi-Granularity Aggregator and Point Instance Interaction. (c) Implementation details of Point Instance Interaction, consisting of P2P attention and P2I attention.

**Decoder** The decoder has $L$ layers. Each layer is composed by a Self-Attention, a Multi-Granularity Attention and a Feed-Forward Network as shown in Fig. 2 (b). Multi-Granularity Attention consists of two components: Multi-Granularity Aggregator and Point Instance Interaction. The instance-level query is initialized by learnable parameters which are updated by querying on BEV features, and the point query is generated dynamically by aggregating BEV features. After that, a Point Instance Interaction is employed to carry out the mutual interaction between local geometry information and the global category information. The details of Multi-Granularity Attention is described in the following section.

### 3.2 MULTI-GRANULARITY ATTENTION

Instance-level queries effectively capture the overall categorical information of road elements but may struggle to accurately represent geometric details, particularly for irregularly shaped or elongated map features. Conversely, point-level queries provide rich, detailed information; however, they can only represent instances by aggregating multiple queries, resulting in a lack of comprehensive descriptions for map elements. To simultaneously capture both detailed and comprehensive instance features, the Multi-Granularity Attention mechanism is designed to effectively maintain and update queries at various granularities. As illustrated in Fig. 2, the Multi-Granularity Attention comprises two primary components: the Multi-Granularity Aggregator and Point Instance Interaction.

#### 3.2.1 MULTI-GRANULARITY AGGREGATOR

In Multi-Granularity Aggregator, instance-level queries interact with the multi-scale BEV features and point-level queries are generated. Specifically, we improve Mutli-Head Deformable Attention (Zhu et al., 2020) with multiple reference points for each query to aggregate long-range features from multi-scale BEV features. To improve readability, we omit the subscript $m$ for the index of multiple heads $M$ in the operator of Multi-Granularity Aggregator.

More specifically, the Multi-Granularity Aggregator takes as input the instance-level queries $\mathbf{Q}_{ins} \in \mathbb{R}^{N_q \times C}$ in the first layer, along with the point-level queries $\mathbf{Q}_{pts} \in \mathbb{R}^{N_q \times N_p \times C}$ and the reference points $\mathbf{RF} \in \mathbb{R}^{N_q \times N_p \times 2}$ in the subsequent layers. $N_q$ is the total number of instance-level queries, and $N_p$ is the total number of points belonging to an instance. Noted that the reference points in the first layer are predicted by $\mathbf{Q}_{ins}$ and the reference points in subsequent layers are updated by reference points from the previous layer in the form of Eq. 1.

$$\begin{cases} \mathbf{RF}^l = \mathrm{MLP}(\mathbf{Q}^l_{ins}), l = 0 \\ \mathbf{RF}^l = \mathrm{sigmoid}(\mathrm{sigmoid}^{-1}(\mathbf{RF}^{l-1}) + \mathrm{MLP}(\mathbf{Q}^l_{pts})), l >= 1 \end{cases} \tag{1}$$

where $l$ represents the current $l$-th layer, sigmoid, $\mathrm{sigmoid}^{-1}$ refers to the sigmoid and inverse sigmoid activation function and MLP stands for Multi-Layer Perceptron layer.

Since an instance is represented as a point sequence, the position encoding is added to the instance-level query. Given the location of reference point $\mathbf{RF}$, we employ $\mathbf{RF}$ to generate positional encoding $\mathbf{PE}_{ref}$:

$$\mathbf{PE}^{l-1}_{ref} = \mathrm{MLP}^{l-1}_{ref}(\mathbf{RF}^{l-1}), \tag{2}$$

where $\mathrm{MLP}^{l-1}_{ref}$ is a projection layer used to generate the positional embedding from reference points.

We allocate $N_{rep}$ sampling points to each reference point where the features are aggregated from to enhance the feature for the reference point. The location offset $\Delta\mathbf{S}$ of sampling points w.r.t the reference point and the associated weights $\mathbf{W}$ are computed by combining the instance-level queries $\mathbf{Q}_{ins}$ and $\mathbf{PE}_{ref}$ as follows:

$$\begin{aligned} \Delta\mathbf{S}^l &= \mathrm{Sampling\_Offset}(\mathbf{Q}^{l-1}_{ins} + \mathbf{PE}^{l-1}_{ref}) \in \mathbf{R}^{N_q \times N_p \times N_{rep} \times 2}, \\ \mathbf{W}^l &= \mathrm{Weight\_Embed}(\mathbf{Q}^{l-1}_{ins} + \mathbf{PE}^{l-1}_{ref}) \in \mathbb{R}^{N_q \times N_p \times N_{rep}}, \\ \mathbf{S}^l &= (\mathbf{RF}^{l-1} + \Delta\mathbf{S}^l) \in \mathbb{R}^{N_q \times N_p \times N_{rep} \times 2}, \end{aligned} \tag{3}$$

where $\mathbf{RF}^{l-1}$ is expanded accordingly to match the shape of $\Delta\mathbf{S}^l$. By utilizing the sampling offset and the reference point, sampling locations $\mathbf{S}^l$ are updated by adding $\mathbf{RF}^{l-1}$ and $\Delta\mathbf{S}^l$.

Subsequently, $\mathbf{Q}_{ins}$ and $\mathbf{Q}_{pts}$ are generated by the weighted sum of sampled features:

$$\begin{aligned} \mathbf{W}^l_{ins} &= \underset{(j,k)\in(N_p,N_{rep})}{\mathrm{softmax}} \left(\mathbf{W}^l_{j,k}\right) \in \mathbb{R}^{N_q \times (N_p \times N_{rep})}, \\ \mathbf{W}^l_{pts} &= \underset{k \in N_{rep}}{\mathrm{softmax}} \left(\mathbf{W}^l_{j,k}\right) \in \mathbb{R}^{N_q \times N_p \times N_{rep}}, \\ \mathbf{Q}^l_{ins} &= \sum_{j=1}^{N_p} \sum_{k=1}^{N_{rep}} \left[\mathbf{W}^l_{ins}\,\mathrm{sampling}(\mathbf{F}_{\mathrm{ms\_bev}}, \mathbf{S}^{\mathrm{l}}_{\mathrm{j,k}})\right] \in \mathbb{R}^{N_q \times C}, \\ \mathbf{Q}^l_{pts} &= \sum_{k=1}^{N_{rep}} \left[\mathbf{W}^l_{pts}\,\mathrm{sampling}(\mathbf{F}_{\mathrm{ms\_bev}}, \mathbf{S}^{\mathrm{l}}_{\mathrm{j,k}})\right] \in \mathbb{R}^{N_q \times N_p \times C}, \end{aligned} \tag{4}$$

where $j$ is the index for the $N_p$ points on an instance, $k$ is the index among the $N_{rep}$ sampling points assigned to the reference point, $\mathbf{W}^l_{ins}$, $\mathbf{W}^l_{pts}$ denotes the softmax normalized weight across $N_p \times N_{rep}$, $N_{rep}$ of $\mathbf{W}^l_{j,k}$, respectively and sampling denotes the bi-linear sampling operator.

Through Multi-Granularity Aggregator, $\mathbf{Q}_{ins}$ and $\mathbf{Q}_{pts}$ are generated from multi-scale BEV features, capturing both global and local information for each map element. Compared with Multi-Point Attention as proposed in StreamMapNet (Yuan et al., 2024), our method incorporates point-level queries directly from the multi-scale BEV features by sampling points, which enhances the accuracy of predicted geometry points. In addition, compared with point-level alone representations such as MapTR (Liao et al., 2022) and MapTRv2 (Liao et al., 2023), our model updates instance-level queries with sampled point features, which effectively captures the overall category and shape information of road elements.

### 3.2.2 POINT INSTANCE INTERACTION

The Point Instance Interaction is designed with the intention of enhancing positional and categorical information interaction between two different granularities of queries. As illustrated in Fig. 2(c), Point Instance Interaction comprises two distinct attention operators: P2P (point-to-point) attention and P2I (Point-to-Instance) attention.

Concurrently, the sampling locations $\mathbf{S}^l$ and attention weights $\mathbf{W}_{ins}^l, \mathbf{W}_{pts}^l$ obtained from Multi-Granularity Aggregator in the $l$-th layer are flattened and concatenated to encode positional information in P2P Attention and P2I Attention:

$$\begin{aligned} \mathbf{PE}_{ins}^l &= \mathrm{MLP}_{ins}^l(\mathbf{S}^l, \mathbf{W}_{ins}^l), \\ \mathbf{PE}_{pts}^l &= \mathrm{MLP}_{pts}^l(\mathbf{S}^l, \mathbf{W}_{pts}^l), \end{aligned} \tag{5}$$

where $\mathrm{MLP}_{ins}^l$ and $\mathrm{MLP}_{pts}^l$ are MLPs for instance-level queries and point-level queries respectively. $\mathbf{PE}_{ins}, \mathbf{PE}_{pts}$ are the corresponding generated position embedding.

**P2P Attention** As the coordinates of map elements are refined based on the point-level queries in previous Multi-Granularity Attention layer, these point-level queries play a pivotal role in predicting coordinates in the current layer. Hence, the P2P Attention module is devised to include point-level queries from both the current $l$-th layer and previous $(l-1)$-th layer as inputs of the attention layer. Formally:

$$\begin{cases} \mathbf{Q}_{pts}^{l'} = \mathrm{SA}(\mathbf{Q}_{pts}^l + \mathbf{PE}_{pts}^l), l = 0 \\ \mathbf{Q}_{pts}^{l'} = \mathrm{CA}(\mathbf{Q}_{pts}^l + \mathbf{PE}_{pts}^l, \mathbf{Q}_{pts}^{l-1} + \mathbf{PE}_{pts}^{l-1}), l >= 1 \end{cases} \tag{6}$$

It is important to note that since the first Multi-Granularity Attention layer does not have previous decoder layer, self-attention operation only conducts in the current point-level queries. And in the P2P Attention of following Multi-Granularity Attention layer, previous point-level queries $\mathbf{Q}_{pts}^{l-1}$ and current generated point-level queries $\mathbf{Q}_{pts}^l$are mixed before P2P Attention.

**P2I Attention** Following the P2P Attention, P2I Attention operation achieves information interaction among different granularities. Point-level queries exchange geometry information with instance-level queries using cross-attention:

$$\mathbf{Q}_{pts}^{l''} = \mathrm{CA}(\mathbf{Q}_{pts}^{l'} + \mathbf{PE}_{pts}^l, \mathbf{Q}_{ins}^l + \mathbf{PE}_{ins}^l). \tag{7}$$

Finally, point-level queries belonging to the same instance-level queries are aggregated to update corresponding instance-level queries as follows:

$$\mathbf{Q}_{ins}^{l'} = \mathrm{MLP}_{agg}^l(\sum_{j=1}^{N_p} \mathbf{Q}_{pts,j}^{l''}), \tag{8}$$

where $j$ represents the index of $N_p$.

**Output** Ultimately, point-granularity queries are utilized for point location prediction using a MLP as the regression head, while instance-granularity queries are employed for predicting the categories of map elements using another MLP. In summary, by utilizing the Multi-Granularity Aggregator and Point Instance Interaction, Multi-Granularity queries are generated and updated. Meanwhile, the geometry and category of each map element can be effectively perceived.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETTINGS

**nuScenes Dataset.** nuScenes (Caesar et al., 2020) is a widely recognized dataset in the field of autonomous driving research, providing 1,000 scenes, each captured over a continuous 20-second

Table 1: Comparison to the state-of-the-art on nuScenes val set.

| Methods | Epoch | $\text{AP}_{ped}$ | $\text{AP}_{div}$ | $\text{AP}_{bou}$ | mAP | FPS | Params. |
|---|---|---|---|---|---|---|---|
| HDMapNet (Li et al., 2022a) | 30 | 14.4 | 21.7 | 33.0 | 23.0 | - | - |
| BeMapNet (Qiao et al., 2023a) | 30 | 62.3 | 57.7 | 59.4 | 59.8 | 4.3 | - |
| PivotNet (Ding et al., 2023a) | 24 | 56.5 | 56.2 | 60.1 | 57.6 | 9.2 | - |
| MapTRv2 (Liao et al., 2023) | 24 | 59.8 | 62.4 | 62.4 | 61.5 | 14.1 | 40.3 |
| MGMap (Liu et al., 2024a) | 24 | 61.8 | 65.0 | 67.5 | 64.8 | 12 | 55.9 |
| MapQR (Liu et al., 2024b) | 24 | **68.0** | 63.4 | 67.7 | 66.4 | 11.9 | 125.3 |
| **MGMapNet** | 24 | 64.7 | **66.1** | **69.4** | **66.8** | 11.7 | 70.1 |
| VectorMapNet (Liu et al., 2023a) | 110 | 42.5 | 51.4 | 44.1 | 46.0 | - | - |
| MapTRv2 (Liao et al., 2023) | 110 | 68.1 | 68.3 | 69.7 | 68.7 | 14.1 | 40.3 |
| MGMap (Liu et al., 2024a) | 110 | 64.4 | 67.6 | 67.7 | 66.5 | 12 | 55.9 |
| MapQR (Liu et al., 2024b) | 110 | **74.4** | 70.1 | 73.2 | 72.6 | 11.9 | 125.3 |
| **MGMapNet** | 110 | 74.3 | **71.8** | **74.8** | **73.6** | 11.7 | 70.1 |

Table 2: Performance comparison on NuScenes with IoU-based AP.

| Methods | $AP^{raster}_{ped}$ | $AP^{raster}_{div}$ | $AP^{raster}_{bou}$ | $mAP^{raster}$ |
|---|---|---|---|---|
| MapVR (Zhang et al., 2024) [NeurIPS2023] | 46.0 | 39.7 | 29.9 | 38.5 |
| MGMap (Liu et al., 2024a) [CVPR2024] | **54.5** | 42.1 | 37.4 | 44.7 |
| MGMapNet(ours) | 54.0 | **42.7** | **44.1** | **46.9** |

interval. Each dataset sample incorporates data from six synchronized RGB cameras and includes detailed pose information. The perception ranges extend from $-15.0m$ to $15.0m$ along the X-axis and from $-30.0m$ to $30.0m$ along the Y-axis. For experimental purposes, the dataset is partitioned into 700 scenes comprising 28,130 samples for training purposes, and 150 scenes containing 6,019 samples designated for validation.

**Argoverse2 Dataset.** The Argoverse2 dataset (Wilson et al., 2023) contains multimodal data from 1000 sequences, including high-resolution images from seven ring cameras and two stereo cameras, as well as LiDAR point clouds and map-aligned 6-DoF pose data. All annotations are densely sampled to facilitate the training and evaluation of 3D perception models. Results are reported on the validation set, with a focus on the same three map categories as identified in the nuScenes dataset.

**Evaluation Metric.** In alignment with MapTR (Liao et al., 2022), we have adopted the widely-accepted metric of mean Average Precision (mAP), predicated on the Chamfer distance, a measure frequently employed in HD map construction task. Evaluation thresholds are set at 0.5m, 1.0m, and 1.5m. Specifically, $AP_{ped}$ and $AP_{div}$, and $AP_{bou}$ refer to the Average Precision for pedestrians, dividers, and boundaries, respectively.

**Auxiliary Loss.** To ensure that each instance exhibits a more reasonable and accurate distribution of map instances, we introduce new auxiliary losses $\mathcal{L}_{aux}$, comprising instance segmentation loss $\mathcal{L}_{ins\_seg}$ and reference point loss $\mathcal{L}_{ref}$ to further improve performance. The implementation details are provided in the appendix A.2. In addition, we incorporate point loss $\mathcal{L}_{pts}$, classification loss $\mathcal{L}_{cls}$, edge direction loss $\mathcal{L}_{dir}$ and dense prediction losses $\mathcal{L}_{dense}$ same with MapTRv2 (Liao et al., 2023).

To increase the precision of the sampling locations within Multi-Granularity Aggregator, we add a reference loss $\mathcal{L}_{ref}$ to supervise the sampling points. Besides, to improve the spatial details of instance-level queries, we generate instance BEV segmentation gt mask $M_{bev}$ to supervise instance-level segmentation prediction. Hungarian matching results are utilized to eliminate negative instance-level queries' segmentation prediction. The instance segmentation loss, denoted as $\mathcal{L}_{ins\_seg}$, is formulated as an ensemble of the cross-entropy loss and the dice loss.

The final loss is defined as the weighted sum of the above losses:

$$\mathcal{L} = \beta_1 \mathcal{L}_{pts} + \beta_2 \mathcal{L}_{cls} + \beta_3 \mathcal{L}_{dir} + \beta_4 \mathcal{L}_{dense} + \beta_5 \mathcal{L}_{aux} \tag{9}$$

**Implementation Details.** Our model is trained on 8 A100 GPUs with batchsize as 2, utilizing an AdamW optimize (Loshchilov & Hutter, 2018) with a learning rate of $4 \times 10^{-4}$. We adopt the ResNet50 (He et al., 2016) as our backbone and employ a LSS transformation (Philion & Fidler, 2020) with a single encoder layer for feature extraction. We also adopt the one-to-many training strategy, consistent with MapTRv2 (Liao et al., 2023). The model trains for 24 epochs on the nuScenes dataset and 6 epochs on Argoverse2 dataset. We conduct a long training schedule (110 epochs) on the nuScenes dataset for a fair comparison with previous methods. We set $N_q = 100$,$N_{rep}$ = 8, $N_p = 20$, $\beta_1 = 5$, $\beta_2 = 2$, $\beta_3 = 0.005$, $\beta_4 = 3$, $\beta_5 = 3$ as the hyperparameters for all settings without further tuning.

### 4.2 COMPARISONS WITH STATE-OF-THE-ART METHODS

**Results on nuScenes.** The default evaluation metric for Vectorized HD Map construction is the Chamfer Distance Average Precision (AP). Tab. 1 presents the results on the nuScenes validation dataset, utilizing multi-view RGB images as input. In comparison to the SOTA method MapTRv2, our MGMapNet has reached an mAP of 66.8, exceeding it by 5.3 mAP with a training duration 24 epochs. After a prolonged training period of 110 epochs, MGMapNet achieved 73.6 mAP, which is still significantly higher than the 68.7 mAP of MapTRv2 by 4.9 mAP and 72.6 mAP of MapQR by 1.0 mAP respectively.

The latest models also use rasterization results and employ IoU-based Average Precision (AP) to evaluate reconstruction performance. As shown in Tab. 2. We evaluate MGMapNet, which achieves an mAP of 46.9, surpassing both MapVR and MGMap in terms of IoU-based AP.

Experimental results substantiate that the proposed multi-granularity representation, which models both local point information and global instance information, significantly enhances predictive performance in both rasterization and vectorization evaluation metrics.

Qualitative results are depicted in Fig. 3. We select three complex scenarios: daytime vehicles with occlusion, nighttime low-light conditions, and low-light situations with occlusion. In the first case, MGMapNet exhibits more precise coordinate predictions compared to StreamMapNet and preserves all road elements compared to MapTRv2. In the second case of nighttime low-light conditions, MapTRv2 struggles to predict the divider on the right side of the vehicle due to its lack of instance-level perception. While StreamMapNet utilizes instance-level queries and identifies the divider, its overall instance positioning accuracy remains inadequate. In contrast, only MGMapNet accurately and completely detects the boundary in these challenging conditions. The third case of nighttime dense vehicle traffic with occlusion highlights StreamMapNet's poor detection performance. MapTRv2 encounters two major issues: mislocating the pedestrian path on the right front and misclassifying the rear divider as a boundary, indicating its limitations in instance-level perception. Conversely, MGMapNet exhibits remarkable robustness, accurately predicting both categories and locations even in low-light conditions and substantial nighttime occlusion.

Qualitative results demonstrate that the proposed MGMapNet effectively mitigates the shortcomings associated with both instance-level and point-level queries, achieving superior accuracy in HD map construction under complex conditions.

**Results on Argoverse2.** On the more complex Argoverse2 dataset, the performance of MGMapNet remains competitive. Tab.3 presents our results on the Argoverse2 validation dataset for 6 epochs. The Argoverse2 dataset provides two configurations for the representation of points: 2D and 3D point coordinates. We conduct experiments on both configurations and achieve mAP scores of 71.2 and 69.1 mAP in 6 epochs, respectively. This represents an improvement of 3.8 and 4.4 mAP respectively comparing with MapTRv2. Compared to the latest HIMap, which achieves 69.6 and 68.4 mAP in 2D and 3D configurations respectively, MGMapNet still surpasses it by 1.6 and 0.7 mAP. The results from other methods are sourced from the original paper, and the experimental results demonstrate the competitiveness of MGMapNet.

**Efficiency comparison.** We conduct a comprehensive efficiency analysis of several open-source models, focusing primarily on frames per second (FPS) and model parameters to substantiate the efficacy of the models. As demonstrated in the last two columns of Tab. 1, our model achieves an FPS of 11.7, which is comparable to the latest models, MapQR and MGMap. It is slightly lower

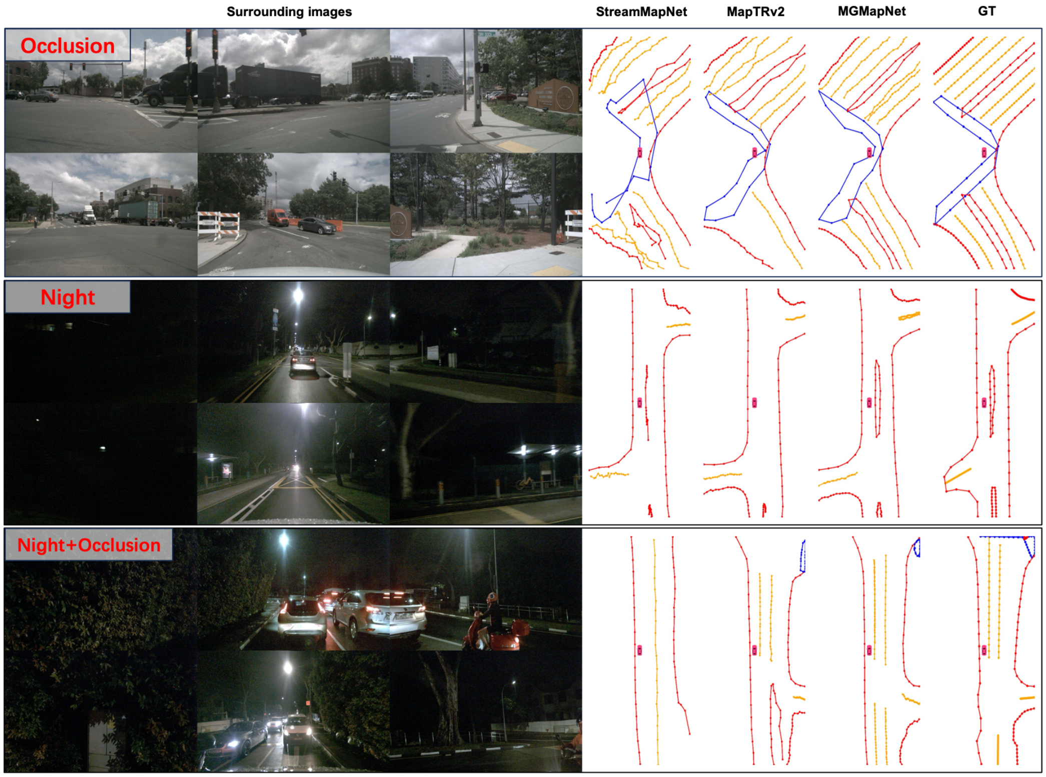


Figure 3: Comparison with SOTAs on qualitative visualization on nuScenes val set.

Table 3: Comparison to the state-of-the-art on Argoverse2 val set.

| Methods | Map dim. | $\text{AP}_{ped}$ | $\text{AP}_{div}$ | $\text{AP}_{bou}$ | mAP |
|---|---|---|---|---|---|
| HDMapNet (Li et al., 2022a) | 2 | 13.1 | 5.7 | 37.6 | 18.8 |
| VectorMapNet (Liu et al., 2023a) | | 38.3 | 36.1 | 39.2 | 37.9 |
| MapTRv2 (Liao et al., 2023) | | 62.9 | 72.1 | 67.1 | 67.4 |
| MapQR (Liu et al., 2024b) | | 64.3 | 72.3 | 68.1 | 68.2 |
| HIMap (Zhou et al., 2024) | | **69.0** | 69.5 | 70.3 | 69.6 |
| **MGMapNet** | | 67.1 | **74.6** | **71.7** | **71.2** |
| VectorMapNet (Liu et al., 2023a) | 3 | 36.5 | 35.0 | 36.2 | 35.8 |
| MapTRv2 (Liao et al., 2023) | | 60.7 | 68.9 | 64.5 | 64.7 |
| MapQR (Liu et al., 2024b) | | 60.1 | 71.2 | 66.2 | 65.9 |
| HIMap (Zhou et al., 2024) | | **66.7** | 68.3 | 70.3 | 68.4 |
| **MGMapNet** | | 64.7 | **72.1** | **70.4** | **69.1** |

than MapTRv2 while outperforming methods like PivotNet. The model parameters are 70.1 MB, which is lower than MapQR's 120.3 MB but slightly higher than MGMap's 55.9 MB.

### 4.3 ABLATION STUDY

We conduct ablation experiments on the nuScenes validation dataset under the 24 epoch training setting, examining the effectiveness of the Multi-Granularity Attention, as well as the incremental impact of the strategy optimization on the model's performance. The influence of each component in MGMapNet is demonstrated in Tab. 4.

**Multi-Granularity Attention.** Tab.4 illustrates the comparison between MPA and MGA, as well as Point Instance Interaction. We initially employed MPA as the fundamental module with strat-

Table 4: Effectiveness of key designs in Multi-Granularity Attention.

| Method | Point Instance Interaction | | mAP |
|---|---|---|---|
| | P2P attention | P2I attention | |
| Multi-Point Attention | - | - | 59.6 |
| Multi-Granularity Attention | - | - | 62.7 |
| | ✓ | - | 64.8 |
| | - | ✓ | 65.0 |
| | ✓ | ✓ | 66.8 |

Table 5: Ablation study of each optimization strategy.

| Exp. | Method | mAP |
|---|---|---|
| | Multi-Point Attention | 55.9 |
| (a) | Multi-Granularity Attention | 63.6 (+7.7) |
| (b) | + Aux. Loss | 64.4 (+0.8) |
| (c) | + Multi-Scale | 65.0 (+0.6) |
| (d) | + Reference Point PE | 66.2 (+1.2) |
| (e) | + Add Query Number | 66.8 (+0.6) |

egy optimizations, achieving an mAP of 59.6. By replacing MPA with MGA and introducing more appropriate queries, it captures fine-grained point and coarse-grained instance features. This enhancement facilitates a more nuanced and precise perception, ultimately achieving 66.8 mAP and leading to **7.2** improvement in mAP. Additionally, only using the Multi-Granularity Aggregator, the mAP is 62.7, indicating that the multi-granularity representation has led to an mAP increase of 3.1 compared to 59.6 for MPA. Further, when the P2P and P2I attention are introduced in the Point Instance Interaction, the mAP increased by 2.1 and 2.3 respectively, reaching 64.8 and 65.0. The simultaneous application of these improvements has boosted the model's performance to 66.8 mAP, an increase of **4.1** mAP. This highlights the significance of both attention modules in enhancing the intrinsic relationships between the two granularities and improving model performance.

**Strategy Optimization.** As shown in Tab. 5, we also investigate the effectiveness of other strategies used in MGMapNet. Experiment (a) demonstrates that Multi-Granularity Attention, as a replacement for Multi-Point Attention, achieves a mAP of 63.6, resulting in a 7.7 mAP increase compared to the 55.9 mAP of Multi-Point Attention. Meanwhile, Experiment (b) reveals that the inclusion of the auxiliary loss results in an improvement of the mAP by 0.8. Experiments (c), (d), and (e) illustrate the effectiveness of using multi-scale approaches, adding the reference point positional encoding, and increasing the number of queries, which yield gains of 0.6, 1.2, and 0.6 mAP, respectively. By optimizing with these strategies, our MGMapNet achieves 66.8 mAP, representing state-of-the-art performance.

## 5 CONCLUSION AND DISCUSSION

In this paper, multi-granularity representation is proposed, enabling the end-to-end vectorized HD Map construction using coarse-grained instance-level and fine-grained point-level queries. Through the designed Multi-Granularity Attention, category and geometry information is exchanged. Our proposed MGMapNet has achieved state-of-the-art (SOTA) single-frame performance on both the nuScenes and Argoverse2 datasets.

However, our primary focus is on improving the quality of HD Map Construction. Addressing real-time performance is a promising direction for future optimization. In addition, exploring some temporal approaches as priors is also a direction worth considering. The mechanism of Multi-Granularity Attention is generic, and it is worth trying to determine its effectiveness in topological prediction or other autonomous driving tasks.

## REFERENCES

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*, pp. 11621–11631, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pp. 550–567. Springer, 2022.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. 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, October 2023a.

Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3672–3682, 2023b.

Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, and Carlos Vallespi-Gonzalez. Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2349–2357, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15273–15282, 2021.

Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17451–17460, 2023.

Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 4628–4634, 2022a. doi: 10.1109/ICRA46639.2022.9812383.

Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. Lanesegnet: Map learning with lane segment perception for autonomous driving. *arXiv preprint arXiv:2312.16108*, 2023.

Xiang Li, Jun Li, Xiaolin Hu, and Jian Yang. Line-cnn: End-to-end traffic line detection with line proposal unit. *IEEE Transactions on Intelligent Transportation Systems*, 21(1):248–258, 2019.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pp. 1–18. Springer, 2022b.

Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 641–656, 2018.

Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7345–7353, 2019.

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.

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*, 2023.

Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. Mgmap: Mask-guided learning for online vectorized hd map construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14812–14821, 2024a.

Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pp. 22352–22369. PMLR, 2023a.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. IEEE, 2023b.

Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. *arXiv preprint arXiv:2402.17430*, 2024b.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

Jonah Philion and Sanja 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*, pp. 194–210. Springer, 2020.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7077–7087, 2021.

Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi 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 (CVPR)*, pp. 13218–13228, June 2023a.

Limeng Qiao, Yongchao Zheng, Peng Zhang, Wenjie Ding, Xi Qiu, Xing Wei, and Chi Zhang. Machmap: End-to-end vectorized solution for compact hd-map construction. *arXiv preprint arXiv:2306.10301*, 2023b.

Juyeb Shin, Francois Rameau, Hyeonjun Jeong, and Dongsuk Kum. Instagram: Instance-level graph modeling for vectorized hd map learning. *arXiv preprint arXiv:2301.04470*, 2023.

Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 294–302, 2021a.

Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Polylanenet: Lane estimation via deep polynomial regression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6150–6156. IEEE, 2021b.

Ruihao Wang, Jian Qin, Kaiying Li, Yaochen Li, Dong Cao, and Jintao Xu. Bev-lanedet: An efficient 3d lane detection based on virtual camera via key-points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1002–1011, June 2023.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1): 537–547, 2020.

Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Neural map prior for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17535–17544, 2023.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.

Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.

Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7356–7365, 2024.

Gongjie Zhang, Jiahao Lin, Shuang Wu, Zhipeng Luo, Yang Xue, Shijian Lu, Zuoguan Wang, et al. Online map vectorization for autonomous driving: A rasterization perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Tu Zheng, Yifei Huang, Yang Liu, Wenjian Tang, Zheng Yang, Deng Cai, and Xiaofei He. Clrnet: Cross layer refinement network for lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 898–907, 2022.

Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. Himap: Hybrid representation learning for end-to-end vectorized hd map construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15396–15406, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

# A APPENDIX

In this appendix material, we provide additional analysis of the proposed MGMapNet, including:

- Visualization.
- Description of the auxiliary loss function.
- Efficiency comparisons.
- Supplementary experiments.
- Broader Impact Statement.

## A.1 VISUALIZATION

We have visualized some comparative results in the supplementary material, as shown in Fig. 4 for the nuScenes dataset and Fig. 5 for the Argoverse2 dataset.

In the visualization for the nuScenes dataset, from left to right, each column corresponds to the surround view images, StreamMapNet, MapTRv2, and the proposed MGMapNet, respectively. From top to bottom, each row represents a sample. It can be observed that in the initial three samples, MapTRv2 was unable to detect all the pedestrian crossings completely. In the fourth sample, the structure of the pedestrian crossing is inaccurate. Meanwhile, the results from StreamMapNet indicate that despite the detection of the majority of map instances, the instability of the shapes compromises their ability to accurately detect map elements.

Similar phenomena can also be observed in the visualization for the Argoverse2 dataset. We can see that in the first and second examples, MapTRv2 missed the middle boundary of the road and the pedestrian crossing in front of the vehicle respectively, while MGMapNet successfully detected both. In the third and fourth samples, MapTRv2 did not correctly detect the exit on the left side and instead interpreted it as a continuous boundary. In the fifth example, in poor lighting conditions, MGMapNet successfully detects the boundary on the left.

In summary, our MGMapNet demonstrates superior performance compared to MapTRv2 on the quality of HD Map Construction.

## A.2 DESCRIPTION OF THE AUXILIARY LOSS FUNCTION.

We introduce auxiliary losses, which comprise two components: the instance segmentation loss $\mathcal{L}_{ins\_seg}$ and the reference point loss $\mathcal{L}_{ref}$. The other $\mathcal{L}_{pts}$, $\mathcal{L}_{cls}$, $\mathcal{L}_{dir}$ and $\mathcal{L}_{dense}$ loss align with MapTRv2.

The instance segmentation loss, denoted as $\mathcal{L}_{ins_{s}eg}$, not only segments BEV features but also retrieves more precise instance localization information for each individual query. First, we compute the instance segmentation masks $M^{pred} \in \mathbb{R}^{H \times W \times N_q}$ ($N_q$ is the total number of instance-level queries) by performing dot product operations between the updated instance-level queries $\mathbf{Q}_{ins} \in \mathbb{R}^{N_q \times C}$ and the BEV features $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$. Subsequently, we utilize the indices of positive samples obtained through the Hungarian algorithm to retrieve their corresponding masks $M^{pred}_{pos}$ and ground truths $M^{gt}_{pos}$. For each positive sample instance mask $M^{pred}_{pos} \in \mathbb{R}^{H \times W \times N_{pos}}$ ($N_{pos}$ is the total number of positive query), we separately compute the segmentation loss by employing both Binary Cross-Entropy loss $\mathcal{L}_{bce}$ and Dice loss $\mathcal{L}_{Dice}$.

The process of generating the $M_{pos}$ is formulated as:

$$M^{pred} = \mathbf{F} \cdot \mathbf{Q}^{T}_{ins},$$

where $\cdot$ denote dot product operations and the $\mathcal{L}_{ins\_seg}$ is formulated as:

$$\mathcal{L}_{ins\_seg} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} (\mathcal{L}_{dice}(M^{pred}_{pos,i}, M^{gt}_{pos,i}) + \mathcal{L}_{bce}(M^{pred}_{pos,i}, M^{gt}_{pos,i})),$$

where $M_i$ denote the $i$-th positive instance mask.

Additionally, the reference point loss $\mathcal{L}_{ref}$ provides auxiliary supervision for reference points during each iteration of the decoder. Similar to the $\mathcal{L}_{pts}$ loss, The $\mathcal{L}_{ref}$ is computed by applying the $\mathcal{L}_{pts}$ loss to the reference points $\mathbf{RF}$ and ground truth points $\mathbf{P}^{gt}$ at each layer. This ensures that each sampling point achieves a more reasonable and accurate distribution.

### A.3 EFFICIENCY COMPARISONS.

Table 6: Efficiency comparison with recent methods.

| | MapTR[ICLR2023] | MapTRv2[IJCV2024] | MGMap[CVPR2024] | MapQR[ECCV2024] | MGMapNet |
|---|---|---|---|---|---|
| FPS | **16.9** | 14.1 | 12.0 | 11.9 | 11.7 |
| GPU mem. (MB) | **2314** | 2656 | 2402 | 2648 | 2790 |
| Params (MB) | **35.9** | 40.3 | 55.9 | 125.3 | 70.1 |
| NuScenes mAP | 50.3 | 61.5 | 64.8 | 66.4 | **66.8** |
| Argoverse2 mAP | 58 | 67.4 | - | 68.2 | **71.2** |

In Table 6, we present a comprehensive comparison of the latest models alongside the primary baseline, detailing GPU memory usage, FPS, parameter counts, and performance. Time and space complexity can be derived from FPS and GPU memory comparisons.

**GPU mem. comparison.** The memory usage (MB) of MapTR, MapTRv2, MGMap, MapQR, and MGMapNet are 2314, 2656, 2402, 2648, and 2790 respectively. Our MGMapNet has a slight increase in memory usage compared to other methods, which is understandable given we retained two types of queries for different output regressions and classifications.

**FPS comparison.** MGMapNet, MapQR, and MGMap show similar performance with FPS scores of 11.7, 11.9, and 12, respectively. Although slightly slower than MapTRv2, MGMapNet's inference time complexity is similar to that of the latest methods.

**Params comparison.** The parameters (MB) of MGMapNet, MapQR, and MGMap are 70.1, 125.3, and 55.9, respectively. Even though MGMapNet has a slightly higher parameter count due to its Multi-Granularity query design and Point Instance Interaction, it still outperforms and has fewer parameters than MapQR's 125MB. We believe there's substantial room for optimization in MGMapNet.

**Performance comparison.** After training for 24 epochs on the NuScenes dataset, MGMapNet achieved a mean average precision (mAP) of 66.8. On the Argoverse2 dataset, it reached an mAP of 71.2 after just 6 epochs. This demonstrates that MGMapNet maintains similar speed while achieving better accuracy across different datasets.

In an overall efficiency analysis, our MGMapNet, thanks to its multi-granularity representation, achieves better performance while maintaining similar parameters, speed, and memory usage compared to the latest methods. The limitations of our method in terms of speed have been mentioned, but we believe there is a significant room for optimization. Therefore, MGMapNet remains a competitive model.

### A.4 SUPPLEMENTARY EXPERIMENTS.

Table 7: Influnence of repeat number $N_{rep}$, the $N_{rep}$ is set as 8.

| Number | AP$_{ped}$ | AP$_{div}$ | AP$_{bou}$ | mAP |
|---|---|---|---|---|
| 4 | 61.5 | 62.3 | 67.3 | 63.7 |
| 8 | 64.7 | 66.1 | 69.4 | 66.8 |
| 12 | 62.8 | 63.6 | 67.9 | 64.8 |

**Hyperparameter Experimentation.** We conducted experiments on the hyperparameters within the model, including those for $N_q$ and $N_{rep}$, as shown in Tab. 7 and Tab. 8. The parameter $N_q = 100$ and $N_{rep} = 8$ that we selected represents the optimal configuration. We believe that too few points are insufficient to describe local detail, while too many points can increase learning difficulty and reduce performance.

Table 8: Influnence of query number $N_q$, the $N_q$ is set as 100.

| Number | AP$_{ped}$ | AP$_{div}$ | AP$_{bou}$ | mAP |
|---|---|---|---|---|
| 50 | 64.9 | 66.1 | 67.3 | 66.2 |
| 75 | 64.6 | 65.9 | 69.1 | 66.5 |
| 100 | 64.7 | 66.1 | 69.4 | 66.8 |
| 125 | 66.8 | 63.0 | 69.6 | 66.4 |

Table 9: Influnence of Multi-Scale BEV feature.

| Shape | AP$_{ped}$ | AP$_{div}$ | AP$_{bou}$ | mAP |
|---|---|---|---|---|
| (200,100) | 64.1 | 64.9 | 68.7 | 65.9 |
| (100,50) | 64.6 | 65.3 | 69.3 | 66.3 |
| Multi-Scale | 64.7 | 66.1 | 69.4 | 66.8 |

Table 10: Long-term training results of Argoverse2.

| Methods | $AP_{ped}$ | $AP_{div}$ | $AP_{bou}$ | mAP |
|---|---|---|---|---|
| MapTRv2 (Liao et al., 2023) | 68.3 | 74.1 | 69.2 | 70.5 |
| HIMap (Zhou et al., 2024) | 72.4 | 72.4 | 73.2 | 72.7 |
| MGMapNet | 71.3 | 76.0 | 73.1 | 73.6 |

**Multi-Scale BEV feature Performance.** In Tab. 9, we conducted ablation experiments on the scale of BEV features in MGMapNet, including two different BEV sizes and multi-scale. The results showed that when the BEV size is $200 \times 100$ and $100 \times 50$, the mAP values are 65.9 and 66.3, respectively. By using multi-scale BEV features, we can capture diverse lengths of map elements, and the mAP reached 66.8.

**Ablation study on the hyperparameters Loss Function $\mathcal{L}_{aux}$.** The configuration of other losses follows the original Maptrv2 scheme. As shown in Tab. 11, when the loss weight is 0, the mAP is 65.8; however, when the weight increases to 3, the mAP increases to 66.8, which is the optimal result, proving the effectiveness of the auxiliary loss. Additionally, the experiments demonstrate that a configuration with a weight of 3 is optimal.

**Long-term training results of Argoverse2.** Most models only report experimental results for 6 epochs, while we present the long-term training results for 24 epochs 2D point coordinates here. As shown in the Tab. 10, MGMapNet still performs exceptionally well.

### A.5 BROADER IMPACT STATEMENT

The deployment of autonomous driving technology brings both opportunities and challenges. It can greatly enhance road safety and improve mobility for those unable to drive. However, it may also lead to job displacement and raises ethical and legal issues. Our research aims to develop safe and efficient autonomous driving technology, while considering its societal implications. We advocate for transparent discussions with all stakeholders to ensure responsible application of this technology.

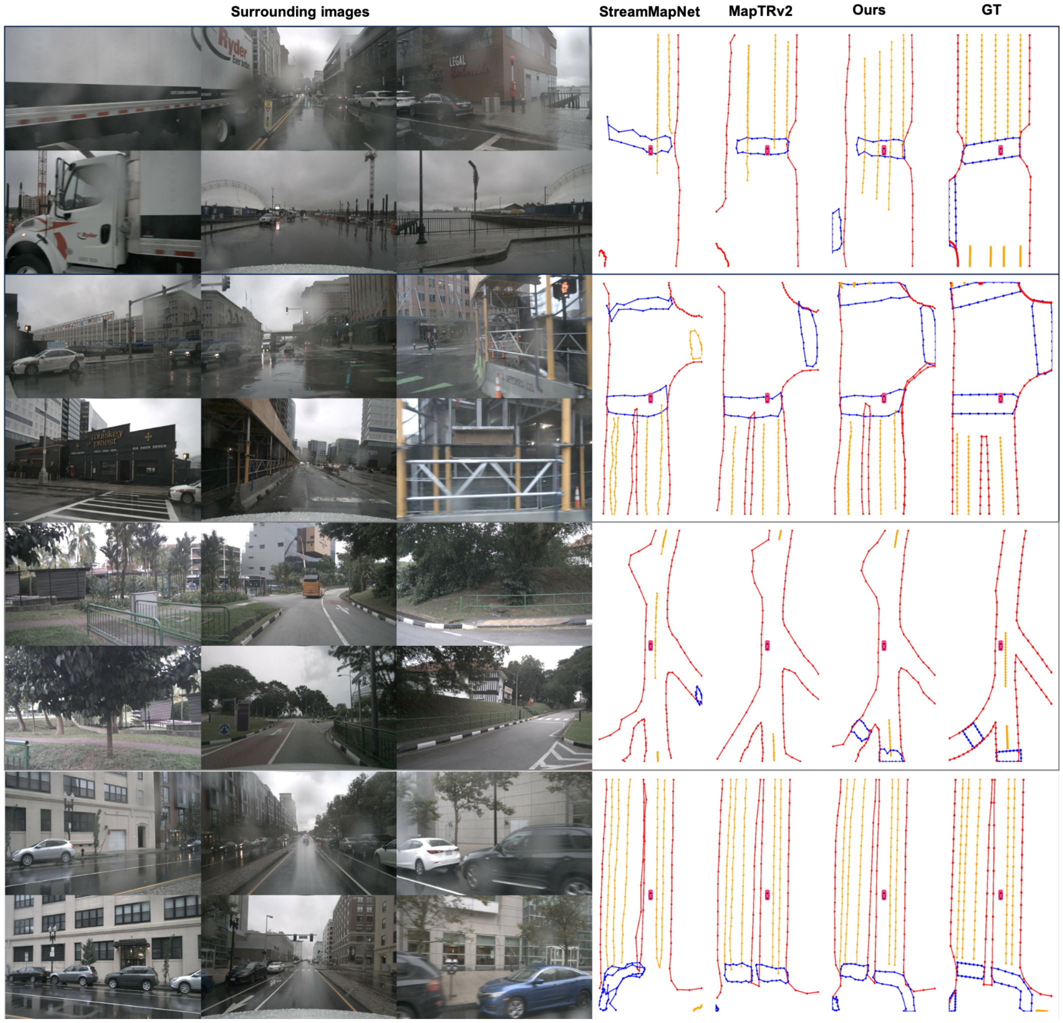


Figure 4: Visualization on nuScenes val set.

Table 11: Ablation Experiments on Loss Function $\mathcal{L}_{aux}$.

| loss weight | $AP_{ped}$ | $AP_{div}$ | $AP_{bou}$ | mAP |
|---|---|---|---|---|
| 0 | 63.7 | 65.0 | 68.7 | 65.8 |
| 1 | 64.0 | 64.8 | 69.5 | 66.1 |
| 2 | 64.0 | 66.3 | 68.9 | 66.4 |
| 3 | 64.7 | 66.1 | 69.4 | 66.8 |
| 4 | 64.5 | 65.9 | 69.1 | 66.5 |

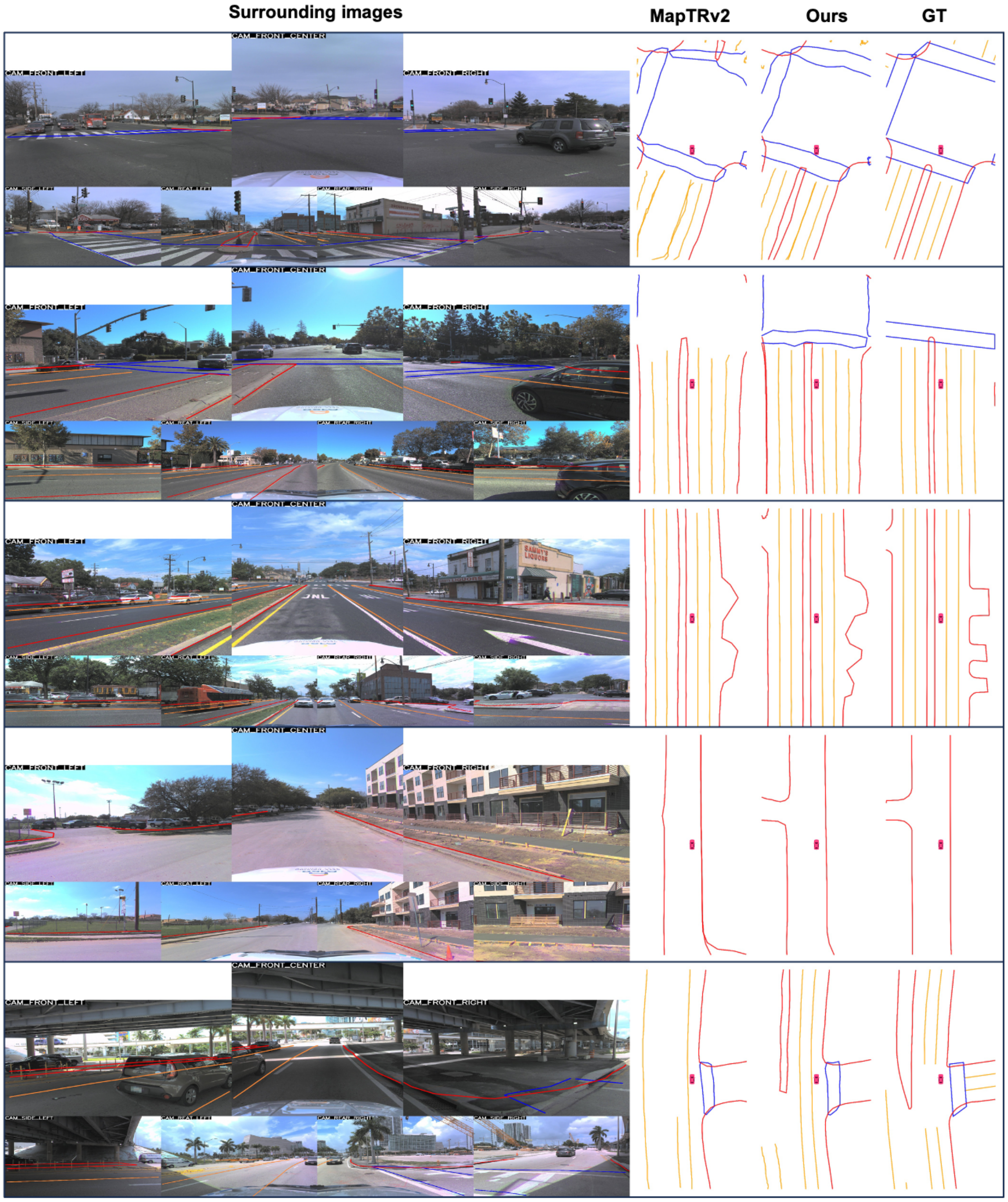


Figure 5: Visualization on Argoverse2 val set.