This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InteractionMap: Improving Online Vectorized HDMap Construction with Interaction

Kuang Wu Chuan Yang * Zhanbin Li
Langge Technology

Abstract

Vectorized high-definition (HD) maps are essential for an autonomous driving system. Recently, state-of-the-art map vectorization methods are mainly based on DETR-like framework to generate HD maps in an end-to-end manner. In this paper, we propose InteractionMap, which improves previous map vectorization methods by fully leveraging local-to-global information interaction in both time and space. Firstly, we explore enhancing DETR-like detectors by explicit position relation prior from point-level to instance-level, since map elements contain strong shape priors. Secondly, we propose a key-frame-based hierarchical temporal fusion module, which interacts temporal information from local to global. Lastly, the separate classification branch and regression branch lead to the problem of misalignment in the output distribution. We interact semantic information with geometric information by introducing a novel geometric-aware classification loss in optimization and a geometric-aware matching cost in label assignment. InteractionMap achieves state-of-the-art performance on both nuScenes and Argoverse2 benchmarks.

1. Introduction

High-Definition (HD) maps are designed for high-precision autonomous driving, which contain instance-level vectorized representation such as lane divider, road boundaries, pedestrian crossing, *etc*. The rich semantic information of road topology and traffic rules is important for the navigation of autonomous driving (AD). HD maps are traditionally constructed offline using LiDAR SLAM-based methods [43, 52] with high maintenance costs, complex pipelines, and notable localization errors. In addition, manual annotation and map updating rely heavily on human labor and time demands.

In recent years, more research endeavors have shifted towards deep-learning-based methods that construct vectorized HD maps around the ego-vehicle at runtime with onboard sensors [19, 20, 24]. With the development of bird's eye view (BEV) representation, recent approaches [4, 13, 19, 22, 38, 41, 55] treat the semantic map learning task as a semantic segmentation problem. However, rasterized maps generated by these methods lack vectorized instance-level information, which is essential to downstream tasks, such as motion forecasting and motion planning [8, 23]. To overcome the limitations of segmentation-based methods, approaches VectorMapNet [29], MapTR [24], MapTRv2 [25] predict point sets to construct end-to-end vectorized local HD maps using transformer [45] decoders inspired by DETR [2].

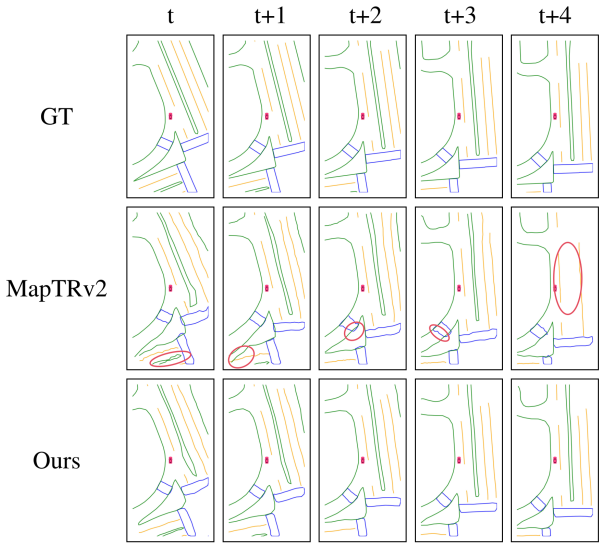


Figure 1. Visual comparison between MapTRv2 [25] and our improved results. Our method effectively eliminates error map elements, leading to better precision and stability.

As shown in Figure 1, the representation of point sets has a limited model capability of instance-level information. Moreover, predicting directly from a single frame input is challenging in complex scenarios, due to temporal inconsistency between frames. To address these issues, we propose InteractionMap, which consists of Relation Embedding Module (REM), Temporal Fusion Module (TFM),

*Corresponding at: ycscience86@gmail.com

and Geometry-aware Alignment Module (GAM). REM establishes position relation representation among feature embeddings at point level and instance level, improving the precision of map construction by explicit geometric priors. TFM facilitates long-range temporal associations by key-frame-based streaming strategy, a hierarchical temporal fusion from local to global. It enhances the stability of map elements among different frames in some complex scenarios, such as map element occlusion caused by moving vehicles. GAM solves the misalignment problem caused by inconsistency of the predictions between the classification score and the instance points precision, *e.g.* a prediction with a high classification score and relatively low localization quality (large Chamfer distance).

Our main contributions can be summarized as follows:

- We introduce an explicit position relation embedding method from point to instance, efficiently employing progressive interaction of point-wise information and instance-wise information.
- We propose a novel key-frame-based hierarchical streaming strategy, which fully leverages long-range temporal information.
- We present a geometry-aware classification loss and a geometry-aware matching cost to overcome the misalignment problem of classification and localization output.

2. Related Work

2.1. HD Map Construction

With the development of view transformation from perspective-view (PV) to bird-eye-view (BEV) methods [18, 33], HD map construction is formulated as a segmentation task [4, 13, 22, 30, 31, 36, 38, 53, 55] based on sensor observations on board. HDMapNet [19] builds a vectorized HD map using semantic segmentation, clustering, and post-processing. VectorMapNet [8] is the first end-to-end framework that utilizes transformers [45], in a two-stage coarse-to-fine manner. However, the auto-regressive model of VectorMapNet leads to a long training schedule. MapTR [24] adopts a one-stage transformer approach based on [2, 22, 58] with a permutation-equivalent point set modeling. The evolved version MapTRv2 [25] adds auxiliary headers, decoupled self-attention in the decoder and one-to-many matching strategy, leading to a large improvement. Instead of a point set representation, BeMapNet [39] utilizes an instance-level representation with a piecewise Bezier head. PivotNet [7] converts point-level representation to instance-level representation using the point-to-line mask module. MapQR [32] implicitly encodes point-level queries within instance-level queries and embeds query positions like Conditional DETR [34] and DAB-DETR [27]. MapVR [50] generates a vectorized map with differentiable rasterization that provides instance-level segmentation supervision. MGMap [28] employs mask-guided features to refine an instance representation with enhanced feature detail. In contrast to the above methods, we introduce more information interaction to improve the reliability and stability of map construction in complex scenarios.

2.2. Online Temporal Fusion

The temporal fusion strategy is effective for the 3D object detection task, such as BEVFormer [22], SOLOFusion [37], StreamPETR [46], VideoBEV [10] and Sparse4D v2 [26]. HDMapNet [19] employs max pooling to fuse temporal information directly into the BEV feature map. StreamMapNet [49] proposes query propagation and BEV feature map fusion. SQD-MapNet [47] designs a stream query denoising approach. MapTracker [3] enhances query propagation and BEV feature map fusion using a tracking strategy. The streaming strategy facilitates longer temporal association as the propagated hidden states encode all historical information. However, a temporal encoder such as the gated recurrent unit (GRU) [5] may still face the problem of forgetting due to limited capacity in complex outdoor environments. For example, occlusion of moving vehicles may lead to temporal fusion of polluted BEV feature map. The stacking strategy may integrate features from specific previous frames, offering flexibility in fusion of long-range information. In this paper, we propose a key-frame-based streaming strategy, leveraging local and global fusion capabilities.

2.3. Map Element Interaction

ADMap [15] explores point-order relationships between and within instances through a cascading approach. InsightMapper [48] performs inner-instance feature aggregation by an additional masked inner-instance self-attention module. GeMap [54] designs a masked decoupled self-attention to handle the shapes and relations of the instance independently. HoMap [1] introduces high-order modeling to capture the correlations between instances using high-order statistics. HIMap [57] proposed a hybrid framework to learn and interact with information at the point-level and the element-level. As proved in [14], modeling relations between objects is beneficial to object recognition. In state-of-the-art methods [9, 12], relationship of queries improves DETR-like models. In this paper, we explore the interaction of map elements between their geometry and appearance feature by our explicit position relation embedding.

2.4. Classification-Localization Alignment

The classification head and the localization head of object detection are implemented separately by two branches in parallel, causing inconsistency in the output distribution. This misalignment problem has been well researched in the 2D object detection field [21, 51]. However, it is overlooked by the map element detectors [7, 24, 25, 32, 39, 50]. In this

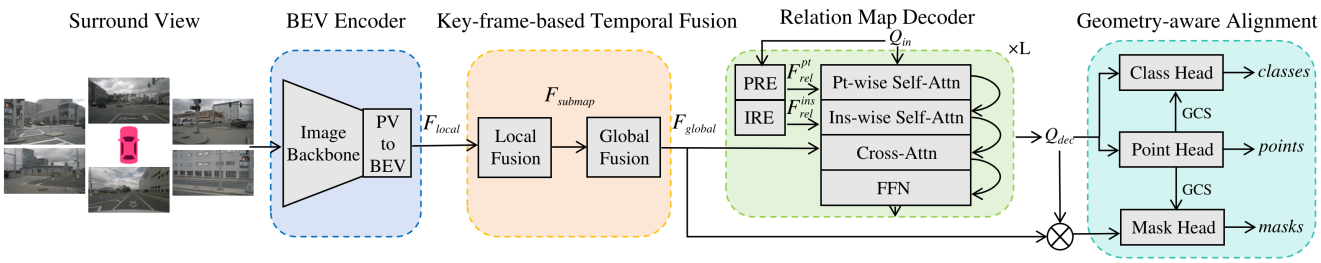


Figure 2. **Overview of InteractionMap framework.** InteractionMap mainly consists of four components: (1) BEV encoder transforms sensor input to a unified BEV representation; (2) Key-frame-based temporal fusion module leverage temporal information from local to global; (3) Relation map decoder utilizes relation embedding in point-level and instance-level; (4) Geometry-aware alignment module is designed to solve the misalignment problem of classification and position output.

paper, we create a geometry-aware classification loss, making use of focal loss [42] with a geometry-aware target in foreground candidates. Moreover, we present a geometry-aware matching cost to suppress the matching of background candidates.

3. Method

Figure 2 illustrates the framework of our proposed InteractionMap. Firstly, BEV features are extracted from multi-view images by the BEV encoder, which consists of a backbone [11] to extract multi-scale image features, and a view transformation module [22, 31] to encode image features into a BEV representation $\mathcal{F}_{local}$. Secondly, key-frame-based temporal fusion leverages long-range temporal information by local fusion and global fusion (Section 3.1). Subsequently, the relation map decoder utilizes point-wise relation embedding (PRE) and instance-wise relation embedding (IRE) for better interaction between queries (Section 3.2). Finally, by introducing a geometric-aware classification loss and a geometric-aware matching cost, the geometry-aware alignment leverages interaction between semantic information and geometry information (Section 3.3).

3.1. Key-frame-based Temporal Fusion

Compared to the single-frame map element prediction method, there are mainly two types of temporal fusion methods, namely stacking-based strategy and streaming-based strategy. The streaming strategy leverages all historical information by hidden state encoding. However, due to limited memory capacity, the streaming strategy has limited performance in long-range perception. The stacking strategy fuse features from specific previous frames, offering flexibility in integration of long-range information. The computational cost is linearly related to the number of fused frames.

We propose a key-frame-based streaming (KFS) strategy to address these problems. The well-known key-frame-based strategy is widely used in robot navigation and mapping, such as sparse mapping [16, 35] and dense mapping [6, 56]. Online reconstruction methods are unstable in long range, but they are accurate in the local regime. The input frames are divided into local frame segments. The submaps are reconstructed by integration of local frames, and then the global map is reconstructed by integration of submaps. We introduce two types of KFS strategy, namely KFS-streaming and KFS-stacking, as shown in Figure 3. Firstly, the BEV feature embedding submaps are generated by a streaming-base temporal encoder at the bottom level, which keeps the stability and temporal consistency of predictions cross latest frames. Secondly, the long-range BEV feature embedding is generated by a streaming-based or stacking-based temporal encoder at the top level. Figure 3 illustrates the KFS strategy that merges one previous submap with an interval of two.

3.1.1 Local BEV Fusion

We warp the memorized BEV feature map from the previous frame to the current frame recurrently based on the ego vehicle's relative pose. Then we employ a GRU [5] to fuse current BEV feature map $\mathcal{F}^t_{local}$ and the warped memorized BEV feature map $\tilde{\mathcal{F}}^{t-1}_{submap}$ into a single BEV feature embedding submap candidate $\mathcal{F}^t_{submap}$.

$$\tilde{\mathcal{F}}^{t-1}_{submap} = Warp(\mathcal{F}^{t-1}_{submap}, \boldsymbol{T}) \tag{1}$$

$$\mathcal{F}^t_{submap} = ResBlock(LN(GRU(\tilde{\mathcal{F}}^{t-1}_{submap}, \mathcal{F}^t_{local}))) \tag{2}$$

Where $ResBlock$ is a residual-based convolution block with the shortcut connection [11]. LN is a layer normalization operation. $\boldsymbol{T}$ denotes a standard 4×4 transformation matrix between the coordinate systems of two frames.

3.1.2 Global BEV Fusion

Our KFS strategy selects a BEV feature embedding as a submap using a specific distance stride d_{stride}, instead of a specific temporal interval customly.

KFS-streaming strategy The previous selected global BEV feature is warped to the current frame. The warped

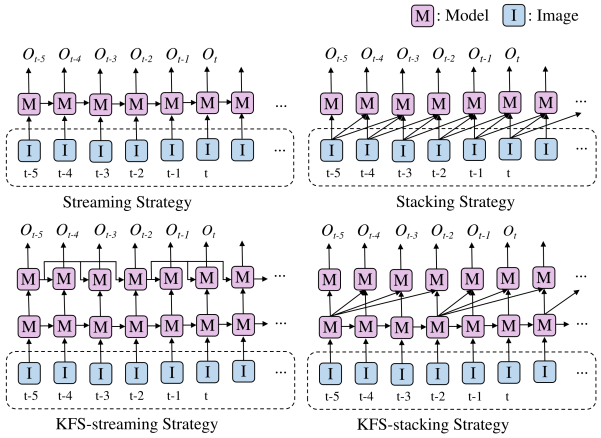


Figure 3. Temporal fusion strategy.

global BEV feature and the current BEV feature map are then fused to create a global BEV feature map.

$$\tilde{\mathcal{F}}^{pre}_{global} = Warp(\mathcal{F}^{pre}_{global}, \boldsymbol{T}) \tag{3}$$

$$\mathcal{F}^{t}_{global} = LN(GRU(\tilde{\mathcal{F}}^{pre}_{global}, \mathcal{F}^{t}_{submap})) \tag{4}$$

KFS-stacking strategy We warp N_{pre} selected submaps to the current frame. Following concatenation and residual-based convolution [11], we obtain the global BEV feature.

$$\tilde{\mathcal{F}}^{t_k}_{submap} = Warp(\mathcal{F}^{t_k}_{submap}, \boldsymbol{T}) \tag{5}$$

$$\mathcal{F}^{t}_{global} = ResBlock(Concat(\tilde{F}_{submap}, \mathcal{F}^{t}_{submap})) \tag{6}$$

Here, $\tilde{\mathcal{F}}^{t_k}_{submap}$ and t_k represent the selected BEV feature submap and its corresponding index. The collection of selected BEV feature submaps is denoted as $\tilde{F}_{submap} = \{\tilde{\mathcal{F}}^{t_k}_{submap}\}^{N_{pre}}_{k=1}$. During the training period, the previous N_{pre} submaps are randomly selected within a range of $N_{pre} \times d_{stride}$.

3.2. Relation map decoder

Previous work has shown the effectiveness of interactions for map element detectors [15, 57]. In contrast to these approaches, we directly construct explicit relation priors by incorporating relation embedding into the self-attention module of the decoder.

3.2.1 Relation Embedding

Here is the self-attention in the decoder of a DETR-like framework.

$$Q_{self_attn} = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

Here d_k is the dimension of the keys. Q, K, V are feature embeddings generated by linear projections. $Q = Linear_q(Q_{in})$. $K = Linear_k(Q_{in})$. $V = Linear_v(Q_{in})$

We incorporate relation embeddings into the self-attention model, inspired by [14].

$$Q_{self_attn} = Softmax(\mathcal{F}_{rel} + \frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

Here $\mathcal{F}_{rel}$ represents the relation embedding. In this paper, point-wise relation embedding and instance-wise relation embedding are added to the decoupled self-attention model at the point level and the instance level, respectively, employing attention refinement of information interaction.

3.2.2 Point-wise Relation Embedding

The point-wise relation encoder represents the high-dimensional point relation embedding as an explicit geometry prior. The point-wise relation embedding is generated by normalized point coordinates and edge directions.

$$Rel_{pt}(i,j) = [\log(x_i - x_j + 1), \log(y_i - y_j + 1)] \tag{9}$$

$$Rel_{dir}(i,j) = cosine_sim(e_i, e_{i+1}) - cosine_sim(e_j, e_{j+1}) \tag{10}$$

Here x_i, x_j are normalized point coordinates. e_i and e_{i+1}, e_j and e_{j+1} are adjacent edges. The cosine similarity is denoted by $cosin_sim$. The point-wise relation embedding is unbiased, since both $Rel_{pt}(i,j)$ and $Rel_{dir}(i,j)$ are equal to 0 when $i = j$.

The point-wise relation embedding $\mathcal{F}^{pt}_{rel}$ is illustrated by the following equation:

$$\mathcal{F}^{pt}_{rel} = ReLU(Linear(SPE(Concat(Rel_{pt}, Rel_{dir})))) \tag{11}$$

Firstly, the point-wise relation embedding is calculated by concatenation of Rel_{pt} and Rel_{dir}. Then the embedding is transformed to high dimension by sine positional encoding (SPE). Lastly, the embedding is transformed by a linear projection and a ReLU function.

3.2.3 Instance-wise Relation Embedding

The instance-wise relation embedding is generated by classification scores and chamfer distance of points.

$$Rel_s(i,j) = sgn(s_i - s_j) \tag{12}$$

$$Rel_{sd}(i,j) = sgn(s_i - s_j) \cdot Chamfer(I_i, I_j) \tag{13}$$

Here, sgn is the sign function. s_i, s_j are classification scores of the i^{th} and j^{th} candidates. $Chamfer(I_i, I_j)$ is the chamfer distance between points of instance i and instance j. The equation of $Rel_s(i,j)$ establishes a simple classification ranking relation among pairs of instances. Then we add the position relation of the map elements to the instance-wise relation embedding $\mathcal{F}^{ins}_{rel}$ by the chamfer distance.

$$\mathcal{F}^{ins}_{rel} = ReLU(Linear(SPE(Rel_{sd}))) \tag{14}$$

The instance-wise relation embedding is unbiased, since $Rel_{sd}(i,j)$ is equal to 0 when $i=j$.

3.3. Geometry-aware Alignment

Current work, such as MapTR [24] uses the classification branch and the regression branch to predict semantic labels and geometric localization separately. MapTR uses the focal loss [42] as the classification loss and the point-to-point L1 loss as the regression loss. Since classification and regression losses are independent, the quality of these two branches is inconsistent, leading to the well-known misalignment problem. In this paper, we introduce the geometry-aware focal loss and the geometry-aware focal cost to overcome this problem.

3.3.1 Geometry-aware Focal Loss

Inspired by IoU-aware Focal Loss [21, 51] in traditional object detection, we design the novel Geometry-aware Focal Loss (GFL). Unlike focal loss, we treat positives and negatives asymmetrically.

$$\mathcal{L}_{GFL} = \sum_{i=1}^{N_{pos}} s_{geo_i} BCE(s_{geo_i}, p_i) + \sum_{j=1}^{N_{neg}} \alpha p_j^{\gamma} BCE(p_j, 0) \tag{15}$$

where p_i and p_j are the predicted probability of the i^{th} and j^{th} candidates, $s_{geo} \in [0,1]$ is an instance Geometry-aware Classification Score (GCS). Here we define three types of GCS, including s_{p2p}, s_{dir} and s_{giou}.

$$s_{p2p} = 1 - \frac{1}{2N_p} \sum_{i=1}^{N_p} D_{Manhattan}(p_{src_i}, p_{tgt_i}) \tag{16}$$

$$s_{dir} = 0.5 + \frac{1}{2N_e} \sum_{i=1}^{N_e} cosine_sim(e_{src_i}, e_{tgt_i}) \tag{17}$$

$$s_{giou} = 0.5 + 0.5 \cdot GIoU(B_{src}, B_{tgt}) \tag{18}$$

where $s_{p2p} \in [0,1]$ is a normalized point-to-point L1 score, $s_{dir} \in [0,1]$ is an edge direction score calculated by cosine similarity and $s_{giou} \in [0,1]$ is a normalized GIoU [40] score of candidate bounding box B_{src} and assigned GT B_{tgt}. The bounding box is computed by minimum enclosing box of instance points. N_p and N_e are the numbers of points and edges. p_{src} and p_{tgt} are normalized points of candidates and assigned GT. e_{src} and e_{tgt} are edges of candidates and assigned GT.

3.3.2 Geometry-aware Focal Cost

The loss function $\mathcal{L}_{GFL}$ supervises the classification scores with geometric metrics GCS s_{geo_i}. And s_{geo_i} accurately selects these high-quality candidates, which leads to better map vectorization performance. Following the spirit of GFL, we make similar modifications to the focal cost, which is a key component of label assignment. The novel matching cost, geometry-aware focal cost (GFC) is as follows.

$$\mathcal{C}_{GFC}(i,j) = s_{geo_i} BCE(s_{geo_i}, p_i) - \alpha p_i^{\gamma} BCE(1-p_i, 1) \tag{19}$$

where $\mathcal{C}_{GFC}(i,j)$ is the geometry-aware focal cost for the i^{th} prediction and the j^{th} ground truth. The GFC is used as a modulated function to suppress candidates with inaccurate prediction localization and promote candidates with an accurate position.

3.4. Training Loss

Auxiliary BEV Supervision. We introduce a query-based instance segmentation (QIS) module using an instance mask prediction branch. Inspired by [17], we make use of the query embeddings Q_{dec} from Transformer to dot-product the BEV feature embedding map $\mathcal{F}_{bev}$ to obtain instance-wise binary masks m with the sigmoid function $\sigma()$.

$$\hat{M}_{ins} = \sigma(MLP(Q_{dec}) \cdot \mathcal{F}_{bev}) \tag{20}$$

For the instance segmentation branch, we extend the mask focal loss and the mask focal cost to mask geometry-aware focal loss (MGFL) and mask geometry-aware focal cost (MGFC) as well, with the same ideal of GFL.

$$\mathcal{L}_{seg} = \lambda_{mgf}\mathcal{L}_{mgf}(\hat{M}_{ins}, M_{ins}) + \lambda_{dice}\mathcal{L}_{dice}(\hat{M}_{ins}, M_{ins}) \tag{21}$$

Instance masks $\hat{M}_{ins}$ are supervised by instance mask annotations M_{ins}, using the mask geometry-aware focal loss $\mathcal{L}_{mgf}$ and the dice loss $\mathcal{L}_{dice}$.

Overall Loss. The total loss L used by InteractionMap is as follow:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{seg} + \mathcal{L}_{aux} \tag{22}$$

where $\mathcal{L}_{det}$ is the loss term for the detection task. $\mathcal{L}_{seg}$ is the loss term for the segmentation task. $\mathcal{L}_{aux}$ is the auxiliary loss term.

The detection loss consists of classification loss $\mathcal{L}_{cls}$, point-to-point distance loss $\mathcal{L}_{p2p}$, edge direction loss $\mathcal{L}_{dir}$.

$$\mathcal{L}_{det} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{p2p}\mathcal{L}_{p2p} + \lambda_{dir}\mathcal{L}_{dir} \tag{23}$$

The loss term $\mathcal{L}_{cls}$ is the geometry-aware focal loss for the classification of instances. The loss term $\mathcal{L}_{p2p}$ is the smooth L1 loss for the regression of vectorized map elements. And the loss term $\mathcal{L}_{dir}$ is the loss of cosine similarity between adjacent edges.

$$\mathcal{L}_{aux} = \lambda_{depth}\mathcal{L}_{depth} + \lambda_{PVSeg}\mathcal{L}_{PVSeg} \tag{24}$$

The auxiliary loss term $\mathcal{L}_{aux}$ is composed of the depth-map prediction loss $\mathcal{L}_{depth}$ and image semantic segmentation loss $\mathcal{L}_{PVSeg}$.

4. Experiment

4.1. Experimental Settings

4.1.1 Datasets

nuScenes Datasets. There are 1000 scenes of 4 locations in nuScenes datasets. Each scene contains 20s of RGB images from six cameras, point clouds from LiDAR sweeps, and a 3D vectorized map. The train set contains 700 scenes with 28130 samples and validation set contains 150 scenes with 6019 samples, respectively. Following previous works [24, 25], we mainly focus on three categories of map elements, including lane divider, pedestrian crossing and road boundary.

Argoverse2 Datasets. Argoverse2 dataset provide 1000 scenes in 6 cities, each lasting around 15 seconds. Each log includes 7 RGB images from surrounding cameras and point cloud from LiDAR sweeps. Follow previous methods [24, 25], 700 scenes are used for training, and 150 scenes are used for validation. For a fair comparison, we report results on its validation set and focus on the same three map categories as the nuScene dataset.

4.1.2 Evaluation Metrics

For a fair comparison, we follow the standard metric used in previous works [19, 24, 25, 29]. With the ego-coordination as the center, the perception ranges are $[-15m, 15m]$ for the X-axis and $[-30m, 30m]$ for the Y-axis. We adopt the mean Average Precision (mAP) to evaluate the map vectorization quality based on the chamfer distance. A candidate is considered as True-Positive (TP) only if the chamfer distance to ground truth is less than certain thresholds ($\tau \in T, T = 0.5, 1.0, 1.5$). The final AP metric is calculated by average across all thresholds and all classes.

4.1.3 Implementation Details

Following previous method [24, 25], we use ResNet50 [11] as the backbone for RGB images. All models are trained with 8 A800 GPUs. The default batch size is 32. The initial learning rate is set to 6×10^{-4} with cosine decay. The nuScenes images resolution is 1600×900. We resize the images with 0.5 ratio. For the Argoverse2 dataset, the image resolutions are 2048×1550 and 1550×2048. We pad the images to 2048×2048, and then resize the images with 0.3 ratio. We define the size of BEV feature as $H \times W$ of 200×100, and the size of each BEV grid as 0.3 meters. The default numbers of instance queries, point queries, and decoder layers are 100, 20, 6, respectively. The model is trained on 24, 110 epochs on the nuScenes and Argoverse2 datasets. We randomly divide each training sequence into 3 splits to generate more diverse data sequences. Inspired by SOLOFusion [37], we train the initial N_{init} epochs with single-frame input to stabilize multi-frame straining. $N_{init} = 0.5N_{total}$. For loss weights, $\lambda_{cls} = 2$, $\lambda_{p2p} = 4$, $\lambda_{dir} = 0.005$, $\lambda_{mgf} = 30$, $\lambda_{dice} = 3$, $\lambda_{depth} = 3$, $\lambda_{PVSeg} = 2$. For temporal fusion, $N_{pre} = 4$, $d_{stride} = 5m$.

4.2. Performance Comparison

Performance on nuScenes Dataset. As shown in Table 1, our methods show consistent improvements over baselines. By integrating our method with MapTRv2 [25], we achieve 71.6(+10.1), 71.8(+10.3) mAP with 24-epoch training schedules and 75.5(+6.8), 76.0(+7.3) mAP with 110-epoch training schedules, respectively. InteractionMap establishes a new state-of-the-art performance, exhibiting significant improvements over existing methods with comparable inference speed, as shown in Table 4.

Performance on Argoverse2 Dataset. Argoverse2 dataset contains 3D map elements, including height information. For fair comparison, the sampling frequency is set to 2Hz and 10Hz, respectively. The amount of training data at a 10Hz sampling frequency is 5 times as many as that at a 2Hz sampling frequency. As demonstrated in Table 2 and Table 3, InteractionMap surpasses all state-of-the-art methods in both 2D and 3D vectorized map perception on Argoverse2 dataset.

4.3. Ablation Study

We examine the efficacy of each component of InteractionMap through ablation studies, utilizing the nuScenes dataset. Initially, we build a simple baseline base on MapTR [24] with the LSS encoder without auxiliary depth-map supervision. The default numbers of instances, points per instance, are 50 and 20, respectively.

Contributions of Main Components. Table 5 demonstrates the impact of each component of InteractionMap. When adding relation embedding into the baseline model, there is a substantial improvement of 4.5 mAP. Furthermore, the addition of geometry alignment results in a significant 7.5 increase in mAP, emphasizing its effectiveness in boosting performance. Lastly, key-frame-based temporal fusion offers 5.0 and 5.1 additional increases in mAP, indicating that the inclusion of long-range hierarchical streaming strategy significantly enhances the performance.

Ablation on Relation Embedding. In Table 6, we explore the effect of relation embedding in point-level and instance-level. Adding the point-wise relation embedding of point coordinates and edge direction results in 2.1 and 0.9 improvements in mAP over the baseline, respectively. The instance-wise relation embedding leverages the classification confidence and points coordinates, leading to enhancements of 1.1 and 3.1 mAP, respectively. The combination of relation embedding in point-level and instance-level enhances performance by 2.8 mAP. Finally, we introduce a

Method	Backbone	Epochs	Temporal	AP_{ped}	AP_{div}	AP_{bou}	mAP
MapTR	R50	24		46.3	51.5	53.1	50.3
MapVR	R50	24		47.7	54.4	51.4	51.2
PivotNet	R50	30		53.8	55.8	59.6	57.4
BeMapNet	R50	30		57.7	62.3	59.4	59.8
MapTRv2	R50	24		59.8	62.4	62.4	61.5
StreamMapNet	R50	30	✓	61.7	66.3	62.1	63.4
MGMap	R50	24		61.8	65.0	67.5	64.8
SQD-MapNet	R50	24	✓	63.6	66.6	64.8	65.0
MapQR	R50	24		63.4	68.0	67.7	66.4
HIMap	R50	30		62.6	68.4	69.1	66.7
HRMapNet-MapTRv2	R50	24	✓	65.8	67.4	68.5	67.2
InteractionMap-R	R50	24	✓	71.3	70.8	72.8	71.6
InteractionMap-C	R50	24	✓	69.7	72.7	73.0	71.8
VectorMapNet	R50	110+ft		42.5	51.4	44.1	46.0
MapTR	R50	110		56.2	59.8	60.1	58.7
MapVR	R50	110		55.0	61.8	59.4	58.8
BeMapNet	R50	110		62.6	66.7	65.1	64.8
MapTRv2	R50	110		68.1	68.3	69.7	68.7
MapQR	R50	110		70.1	74.4	73.2	72.6
HRMapNet-MapTRv2	R50	110	✓	72.0	72.9	75.8	73.6
HIMap	R50	110		71.3	75.0	74.7	73.7
InteractionMap-R	R50	110	✓	75.2	75.0	76.1	75.5
InteractionMap-C	R50	110	✓	75.3	75.6	77.0	76.0

Table 1. Comparison with SOTA methods on the nuScenes validation set at $60m \times 30m$ perception range. "EB0", "EB4", "R50" correspond to the backbones Efficient-B0, Efficient-B4 [44], ResNet50 [11]. "ft" means the two-stage fine-tune strategy. "C" means KFS-stacking strategy and "R" means KFS-streaming strategy.

Method	Dim	AP_{ped}	AP_{div}	AP_{bou}	mAP
StreamMapNet	2d	62.0	59.5	63.0	61.5
SQD-MapNet	2d	64.9	60.2	64.9	63.3
MapTRv2	2d	60.0	68.7	64.2	64.3
HRMapNet-MapTRv2	2d	65.1	71.4	68.6	68.3
InteractionMap-C	2d	68.5	77.9	74.6	73.7
InteractionMap-R	2d	70.1	77.0	74.0	73.7

Table 2. Comparison with SOTA methods on the Argoverse2 validation set at $60m \times 30m$ perception range with a 30-epoch training schedule, at a 2Hz sampling frequency.

alternating training strategy to randomly turn off one of these two components, leading the overall design achieves 54.7(+4.5) mAP.

Ablation on Key-frame-based Temporal Fusion. To investigate the effectiveness of the key-frame-based streaming strategy, we conduct ablation experiments to compare it with the baseline (without temporal fusion) and the streaming strategy, in Table 7. Employing the streaming strategy results in 3.8 mAP over the baseline. Furthermore, when the KFS-stacking strategy is introduced, the mAP increased by 7.4, reaching 57.6. Additionally, the inclusion of KFS-streaming strategy leads to an improvement of 7.2 mAP.

Ablation on Geometry-aware Alignment. As shown in

Method	Dim	AP_{ped}	AP_{div}	AP_{bou}	mAP
MapTRv2	2d	63.6	71.5	67.4	67.5
MapQR	2d	64.3	72.3	68.1	68.2
HIMap	2d	69.0	69.5	70.3	69.6
InteractionMap-C	2d	69.8	78.6	74.8	74.4
InteractionMap-R	2d	70.8	78.1	74.9	74.6
MapTRv2	3d	60.7	68.9	64.5	64.7
MapQR	3d	60.1	71.2	66.2	65.9
HIMap	3d	66.7	68.3	70.3	68.4
InteractionMap-C	3d	66.6	75.6	72.7	71.6
InteractionMap-R	3d	67.7	75.5	73.1	72.1

Table 3. Comparison with SOTA methods on the Argoverse2 validation set at $60m \times 30m$ perception range with a 6-epoch training schedule, at a 10Hz sampling frequency.

Method	FPS	Params(MB)	mAP
MapTRv2	14.9	40.6	61.5
InteractionMap-C	11.0	116.7	71.8
InteractionMap-R	12.1	49.9	71.6

Table 4. Comparison with SOTA methods on the nuScenes validation set. FPS is measured on a single A800 GPU.

REM	GAM	TFM	AP_{ped}	AP_{div}	AP_{bou}	mAP
			46.7	50.5	53.5	50.2
✓			50.1	56.6	57.3	54.7
	✓		57.1	63.8	63.7	61.5
		C	50.9	60.0	61.9	57.6
		R	50.6	61.3	60.3	57.4
✓	✓		58.6	65.1	62.9	62.2
✓	✓	C	64.2	68.6	68.8	67.2
✓	✓	R	64.1	68.1	69.8	67.3

Table 5. Ablation study of each component, including Relation Embedding Module (REM), Temporal Fusion Module (TFM), Geometry-aware Alignment Module (GAM). "C" means KFS-stacking strategy and "R" means KFS-streaming strategy.

Pt_{xy}	Pt_{dir}	Ins_s	Ins_{sd}	PP	AP_{ped}	AP_{div}	AP_{bou}	mAP
					46.7	50.5	53.5	50.2
✓					47.8	52.7	56.3	52.3
	✓				47.0	52.4	53.9	51.1
✓	✓				47.8	55.6	56.4	53.3
		✓			46.9	52.1	54.8	51.3
			✓		47.9	55.7	56.4	53.3
✓	✓		✓		47.6	55.9	55.5	53.0
✓	✓		✓	✓	50.1	56.6	57.3	54.7

Table 6. Ablation study of relation embedding by employing point-wise coordinate embedding Pt_{xy}, point-wise edge direction embedding Pt_{dir}, instance-wise classification score embedding Ins_s, instance-wise classifier score and chamfer distance embedding Ins_{sd} and alternating training strategy of relation embedding, namely Ping-Pong (PP).

Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
Baseline	46.7	50.5	53.5	50.2
Streaming	48.6	54.7	58.8	54.0
KFS-stacking	50.9	60.0	61.9	57.6
KFS-streaming	50.6	61.3	60.3	57.4

Table 7. Ablation study of streaming strategy, KFS-streaming strategy and KFS-stacking strategy.

Table 8, geometry-aware alignment comprised five main components: query-based instance segmentation (QIS), geometry-aware focal loss (GFL), geometry-aware focal cost (GFC), mask geometry-aware focal loss (MGFL) and mask geometry-aware focal cost (MGFC). The QIS leverages geometry priors by instance segmentation mask supervision. Compared to the baseline, the QIS leads to an improvement of 6.4 mAP. The GFL encourages predictions with high localization metrics to obtain a better classification score, with an improvement of 3.1 mAP compared to mask focal loss. The GFC suppresses the matching cost of candidates with an inaccurate prediction position, which brings an improvement of 0.9 mAP. For the segmentation

QIS	GFL	GFC	MGFL	MGFC	AP_{ped}	AP_{div}	AP_{bou}	mAP
					46.7	50.5	53.5	50.2
✓					51.7	58.7	59.4	56.6
✓	✓				55.1	62.0	62.0	59.7
✓	✓	✓			55.8	63.8	62.2	60.6
✓	✓	✓	✓		57.0	63.6	62.3	61.0
✓	✓	✓	✓	✓	57.1	63.8	63.7	61.5

Table 8. Ablation study of geometry-aware alignment by investigating the performance of query-based instance segmentation (QIS), geometry-aware focal loss (GFL), geometry-aware focal cost (GFC), mask geometry-aware focal loss (MGFL) and mask geometry-aware focal cost (MGFC).

S_{giou}	S_{p2p}	S_{dir}	AP_{ped}	AP_{div}	AP_{bou}	mAP
			51.7	58.7	59.4	56.6
✓			52.6	56.2	58.0	55.6
	✓		54.8	60.2	63.1	59.4
		✓	58.3	63.1	62.5	61.3
	✓	✓	57.1	63.8	63.7	61.5

Table 9. Ablation study of geometry-aware classification score.

branch, the result shows that the position-supervised segmentation loss MGFL and the position-modulated segmentation cost MGFC improve the final results, with 0.4 mAP and 0.5 mAP gains, individually.

We further investigate the geometry-aware classification score S_{geo}. As depicted in Table 9, using S_{giou}, S_{p2p} and S_{dir} leads to individual improvement of -1.0, 2.8 and 4.7 mAP, respectively. Furthermore, the combination of S_{p2p} and S_{dir} boosts the performance, improving by 4.9 mAP.

5. Conclusion

In this paper, we introduce InteractionMap, a novel approach to end-to-end online vectorized HD map utilizing temporal and spatial information interaction from local to global. By leveraging relation embedding at both point level and instance level, the network learns the geometric priors of map elements better. Subsequently, key-frame-based streaming strategy enhances the network by utilizing temporal feature embedding from near to distant. Finally, the geometric-aware focal loss and the geometric-aware focal cost build a strong correlation between classification score and instance-point location precision. Across various experimental settings, our proposed InteractionMap yields significant performance improvements over the baseline on various datasets.

References

[1] Yingfeng Cai, Wei Dong, Ze Liu, Hai Wang, and Long Chen. Homap: End-to-end vectorized hd map construction with

high-order modeling. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[2] 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*, pages 213–229. Springer, 2020. 1, 2

[3] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. In *European Conference on Computer Vision*, pages 90–107. Springer, 2025. 2

[4] Shaoyu Chen, Tianheng Cheng, Xinggang Wang, Wenming Meng, Qian Zhang, and Wenyu Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*, 2022. 1, 2

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2, 3

[6] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4): 1, 2017. 3

[7] 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*, pages 3672–3682, 2023. 2

[8] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11525–11533, 2020. 1, 2

[9] Yulu Gao, Yifan Sun, Xudong Ding, Chuyang Zhao, and Si Liu. Ease-detr: Easing the competition among object queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17282–17291, 2024. 2

[10] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics and Automation Letters*, 2024. 2

[11] 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*, pages 770–778, 2016. 3, 4, 6, 7

[12] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, Badong Chen, and Xuguang Lan. Relation detr: Exploring explicit position relation prior for object detection. *arXiv preprint arXiv:2407.11699*, 2024. 2

[13] 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*, pages 15273–15282, 2021. 1, 2

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 2, 4

[15] Haotian Hu, Fanyi Wang, Yaonong Wang, Laifeng Hu, Jingwei Xu, and Zhiwang Zhang. Admap: Anti-disturbance framework for reconstructing online vectorized hd map. *arXiv preprint arXiv:2401.13172*, 2024. 2, 4

[16] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007. 3

[17] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. 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*, pages 3041–3050, 2023. 5

[18] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[19] 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)*, pages 4628–4634. IEEE, 2022. 1, 2, 6

[20] 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. 1

[21] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 2, 5

[22] 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*, pages 1–18. Springer, 2022. 1, 2, 3

[23] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020. 1

[24] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022. 1, 2, 5, 6

[25] 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. 1, 2, 6

[26] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 2

[27] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2

[28] 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*, pages 14812–14821, 2024. 2

[29] 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*, pages 22352–22369. PMLR, 2023. 1, 6

[30] Zhi Liu, Shaoyu Chen, Xiaojie Guo, Xinggang Wang, Tianheng Cheng, Hongmei Zhu, Qian Zhang, Wenyu Liu, and Yi Zhang. Vision-based uneven bev representation learning with polar rasterization and surface estimation. In *Conference on Robot Learning*, pages 437–446. PMLR, 2023. 2

[31] 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)*, pages 2774–2781. IEEE, 2023. 2, 3

[32] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *European Conference on Computer Vision*, pages 461–477. Springer, 2025. 2

[33] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[34] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. 2

[35] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3

[36] Cong Pan, Yonghao He, Junran Peng, Qian Zhang, Wei Sui, and Zhaoxiang Zhang. Baeformer: Bi-directional and early interaction transformers for bird's eye view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9599, 2023. 2

[37] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 2, 6

[38] 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*, pages 194–210. Springer, 2020. 1, 2

[39] 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*, pages 13218–13228, 2023. 2

[40] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[41] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147, 2020. 1

[42] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 3, 5

[43] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 1

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[45] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1, 2

[46] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 2

[47] Shuo Wang, Fan Jia, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao. Stream query denoising for vectorized hd map construction. *arXiv preprint arXiv:2401.09112*, 2024. 2

[48] Zhenhua Xu, Kenneth KY Wong, and Hengshuang Zhao. Insightmapper: A closer look at inner-instance information for vectorized high-definition mapping. *arXiv preprint arXiv:2308.08543*, 2023. 2

[49] 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*, pages 7356–7365, 2024. 2

[50] Gongjie Zhang, Jiahao Lin, Shuang Wu, Zhipeng Luo, Yang Xue, Shijian Lu, Zuoguan Wang, et al. Online map vec-

torization for autonomous driving: A rasterization perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[51] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523, 2021. 2, 5

[52] Ji Zhang, Sanjiv Singh, et al. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, pages 1–9. Berkeley, CA, 2014. 1

[53] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2

[54] Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue. Online vectorized hd map construction using geometry. *arXiv preprint arXiv:2312.03341*, 2023. 2

[55] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022. 1, 2

[56] Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. Elastic fragments for dense scene reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 473–480, 2013. 3

[57] 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*, pages 15396–15406, 2024. 2, 4

[58] 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. 2