








# P-MapNet: Far-Seeing Map Generator Enhanced by Both SMap and HMap Priors

Zhou Jiang , Zhenxin Zhu , Pengfei Li , Huan-ang Gao , Tianyuan Yuan , Yongliang Shi ,  
Hang Zhao, and Hao Zhao 

**Abstract**—Autonomous vehicles are gradually entering city roads today, with the help of high-definition maps (HDMaps). However, the reliance on HDMaps prevents autonomous vehicles from stepping into regions without this expensive digital infrastructure. This fact drives many researchers to study online HMap generation algorithms, but the performance of these algorithms at far regions is still unsatisfying. We present P-MapNet, in which the letter P highlights the fact that we focus on incorporating map priors to improve model performance. Specifically, we exploit priors in both standard-definition map (SMap) and HMap. On one hand, we extract weakly aligned SMap from OpenStreetMap and encode it as an additional conditioning branch. Despite the misalignment challenge, our attention-based architecture adaptively attends to relevant SMap skeletons and significantly improves performance. On the other hand, we exploit a masked autoencoder to capture the prior distribution of HMap, which can serve as a refinement module to mitigate occlusions and artifacts. We benchmark on the nuScenes and Argoverse2 datasets.

**Index Terms**—Computer vision for transportation, semantic scene understanding, intelligent transportation systems.

## I. INTRODUCTION

WHILE we still don't know the ultimate answer to fully autonomous vehicles that can run smoothly in each and every corner of the earth, the community does have seen some impressive milestones, e.g., robotaxis are under steady operation in some big cities now. Yet current autonomous driving stacks heavily depend on an expensive digital infrastructure: HDMaps. With the availability of HDMaps, local maneuvers are reduced to lane following and lane changing coupled with dynamic obstacle avoidance, significantly narrowing down the space of decision making. But the generation of HDMaps, which is shown in the left-top panel of Fig. 1, is very cumbersome and expensive. And

Manuscript received 13 March 2024; accepted 22 July 2024. Date of publication 21 August 2024; date of current version 29 August 2024. This article was recommended for publication by Associate Editor Y. Sun and Editor H. Moon upon evaluation of the reviewers' comments. (Zhou Jiang and Zhenxin Zhu contributed equally to this work.) (Corresponding author: Hao Zhao.)

Zhou Jiang is with the School of Mechatronic Engineering, Beijing Institute of Technology, Beijing 100081, China, and also with the AIR, Tsinghua University, Beijing 100190, China.

Zhenxin Zhu is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China, and also with the AIR, Tsinghua University, Beijing 100190, China.

Pengfei Li, Huan-ang Gao, Yongliang Shi, and Hao Zhao are with the AIR, Tsinghua University, Beijing 100190, China (e-mail: zhaohao@air.tsinghua.edu.cn).

Tianyuan Yuan and Hang Zhao are with the IIIS and MARS Lab, Tsinghua University, Beijing 100084, China.

Codes and models are publicly available at <https://jike5.github.io/P-MapNet/>. Digital Object Identifier 10.1109/LRA.2024.3447450

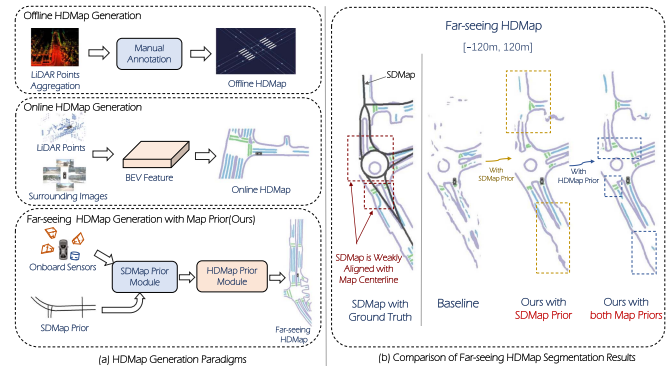


Fig. 1. Left: Since offline HMap generation is cumbersome and expensive, people are pursuing online HMap generation algorithms and our P-MapNet is an online HMap generator enhanced by both SMap and HMap priors. Right: Despite the misalignment between SMaps and HDMaps, our P-MapNet can significantly improve map generation performance, especially on the far side.

what's worse, HDMaps cannot be generated for good and all, because they must be updated every three months on average. It is widely recognized that reducing reliance on HDMaps is critical.

Thus, several recent methods [1], [2] generate HDMaps using multi-modal online sensory inputs like LiDAR point clouds and panoramic multi-view RGB images, and a conceptual illustration of this paradigm is given in the left-middle panel of Fig. 1. Despite promising results achieved by these methods, long range distance online HMap generators still report limited quantitative metrics and this study focuses on promoting their performance using priors. Specifically, two sources of priors are exploited: SMap and HMap, as demonstrated in the left-bottom panel of Fig. 1.

**SMap Prior:** Before the industry turns to build the digital infrastructure of HDMaps on a large scale, SMaps have been used for years and have largely promoted the convenience of our daily lives. Commercial SMap applications provided by Google or Baidu help people navigate big cities with complex road networks, telling us to make turns at crossings or merge into main roads. SMaps are not readily useful for autonomous cars because they only provide center-line skeletons (noted as SMap Prior in the left-bottom panel of Fig. 1). So we aim to exploit SMap priors to build better online HMap generation algorithms, which can be intuitively interpreted as *drawing* HMaps around the skeleton of SMaps. However, this intuitive idea faces a primary challenge: misalignment. Per implementation, we extract SMaps from OpenStreetMap using GPS signals but unfortunately they are, at best, weakly aligned with the ground truth HMap in a certain scenario. An

TABLE I  
QUANTITATIVE RESULTS OF IOU SCORES AND AP SCORES

Range	Method	S	S+H	M	Epoch	Div.	Ped.	Bound.	mIoU	Div.	Ped.	Bound.	mAP	FPS
60 × 30 m	HMapNet			C	30	40.5	19.7	40.5	33.57	27.68	10.26	45.19	27.71	35.4
	P-MapNet	✓		C	30	44.1	22.6	43.8	36.83 (+3.26)	32.11	11.33	48.67	30.70 (+2.99)	30.2
	P-MapNet	✓	✓	C	10	44.3	23.3	43.8	37.13 (+3.56)	26.08	17.66	48.43	30.72 (+3.01)	12.2
	HMapNet			C+L	30	45.9	30.5	56.8	44.40	29.46	13.89	54.07	32.47	21.4
	P-MapNet	✓		C+L	30	53.3	39.4	63.1	51.93 (+7.53)	36.56	20.06	60.31	38.98 (+6.51)	19.2
	P-MapNet	✓	✓	C+L	10	<b>54.2</b>	<b>41.3</b>	<b>63.7</b>	<b>53.07 (+8.67)</b>	<b>37.81</b>	<b>24.96</b>	<b>60.90</b>	<b>41.22 (+8.75)</b>	9.6
120 × 60 m	HMapNet			C	30	39.2	23.0	39.1	33.77	14.40	8.98	34.99	19.46	34.2
	P-MapNet	✓		C	30	44.8	30.6	45.6	40.33 (+6.56)	19.39	14.59	38.69	24.22 (+4.76)	28.7
	P-MapNet	✓	✓	C	10	45.5	30.9	46.2	40.87 (+7.10)	19.50	24.72	42.48	28.90 (+9.44)	12.1
	HMapNet			C+L	30	53.6	37.8	57.1	49.50	21.11	18.90	47.31	29.11	21.2
	P-MapNet	✓		C+L	30	63.6	50.2	66.8	60.20 (+10.70)	28.30	25.67	52.51	35.49 (+6.38)	18.7
	P-MapNet	✓	✓	C+L	10	<b>65.3</b>	<b>52.0</b>	<b>68.0</b>	<b>61.77 (+12.27)</b>	<b>30.63</b>	<b>28.42</b>	<b>53.27</b>	<b>37.44 (+8.33)</b>	9.6
240 × 60 m	HMapNet			C	30	31.9	17.0	31.4	26.77	7.37	5.09	21.59	11.35	22.3
	P-MapNet	✓		C	30	46.3	35.7	44.6	42.20 (+15.43)	10.86	12.74	25.52	16.38 (+5.03)	19.2
	P-MapNet	✓	✓	C	10	49.0	40.9	46.6	45.50 (+18.73)	14.51	<b>25.63</b>	28.11	22.75 (+11.40)	9.1
	HMapNet			C+L	30	40.0	26.8	42.6	36.47	11.29	11.40	29.05	17.25	13.1
	P-MapNet	✓		C+L	30	52.0	41.0	53.6	48.87 (+12.40)	17.87	20.00	<b>35.89</b>	24.59 (+7.34)	10.9
	P-MapNet	✓	✓	C+L	10	<b>53.0</b>	<b>42.6</b>	<b>54.2</b>	<b>49.93 (+13.46)</b>	<b>21.47</b>	24.14	34.23	<b>26.61 (+9.36)</b>	6.6

Performance comparison of HMapNet [1] baseline and ours on the nuScenes val set [32]. “S” indicates that our method utilizes only the SDMap priors, while “S+H” indicates the utilization of the both priors. “M” represents the modality of our method and “Epoch” represents the number of refinement epochs.

illustration is given in the right panel of Fig. 1, noted as SDMap with Ground Truth. To this end, we leverage an attention based network architecture that adaptively attends to relevant SDMap features and successfully improves the performance by large margins in various settings (see Table I).

*HMap Prior:* Although useful, SDMap priors cannot fully capture the distribution of HMap output space. As noted by Ours with SDMap Prior in the right panel of Fig. 1, HMap generation results are broken and unnecessarily curved. This is credited to the fact that our architecture is, like prior methods, designed in a BEV dense prediction manner and the structured output space of BEV HMap cannot be guaranteed. Consequently, the HMap prior is introduced as a solution. The intuition is that if the algorithm explicitly models the structured output space of HMaps, it can naturally correct the unnatural artifacts (such as the broken and unnecessarily curved results mentioned earlier). On the implementation side, we train a masked autoencoder (MAE) on a large set of HMaps to capture the HMap prior and use it as a refinement module. As shown in the right panel of Fig. 1, our MAE, utilizing both Map Priors, successfully addresses the aforementioned issues.

*P-MapNet as a far-seeing solution:* A closer look at the positive margins brought by incorporating priors reveals that P-MapNet is a far-seeing solution. As shown in the right panel of Fig. 1, after incorporating the SDMap prior, missing map elements far from the ego vehicle (denoted by the car icon) are successfully extracted. This is understandable as the road center-line skeletons on the far side are already known in the SDMap input. Meanwhile, the HMap prior brings improvements in two kinds of regions: crossings with highly structured repetitive patterns and lanes on the far side. This is credited to the fact that our masked autoencoder can incorporate the priors about how typical HMaps look like, e.g., lanes should be connected and largely straight, and crossings are drawn in a repetitive manner. As Table I demonstrates, positive margins steadily grow along with the sensing range. We believe P-MapNet, as a far-seeing solution, is potentially helpful in deriving more intelligent decisions that are informed of maps on the far side.

In summary, our contributions are three-fold: (1) We incorporate SDMap priors into online map generators by attending to weakly aligned SDMap features and achieve significant

performance improvements; (2) We also incorporate HMap priors using a masked autoencoder as a refinement module, correcting artifacts that deviate the structured output space of HMaps; (3) We achieve state-of-the-art results of *far-seeing* HMap generation on public benchmarks and present in-depth ablative analyses revealing the mechanism.

## II. RELATED WORK

### A. Online HD Map Generation

Online map generators are important for autonomous driving [3], [4], [5], [6], [7], [8], [9], [10], [11], which is similar in spirit to room layout estimation [12], [13], [14], [15] for indoor scenes. Traditionally, HD maps are manually annotated offline, combining point cloud maps via SLAM algorithms [16], [17] for high accuracy but at a high cost and without real-time updates. In contrast, recent efforts have focused on utilizing onboard sensors for the efficient and cost-effective generation of online HD maps [1], [2], [18], [19], [20], [21]. HMapNet [1] employs pixel-wise labeling and heuristic post-processing, using Average Precision (AP) and Intersection over Union (IoU) as metrics. More recent approaches [2], [22], [23], [24], [25], have adopted end-to-end vectorized HD map generation techniques, leveraging Transformer architectures [26]. However, these methods rely solely on onboard sensors and have limitations in handling challenging environmental conditions like occlusions or adverse weather.

### B. Long-Range Map Perception

To enhance the practicality of HD maps for downstream tasks, some studies aim to extend their coverage to longer perception ranges. SuperFusion [20] combines LiDAR point clouds and camera images for depth-aware BEV transformation, yielding forward-view HD map predictions up to 90 m. NeuralMapPrior [27] maintains and updates a global neural map prior, enhancing online observations to generate higher-quality, extended-range HD map predictions. [28] proposes using satellite maps to aid online map generation. Features from onboard sensors and satellite images are aggregated through a hierarchical fusion module to obtain the final BEV features. MV-Map [29]

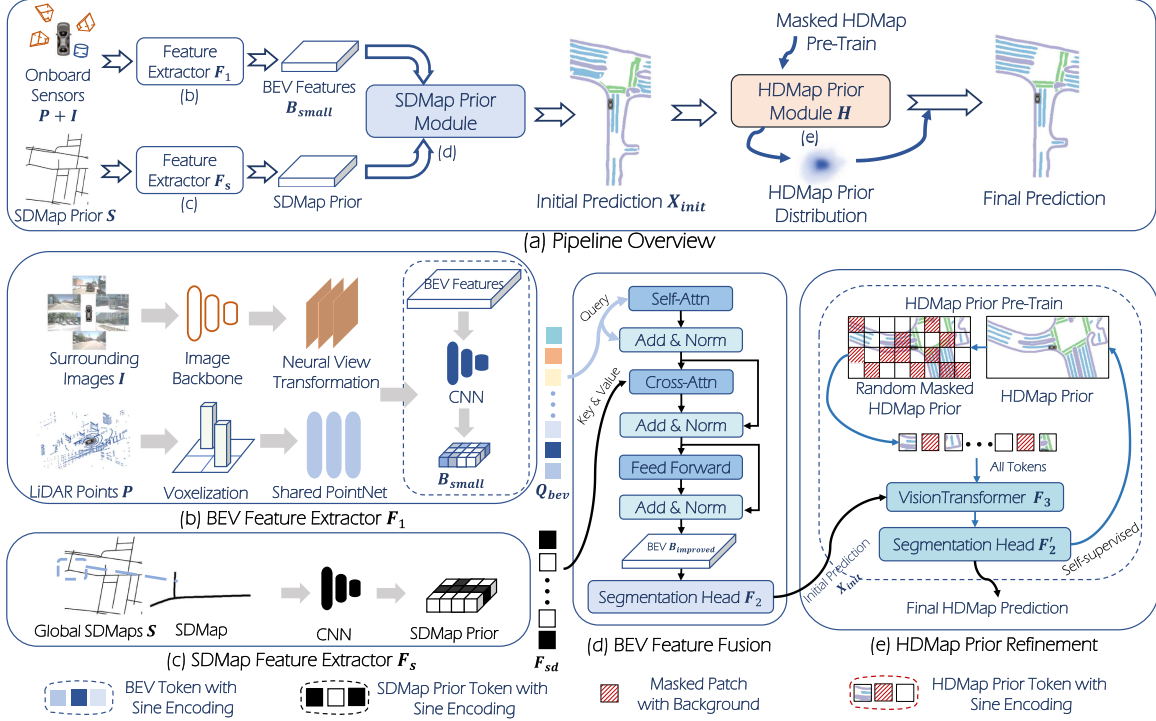


Fig. 2. **P-MapNet overview.** P-MapNet is designed to accept either surrounding cameras or multi-modal inputs. It processes these inputs to extract sensor features and SDMap priors features, both represented in the Bird’s Eye View (BEV) space. These features are then fused using an attention mechanism and subsequently refined by the HDMAP prior module to produce results that closely align with real-world map data.

specializes in offline, long-range HD map generation. It aggregates all relevant frames during traversal and optimizes neural radiance fields for improved BEV feature generation.

### III. FORMULATION

Given the LiDAR point cloud  $\mathcal{P}$  and panoramic images  $\{\mathcal{I}_i \mid i = 1, 2, \dots, N\}$ , where  $N$  is typically six for a panoramic rig, a common online HDMAP generation task (e.g., HDMAPNet [11]) can be formulated as:

$$\mathcal{M} = \mathcal{F}_2(\mathcal{F}_1(\mathcal{P}, \mathcal{I})), \quad (1)$$

where  $\mathcal{F}_1$  represents the feature extractor that takes multi-modal inputs and generates BEV features, while  $\mathcal{F}_2$  is a segmentation head that predicts a semantic category label for each grid in the BEV. And  $\mathcal{M}$  is the HDMAP prediction.

However, this common formulation fails to leverage the rich priors in SDMap and HDMAP. So we formulate a new task to incorporate these priors to produce a more accurate and *far-seeing* HDMAP, thereby effectively addressing issues pertaining to occlusion as well as super long range sensing:

$$\mathcal{M}' = \mathcal{H}(\mathcal{F}_2(\mathcal{F}_1(\mathcal{P}, \mathcal{I}) \otimes \mathcal{F}_s(\mathcal{S}))), \quad (2)$$

Here  $\mathcal{S}$  is the SDMap prior that comes in the form of road center-line skeletons and  $\mathcal{F}_s$  represents the convolutional encoder of the SDMap. The symbol  $\otimes$  indicates the cross-attention operation and  $\mathcal{H}$  represents the refinement module which is a pre-trained model capturing the distribution characteristics of HDMAP. Similarly,  $\mathcal{M}'$  is the *far-seeing* HDMAP prediction over 100 meters on the front/back side.

**Output Format:** There are two typical output formats for on-line HDMAP generation: rasterized and vectorized. In this study, we focus on the rasterized representation (e.g. HDMAPNet [11]), as it is more suitable for designing our two prior modules (than vectorized counterpart). Specifically, how to effectively encode vectorized representation for input/output is not as natural as rasterized representation.

**S-only setting:** Shown in Fig. 2(a), we incorporate SDMap prior by encoding center-line skeletons as an additional input branch. In this S-only setting, the formulation is:

$$\mathcal{M}_S = \mathcal{F}_2(\mathcal{F}_1(\mathcal{P}, \mathcal{I}) \otimes \mathcal{F}_s(\mathcal{S})), \quad (3)$$

where the procedure of encoding  $\mathcal{S}$  is illustrated in Fig. 2(c).

**S+H setting:** While the SDMap prior  $\mathcal{S}$  is naturally incorporated as an additional input, leveraging the HDMAP prior is challenging. Our innovative proposal is to incorporate HDMAP prior using a masked auto-encoder (MAE) as a refinement module. The core idea is to use MAE for the reconstruction of HDMaps so that this MAE intrinsically captures the distribution of HDMAP prior. However, this is not trivial as the vanilla MAE cannot achieve this goal.

**Vanilla MAE:** A vanilla MAE [30] would treat HDMaps as images and trains under an MSE loss for image reconstruction. The problem is that this MAE would predict images thus cannot be used as a refinement module, as our HDMAP generator actually needs a segmentation head at last.

**Our MAE variant:** Our MAE variant takes rasterized HDMAP (which is in nature images) as input but predicts the semantic label of each grid using a segmentation head. This is still an auto-encoding process since the module reconstructs the HDMAP of interest. However, this MAE’s input and output come in different

formats: images and segmentation masks. This readily allows refinement when attached after the  $\mathcal{M}_S$  output mentioned above.

*Formal notation:* Our HDMap refinement module has two training steps. The first step is to pre-train the HDMap Prior Module on a large set of HDMaps, as shown in Fig. 2(e).

$$\begin{aligned} \mathcal{M}_{\text{self}} &= \mathcal{H}(\mathcal{M}_{\text{masked}}) \\ &\triangleq \mathcal{F}'_2(\mathcal{F}_3(\mathcal{M}_{\text{masked}})), \end{aligned} \quad (4)$$

Here the HDMap Prior Module  $\mathcal{H}(\cdot)$  is specifically defined as  $\mathcal{F}'_2(\mathcal{F}_3(\cdot))$ , of which  $\mathcal{F}_3$  denotes a ViT model used in typical MAEs. But  $\mathcal{F}'_2$  denotes another segmentation head, as mentioned above. This  $\mathcal{F}'_2$  makes our MAE a variant ready for refinement.  $\mathcal{M}_{\text{masked}}$  is the randomly masked HDMap from the training dataset and  $\mathcal{M}_{\text{self}}$  is the unmasked version.

*Fine-tuning:* The second step is end-to-end fine-tuning:

$$\begin{aligned} \mathcal{M}' &= \mathcal{H}(\mathcal{M}_S) \\ &\triangleq \mathcal{F}'_2(\mathcal{F}_3()), \end{aligned} \quad (5)$$

where  $\mathcal{M}_S$  is the initial prediction from SDMap Prior Module, as depicted in Fig. 2(a) and Equation.3.

As such, the formal  $\mathcal{S}+\mathcal{H}$  setting (by integrating Equation.3 and Equation.5) is shown as:

$$\mathcal{M}' = \mathcal{F}'_2(\mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\mathcal{P}, \mathcal{I}) \otimes \mathcal{F}_s(\mathcal{S})))) \quad (6)$$

## IV. METHOD

### A. SDMap Prior Module

Now we elaborate on the implementation of our SDMap prior module, and first we recap the motivation again: Given the intrinsic challenges associated with onboard sensing, such as distant road invisibility and adverse weather conditions, incorporating SDMap prior becomes a promising technique, as SDMap provides a stable and consistent outlining of the environment (agnostic of those challenges).

*SDMap Generation:* We first introduce our approach to generate SDMap priors by leveraging OpenStreetMap (OSM) [31] data. We specifically employ the nuScenes [32] and Argoverse2 [33] datasets for our research, as these datasets hold a prominent position within the autonomous driving domain. These datasets are richly equipped with sensors but do not include the corresponding SDMap information for the captured regions. To address this limitation, we leverage OpenStreetMap to obtain the relevant SDMap data for these regions. Specifically, we first obtain the localized SDMap data of the corresponding area from the OSM<sup>1</sup> based on the onboard GPS information. Then these SDMap data are transformed to the ego vehicle's coordinate system. Although we obtain the SDMap priors, the problem of misalignment due to the low accuracy of the OSM and the bias of the GPS will pose a challenge to the fusion of SDMap priors.

*Incorporating SDMap Prior:* After extraction and rasterization, the rasterized SDMap prior inevitably faces spatial misalignment, where the SDMap prior doesn't align precisely with the current operational location, often resulting from inaccurate GPS signals or rapid vehicle movement. This misalignment renders the straightforward method of directly concatenating BEV features with SDMap features in the feature dimension

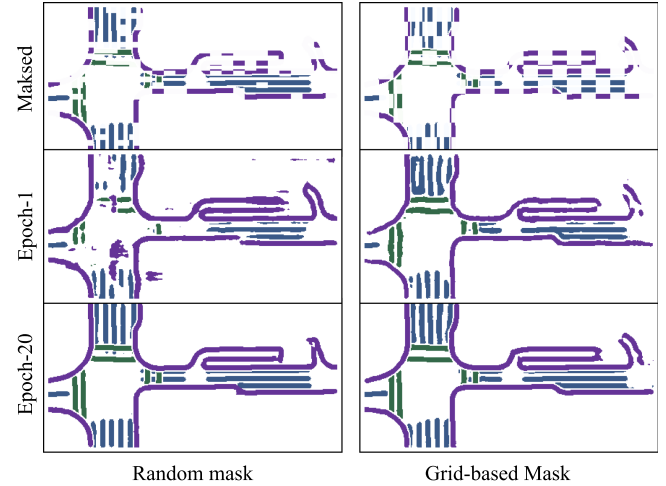


Fig. 3. **Different mask strategies.** “Masked” refers to the pre-training inputs after applying various masking strategies, and “Epoch-1” and “Epoch-20” denote the reconstruction results at the first and twentieth epochs of the pre-training process, respectively.

ineffective, as detailed in Table VI. To tackle this challenge, we adopt a multi-head cross-attention module. This allows the network to utilize cross-attention to determine the most suitably aligned location, thereby effectively enhancing the BEV feature with the SDMap prior.

*BEV Query:* As illustrated in Fig. 2(b), we first utilize a convolutional network to downsample the BEV features. This not only averts excessive memory consumption on low-level feature maps but also partially alleviates the misalignment between the image BEV features and the LiDAR BEV features. The downsampled BEV features are represented as  $\mathcal{B}_{\text{small}} \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times C}$ , where  $d$  is the downsampling factor. These features, combined with sine positional embedding and squeezed into 1D, result in the BEV queries  $\mathcal{Q}_{\text{bev}}$ .

*SDMaps Prior attended:* The associated (though misaligned) SDMap undergoes processing via a convolutional network in conjunction with sine positional embedding, producing the SDMap prior tokens  $\mathcal{F}_{\text{sd}}$ , as shown in Fig. 2(c). Subsequently, the multi-head cross-attention is employed to enhance the BEV queries by integrating the information from SDMap priors. The formal representation is,

$$\mathcal{Q}' = \text{Concat}(\text{CA}_1(\mathcal{Q}_{\text{bev}}, \mathcal{F}_{\text{sd}}), \dots, \text{CA}_m(\mathcal{Q}_{\text{bev}}, \mathcal{F}_{\text{sd}})),$$

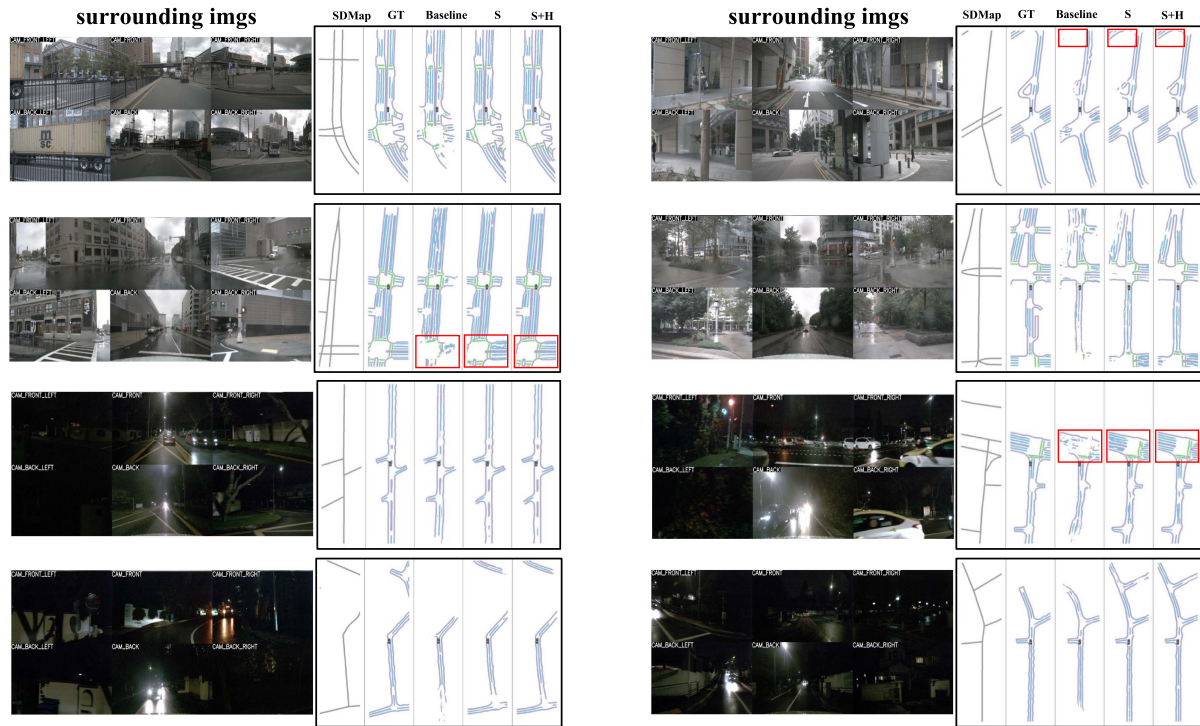
$$\mathcal{B}_{\text{improved}} = \text{layernorm}(\mathcal{Q}_{\text{bev}} + \text{Dropout}(\text{Proj}(\mathcal{Q}'))), \quad (7)$$

where the  $\text{CA}_i$  is the  $i$ -th single head cross-attention,  $m$  is the number of head, key and value embeddings,  $\text{Proj}$  is a projection layer and  $\mathcal{B}_{\text{improved}}$  represents the resized BEV feature derived from the multi-head cross-attention that incorporates the SDMap prior. Subsequently, the improved BEV features pass through a segmentation head to get the initial HDMap element prediction, denoted as  $X_{\text{init}} \in \mathbb{R}^{H \times W \times (N_c + 1)}$ . Here, the  $(N_c + 1)$  channels denote the total number of map element classes, including an additional background class.

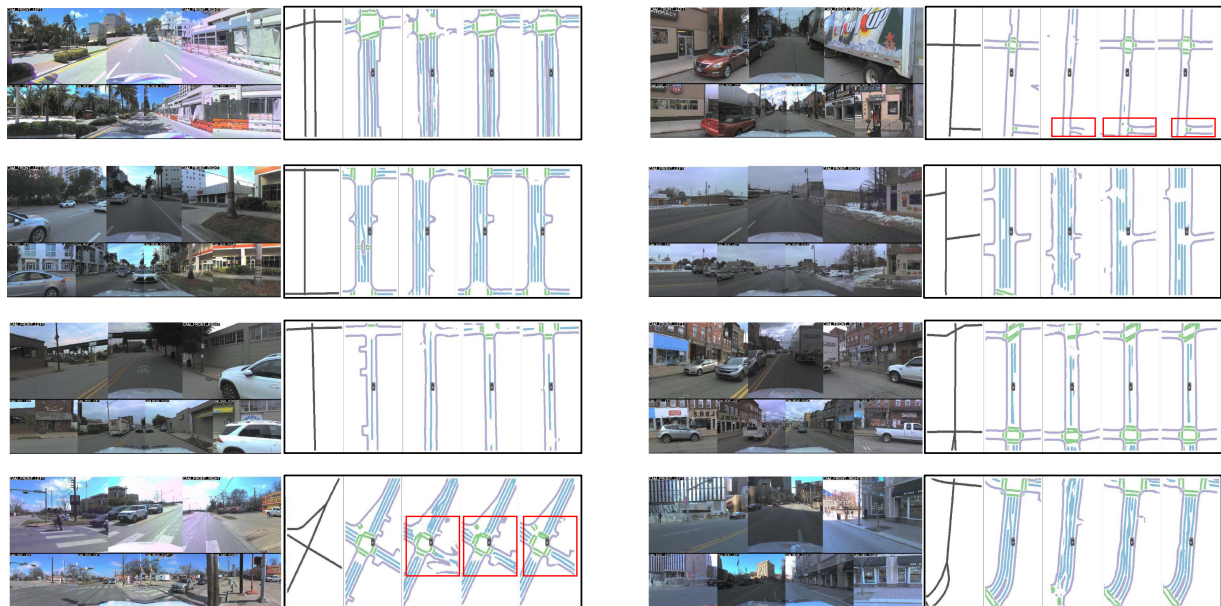
### B. HDMap Prior Module

Then we describe the HDMap prior module, which  $\mathcal{H}$  is computationally heavy (see Fig. 5) which is indeed optional.

<sup>1</sup>[Online]. Available: <https://www.openstreetmap.org/>



(a) Qualitative Results on nuScenes within the range of 240m\*60m.



(b) Qualitative Results on Argoverse2 within the range of 120m\*60m.

Fig. 4. **Qualitative results.** We conduct a comparative analysis within a range of  $240 \times 60$  m on the nuScenes dataset and  $120 \times 60$  m on the Argoverse2 dataset, utilizing C+L as input. In our notation, “S” indicates that our method utilizes only the SDMap priors, while “S+H” indicates the utilization of both. Our method consistently outperforms the baseline method under various weather conditions and in scenarios involving viewpoint occlusion.

Our goal is to acquire a more precise and realistic *far-seeing* HDMap, particularly in challenging scenarios such as inclement weather, areas of occlusion, and regions of invisibility. To enhance the continuity and realism of HDMap generation in these scenarios, closely approximating the distribution of HDMap, we employ an adapted pre-trained MAE module to capture the distribution. Training the HDMap Prior Module has two training steps: the first step involves training the MAE module using

self-supervised learning to capture the HDMap distribution and the second step is fine-tuning by loading weights in the first step and using the initial HDMap prediction  $X_{\text{init}}$  as input, as shown in Function. 5.

*Pre-trained MAE module:* We utilize self-supervised learning to pre-train the masked autoencoder to capture the data distribution of HDMap. As shown in Fig. 2(e), this module consists of a Vision Transformer model [34] and a fully convolutional

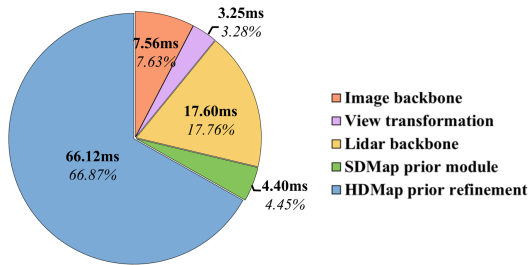


Fig. 5. **Detailed runtime.** We conduct runtime profiling of each component in P-MapNet at a range of  $60 \times 120$  m on one RTX 3090 GPU.

segmentation head. As illustrated in Function. 4, we mask the HDMMap ground truth in the *training* set of the dataset, and then encode this masked HDMMap using the ViT model. Subsequently, given that our reconstruction target is indeed a semantic mask (although treated as images for input), we employ the segmentation head to revert the masked HDMMap back to its original HDMMap ground truth. This process is self-supervised using pixel-wise cross-entropy loss between the HDMMap ground truth and the masked HDMMap. Specifically, we tried two different mask strategies to pre-train the module, namely grid mask and random mask, as shown in Fig. 3. In the random mask strategy, we randomly select both the mask patch size and the mask ratio from a set of candidates to mitigate the over-fitting issue during pre-training.

*End-to-end fine-tuning:* Next we apply the pre-trained MAE module on the initial HDMMap prediction  $X_{init}$  as a refinement plug-in to improve the initially predicted HDMMap, addressing potential issues such as broken or missing lanes in challenging scenarios. We then take the entire model through some lightweight fine-tuning for 10 epochs to better align the distribution of the initial prediction with that of the HDMMap distribution, as illustrated in Function 5.

## V. EXPERIMENTS

### A. Dataset and Metrics

We evaluate P-MapNet on two popular datasets in autonomous driving research, nuScenes [32] and Argoverse2 [33]. To demonstrate our method is a far-seeing solution, we set three distinct perception ranges along the direction of vehicle travel:  $60 \times 30$  m,  $120 \times 60$  m,  $240 \times 60$  m. Additionally, we utilize different map resolutions, specifically 0.15 m for the short range of  $60 \times 30$  m and 0.3m for the rest two longer ranges. We use intersection-over-union (IoU) as the metrics for segmentation results and incorporate a post-processing step to get the vectorized map and evaluate it using the average precision (AP). Following [20], we set the threshold of IoU as 0.2 and threshold of CD as 0.5m, 1.0m, 1.5m. Furthermore, to evaluate the *realism* of the HDMMap prior refinement module output, we utilize a perceptual metric LPIPS [37], which leverages deep learning techniques to more closely simulate human visual perception differences, providing a more precise and human vision-aligned image quality assessment than traditional pixel-level or simple structural comparisons. Implementation details can be found in the supplementary material.

### B. Results

*Comparisons with State-of-the-arts:* We conducted a comparative analysis of our approach with current state-of-the-art

TABLE II  
P-MAPNET ACHIEVES STATE-OF-THE-ART ON NUSCENES VAL SET

Range	Method	Modality	Div.	Ped.	Bound.	mIoU
$60 \times 30$ m	VPN [36] †	C	36.5	15.8	35.6	29.30
	Lift-Splat-Shoot [18] †	C	38.3	14.9	39.3	30.83
	HDMMapNet [1]	C	40.5	19.7	40.5	33.57
	NMP [27] †	C	44.1	21.0	46.1	37.05
	HDMMapNet	C+L	45.9	30.5	56.8	44.40
	P-MapNet (Ours)	C	<b>44.3</b>	<b>23.3</b>	43.8	<b>37.13</b>
P-MapNet (Ours)	C+L	<b>54.2</b>	<b>41.3</b>	<b>63.7</b>	<b>53.07</b>	
$90 \times 30$ m	BEVFusion [35] †	C+L	33.9	18.8	38.8	30.50
	SuperFusion [20] †	C+L	37.0	24.8	41.5	34.43
	P-MapNet (Ours)	C+L	<b>44.73</b>	<b>31.03</b>	<b>45.5</b>	<b>40.64</b>

The symbol “†” denotes results reported in [29], [20], while “NMP” represents the “HDMMapNet+NMP” configuration as described in [27]. For superlong-range perception, we compared with superfusion [20] and BEVfusion [35]. “C” and “L” respectively refer to the surround-view cameras and LiDAR inputs. Ours uses both SDMap and HDMMap priors.

TABLE III  
QUANTITATIVE RESULTS OF MAP SEGMENTATION ON ARGOVERSE2 VAL SET

Range	Method	Div.	Ped.	Bound.	mIoU
$60 \times 30$ m	HDMMapNet	53.0	27.9	44.5	41.80
	P-MapNet (S)	52.9	29.7	46.8	43.13 (+1.33)
	P-MapNet (S+H)	<b>53.5</b>	<b>30.1</b>	<b>47.3</b>	<b>43.63 (+1.83)</b>
$120 \times 60$ m	HDMMapNet	48.3	27.8	40.0	38.70
	P-MapNet (S)	52.5	34.3	48.7	45.17 (+6.47)
	P-MapNet (S+H)	<b>53.1</b>	<b>34.7</b>	<b>49.0</b>	<b>45.60 (+6.90)</b>

We conducted a comparison between the P-MapNet method and HDMMapNet [1], using only surround-view cameras as input, demonstrating superior performance.

(SOTA) approaches in both short-range ( $60m \times 30m$ ) perception and long-range ( $90m \times 30m$ ) with a resolution of 0.15 m. As indicated in Table II, our method exhibits superior performance compared to both existing vision-only and multi-modal (RGB+LiDAR) methods.

*Far-seeing Experiments:* We performed a performance comparison with HDMMapNet [1] at various distances and using different sensor modalities, with the results summarized in Tables I and III. Our method achieves a remarkable 13.4% improvement in mIOU at a range of  $240$  m  $\times$   $60$  m. It is noteworthy that the effectiveness of SD Map priors becomes more pronounced as the perception distance extends beyond or even surpasses the sensor detection range, thus validating the efficacy of SD Map priors. Lastly, our utilization of HD Map priors contributes to additional performance improvements by refining the initial prediction results to be more realistic and eliminating results that are broken and unnecessarily curved, as demonstrated in Fig. 4.

*Perceptual Metric of HDMMap Prior:* The HDMMap Priors Module endeavors to map the network output onto the distribution of HDMMaps to make it more *realistic*. To evaluate the *realism* of the HDMMap prior refinement Module output, we utilize a perceptual metric LPIPS [37] (lower values indicate better performance). The enhancements achieved in the S+H setting are considerably greater when compared to those in the S-only setting as demonstrated in Table IV.

*Vectorization Results:* We also conducted a comparison of vectorization results by employing post-processing to obtain vectorized HD Maps. As detailed in Table I, we achieve the best instance detection AP results across distance ranges.

*Does SDMap prior work for direct vectorized map prediction?* As demonstrated in Table V, to confirm the universality of our SDMap prior, we integrated our SDMap Prior Module into MapTR [22] (with only minor modifications), an end-to-end framework, referred to as the MapTR-SDMap method. Our MapTR-SDMap method also led to a significant improvement in

TABLE IV  
PERCEPTUAL METRIC OF HDMA PRIORITY

Range	Method	Modality	mIoU $\uparrow$	LPIPS $\downarrow$	Modality	mIoU $\uparrow$	LPIPS $\downarrow$
120 $\times$ 60 m	Baseline	C	33.77	0.8050	C+L	49.50	0.7872
	P-MapNet (S)	C	40.33 (+6.56)	0.7926 (1.54%)	C+L	60.20 (+10.70)	0.7607 (3.37%)
	P-MapNet (S+H)	C	<b>40.87 (+7.10)</b>	<b>0.7717 (4.14%)</b>	C+L	<b>61.77 (+12.27)</b>	<b>0.7124 (9.50%)</b>
240 $\times$ 60 m	Baseline	C	26.77	0.8484	C+L	36.47	0.8408
	P-MapNet (S)	C	42.20 (+15.43)	0.8192 (3.44%)	C+L	48.87 (+12.40)	0.8097 (3.70%)
	P-MapNet (S+H)	C	<b>45.50 (+18.73)</b>	<b>0.7906 (6.81%)</b>	C+L	<b>49.93 (+13.46)</b>	<b>0.7765 (7.65%)</b>

We utilize the LPIPS metric to evaluate the realism of S+H models on 120m  $\times$  60m perception range. And the improvements in the S+H setting are more significant compared to those in the S-only setting.

TABLE V  
COMPARISONS WITH MAPTR [22] ON NUSCENES VAL SET

Range	Method	Div.	Ped.	Bound.	mAP	mAP <sub>raster</sub>
60 $\times$ 30 m	MapTR [22]	49.50	41.17	51.08	47.25	24.33
	MapTR-SDMap	<b>50.92</b>	<b>43.71</b>	<b>53.49</b>	<b>49.37 (+2.21)</b>	<b>26.60 (+2.27)</b>
	P-MapNet	26.08	17.66	48.43	30.72	-
120 $\times$ 60 m	MapTR [22]	26.00	18.89	15.73	20.20	14.90
	MapTR-SDMap	<b>27.23</b>	21.95	19.50	22.89 (+2.69)	<b>19.37 (+4.47)</b>
	P-MapNet	19.50	<b>24.72</b>	<b>42.48</b>	<b>28.90</b>	-
240 $\times$ 60 m	MapTR [22]	12.69	7.17	4.23	8.03	7.50
	MapTR-SDMap	<b>22.74</b>	16.34	10.53	16.53 (+8.50)	<b>11.73 (+4.23)</b>
	P-MapNet	14.51	<b>25.63</b>	<b>28.11</b>	<b>22.75</b>	-

We conducted a comparison between MapTR fused with the SDMap prior method (MapTR-SDMap) and the vanilla MapTR [22], using only cameras as input. We used CD-based metric [1] and rasterization-based metric (mAP<sub>raster</sub>) from [25], which is more sensitive to vectorized map accuracy.

TABLE VI  
ABLATIONS ABOUT SD MAPS FUSION STRATEGIES

Fusion Method	Div.	Ped.	Bound.	mIoU
w/o SDMap	53.2	36.9	57.1	49.07
w/o SDMap, w/ Self.Attn.	57.7	42.0	60.6	53.43
Simply-concat	59.4	43.2	61.6	54.73
CNN-concat	60.2	45.5	63.1	56.27
Cross.Attn.	<b>63.6</b>	<b>50.2</b>	<b>66.8</b>	<b>60.20</b>

The experiments are conducted with range of 120  $\times$  60m and C+L as inputs. "w/o SDMap" is the baseline [1]. "w/o SDMap, w self.Attn" only employed BEV queries self-attention.

mean Average Precision (mAP). We also use the AP<sub>raster</sub> metric from MapVR [25], which is more objective and accurate for vectorized maps. Since this metric is unsuitable for comparing rasterized and vectorized representation methods, we did not report results for our method.

### C. Ablation Study

All ablative experiments are conducted on nuScenes val set with a perception range of 120m  $\times$  60m and the camera-LiDAR fusion(C+L) configuration.

*Detailed Runtime:* In Fig. 5, we provide the detailed wallclock runtime of each component in P-MapNet with both camera and LiDAR inputs. As a recap, a complete FPS evaluation is reported in Table I. As shown by the profiling, the HDMap prior is computationally heavy but it is indeed optional. Practitioners can switch between the SDmap only setting or the SDMap+HDMap setting, depending on the computation overhead (e.g., onboard or offboard).

*SDMap Prior Fusion Strategies:* To validate the effectiveness of our proposed fusion approach for SDMap priors, we experimented with various fusion strategies, the details of which are summarized in Table VI. In an initial evaluation, a straightforward concatenation (termed "Simple-concat") of the

TABLE VII  
ROBUSTNESS TO EGO-POSE LOCALIZATION ERROR

Noise type	$\sigma$	Div.	Ped.	Bound.	mIoU
Baseline	$\circ$	53.6	37.8	57.1	49.50
	P-MapNet (S)	<b>63.6</b>	<b>50.2</b>	<b>66.8</b>	<b>60.20</b>
Translation	3m	62.0	47.9	65.5	58.51
	6m	58.9	43.8	62.9	55.19
	9m	55.3	39.7	60.0	51.69
	12m	52.1	36.2	57.4	48.56
Rotation	3 $^\circ$	60.7	47.8	64.5	57.65
	6 $^\circ$	55.7	44.0	60.3	53.34
	9 $^\circ$	50.1	39.7	55.9	48.55
All	3m, 3 $^\circ$	59.4	45.9	63.4	56.24
	6m, 3 $^\circ$	56.7	42.7	61.2	53.51
	6m, 6 $^\circ$	52.4	39.5	57.7	49.87

To validate the robustness of our SDMap prior module against SDMap misalignment, we add translation and rotation gaussian noise,  $\mathcal{N}(0, \sigma^2)$ , to the ego-poses, respectively.

rasterized SDMap with BEV features led to a mIoU boost of about 5%. A better approach, where we exploit CNNs to encode and concatenate the rasterized SDMap, furthered this improvement to about 7%. Nonetheless, the straightforward concatenation techniques were hampered by spatial misalignment issues, preventing the full capitalization of the SDMap priors' potential. Interestingly, leveraging self-attention solely for BEV queries also enhanced performance. Among all the approaches tested, our method anchored on cross-attention demonstrated the most substantial gains.

*Robustness to Ego-pose Localization Error:* To validate the robustness of the SDMap Prior Module against SDMap misalignment, we introduce Gaussian noise,  $\mathcal{N}(0, \sigma^2)$ , to the ego-poses. This includes adding Gaussian noise to both the translation and rotation, to simulate real-world localization errors. As shown in Table VII, the SDMap Prior Module can perform comparable to the baseline under the deviation of 12m for translation Gaussian error, 9 $^\circ$  for rotation Gaussian error, and 6m with 6 $^\circ$  for combined error.

*Ablation of BEV-SDPrior Cross-attention Layers:* As the number of transformer layers increases, performance of our method improves, but it eventually reaches a point of saturation since the SDMap priors contain low-dimensional information, and excessively large network layers are susceptible to overfitting, as shown in Table VIII.

*The generalization capability of HDMap MAE:* In order to verify the generalizability of our HDMap Prior refinement module, we pre-train on Argoverse2 and nuScenes datasets respectively, and fine-tune on nuScenes dataset and test the prediction results mIOU. The results are shown in the Table IX, and it can be seen that the model pre-trained on Argoverse2 is only

TABLE VIII  
ABLATIONS ABOUT THE NUMBER OF BEV-SDPRIOR  
CROSS-ATTENTION LAYERS

Attention Layer	Div.	Ped.	Bound.	mIoU	Memory (GB)	FPS
1	62.6	48.4	65.6	58.87	19.03	19.60
2	<b>63.6</b>	<b>50.2</b>	<b>66.8</b>	<b>60.20</b>	20.20	18.56
4	60.6	44.9	63.2	56.23	23.24	18.45
6	58.7	42.4	61.8	54.30	OOM	-

During training, we evaluated memory usage with a batch size of 4, while for inference, we measured frames per second (FPS) with a batch size of 1.

TABLE IX  
CROSS-DATASET EXPERIMENT OF HDMAP PRIORS

Pre-Train Dataset	Div.	Ped.	Bound.	mIoU $\uparrow$	LPIPS $\downarrow$
Argoverse v2	64.5	51.3	67.6	61.13 (+0.93)	0.7203 (8.49%)
Nuscense	65.3	52.0	68.0	61.77 (+1.57)	0.7124 (9.50%)

We pre-trained the HDMAP prior module on argoverse2 and nuScenes datasets, respectively, and tested it on nuScenes val, using a range of  $120 \times 60m$  and RGB+LiDAR inputs.

0.64% mIOU lower than the pre-trained model on nuScenes, which can prove that our refinement module indeed captures the HDMAP priors information with generalization capability rather than overfitting to the dataset.

## REFERENCES

- Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMAPNet: An online HD map construction and evaluation framework," in *2022 IEEE Int. Conf. Robot. Automat.*, 2022, pp. 4628–4634.
- Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized HD map learning," 2022, *arXiv:2206.08920*.
- Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.
- H. Wang et al., "Openlane-V2: A topology reasoning benchmark for unified 3D HD mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36.
- Y. Zheng et al., "Steps: Joint self-supervised nighttime image enhancement and depth estimation," in *2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 4916–4923.
- Z. Wu et al., "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *Proc. CAAI Int. Conf. Artif. Intell.*, 2023, pp. 3–15.
- B. Jin et al., "Adapt: Action-aware driving caption transformer," in *2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 7554–7561.
- P.-E. Sarlin et al., "Orienternet: Visual localization in 2D public maps with neural matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21632–21642.
- Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Comput. Vis.—ECCV 2022, 17th Eur. Conf.*, Oct. 23–27, 2022, pp. 1–18.
- B. Tian, M. Liu, H.-a.-a. Gao, P. Li, H. Zhao, and G. Zhou, "Unsupervised road anomaly detection with language anchors," in *2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 7778–7785.
- B. Tian et al., "Latency-aware road anomaly segmentation in videos: A photorealistic dataset and new metrics," 2024, *arXiv:2401.04942*.
- H. Zhao, M. Lu, A. Yao, Y. Guo, Y. Chen, and L. Zhang, "Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 10–18.
- X. Chen, H. Zhao, G. Zhou, and Y.-Q. Zhang, "PQ-Transformer: Jointly parsing 3D objects and layouts from point clouds," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2519–2526, Apr. 2022.
- H.-A. Gao et al., "From semi-supervised to omni-supervised room layout estimation using point clouds," in *2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2803–2810.
- H.-A. Gao, B. Tian, P. Li, H. Zhao, and G. Zhou, "DQS3D: Densely-matched quantization-aware semi-supervised 3D detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21905–21915.
- Z. Bao, S. Hossain, H. Lang, and X. Lin, "High-definition map generation technologies for autonomous driving," 2022, *arXiv:2206.05400*.
- J. Houston et al., "One thousand and one hours: Self-driving motion prediction dataset," in *Proc. Conf. Robot Learn.*, 2021, pp. 409–418.
- J. Phillion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Comput. Vis.—ECCV 2020, 16th Eur. Conf.*, Aug. 23–28, 2020, pp. 194–210.
- A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 IEEE Int. Conf. Robot. Automat.*, 2022, pp. 9200–9206.
- H. Dong et al., "Superfusion: Multilevel LiDAR-camera fusion for long-range HD map generation and prediction," 2022, *arXiv:2211.15656*.
- B. Liao et al., "Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction," 2023, *arXiv:2303.08815*.
- B. Liao et al., "MAPTR: Structured modeling and learning for online vectorized HD map construction," 2022, *arXiv:2208.14437*.
- W. Ding, L. Qiao, X. Qiu, and C. Zhang, "PivotNet: Vectorized pivot learning for end-to-end HD map construction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3672–3682.
- T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Streammapnet: Streaming mapping network for vectorized online HD map construction," 2023, *arXiv:2308.12570*.
- G. Zhang et al., "Online map vectorization for autonomous driving: A rasterization perspective," 2023, *arXiv:2306.10502*.
- A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
- X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Neural map prior for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17535–17544.
- W. Gao, J. Fu, Y. Shen, H. Jing, S. Chen, and N. Zheng, "Complementing onboard sensors with satellite map: A new perspective for HD map construction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 11103–11109.
- Z. Xie, Z. Pang, and Y.-X. Wang, "MV-Map: Offboard HD-map generation with multi-view consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8658–8668.
- K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
- M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct.–Dec. 2008.
- H. Caesar et al., "nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.
- B. Wilson et al., "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proc. NeurIPS Datasets Benchmarks*, 2021.
- A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- Z. Liu et al., "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," 2022, *arXiv:2205.13542*.
- B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Automat. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.
- R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.