Semantic Map Guided Bird's-Eye View Learning for Online HD Map Construction

Huantao Ren*
Syracuse University, NY, USA
hren11@syr.edu

Hesham M. Eraqi
Last Mile, Amazon, WA, USA
heraqi@amazon.com

ABM Musa
Last Mile, Amazon, WA, USA
musaabm@amazon.com

Mohamed Moustafa
Last Mile, Amazon, WA, USA
mmoustm@amazon.com

Vectorized High-Definition (HD) maps offer rich and precise environmental information about driving scenes, playing a crucial role in improving driver safety by supporting autonomous driving and advanced driver-assistance systems (ADAS). Processing individual camera images creates fragmented view of the world requiring complex and error-prone merging. Existing multi-view camera methods train deep neural networks to directly generate a unified bird's-eye view (BEV) features used to learn HD map construction. Nevertheless, a significant limitation is the lack of direct supervision of the learned BEV features based on the ground-truth map elements. To overcome this limitation, we propose a novel method, referred to as Semantic Map Guidance (SMG), for explicit alignment of the learned BEV features and the corresponding semantic representations by utilizing ground-truth label during training. We demonstrate the effectiveness of the proposed SMG method by incorporating it into multiple state-of-the-art BEV-based methods for online HD map construction task. We perform extensive experiments on two widely used HD map datasets, nuScenes and Argoverse 2, demonstrating that SMG, without any bells and whistles, consistently improves the accuracy of all the tested networks by using the same base network implementation and hyperparameters without any additional inference time.

1. Introduction

High-definition (HD) maps provide highly instance-level vectorized representations, such as lane dividers, road boundaries, pedestrian crossings, and more features. The rich semantic information about road topology and traffic rules is crucial for improving navigation efficiency and enabling both human drivers and autonomous systems to make safer and more efficient decisions on the road. Traditionally, HD maps were built offline using SLAM-based methods [27, 28, 33], which involve complex pipelines and high maintenance. Recent online approaches generate maps dynamically at runtime with onboard sensors, reducing manual effort.

Early works [3, 17] use line-shape priors to detect lanes from front-view cameras, while other single-view methods [9, 29] detect driving scene elements from one camera. Multi-camera setups require merging fragmented views, which is complex and error-prone. With advancements in bird's-eye view (BEV) representation learning [14, 21, 22], recent multi-view methods [18, 24, 25, 31] generate unified BEV maps directly from multiple cameras, leveraging complementary views to improve depth estimation, occlusion handling, and scene unification, making them more accurate and practical for safety-critical applications.

However, current HD map construction approaches that use BEV representations, such as MapTR [15], ensure the quality of BEV features indirectly by reconstructing specific map elements, such as lanes or road boundaries, through a transformer decoder. While these models are trained end-to-end and intermediate BEV features are optimized indirectly, the supervision signal comes primarily from the final map reconstruction. As a result, the learned BEV features may not be explicitly encouraged to align with the ground-truth map structure. To address this, we propose aligning generated BEV features directly with ground-truth semantic and spatial information, providing an additional training signal that can help the model better capture map element classes and boundaries.

To demonstrate the value of aligning BEV features with ground-truth BEV output, we conduct a preliminary experiment in nuScenes dataset [1], using MapTR as the baseline model. We convert the ground-truth semantic maps into masked binary class images and use UNet [26] to process these images. This constrains the output feature map to be

*Work done while an intern at Amazon.

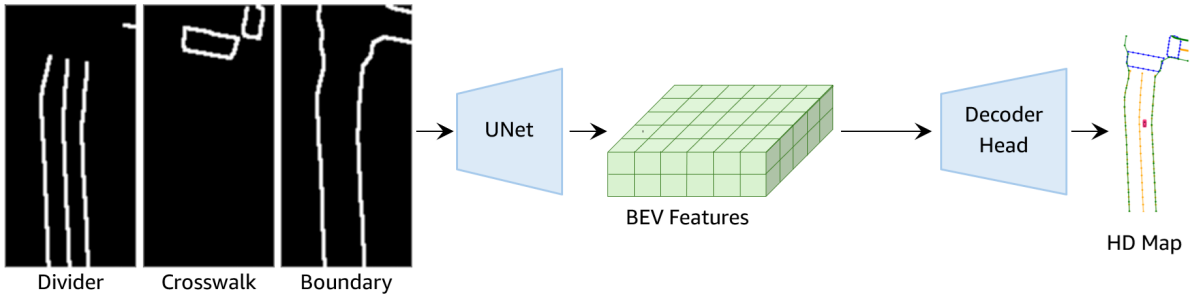

Figure 1. Preliminary experiment of replacing MapTR BEV encoder output features with perfectly aligned masked binary class map images from the ground-truth BEV HD maps, processed by UNet. The HD map decoder and prediction head are kept unchanged. The significant mAP increase across all map elements demonstrates the advantage of aligning the BEV features with the target ground-truth semantic maps.

	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	mAP
MapTR	51.06	43.92	52.55	49.18
MapTR - perfectly-aligned BEV feature	86.15	47.91	71.27	68.44

Table 1. Comparison of the original MapTR with its variant that replaces BEV features with features from UNet that are perfectly aligned with BEV HD maps

perfectly aligned with the ground-truth HD map. Then we pass that output, instead of original BEV encoder output in MapTR, to the HD map generation decoder head while keeping the decoder and prediction head unchanged. The pipeline is illustrated in Fig. 1. Feeding the model with this enriched ground-truth BEV features, we show that ground-truth BEV information provides valuable spatial and semantic cues, highlighting the natural alignment between HD maps and BEV representation. The results, summarized in Tab.1, show that using masked class images improves mean Average Precision (mAP) by 19.26% over the original MapTR and significantly outperforms it across all map elements, demonstrating the significant advantage of aligning the BEV features with the target ground-truth semantic maps.

However, in the experiment above, ground-truth information is used during inference, albeit indirectly, which is not permissible. To address this, we propose the Semantic Map Guidance (SMG) Module, which incorporates direct supervision during the encoding phase but is discarded during inference, enhancing BEV feature generation without relying on ground-truth information at test time. Specifically, SMG guides the BEV encoding process by extracting ground-truth semantic features and using spatial information to align them with the corresponding regions in the generated BEV features. Here, we employ a contrastive learning technique [5, 25] to align the generated BEV with the ground-truth BEV, ensuring precise alignment of BEV elements according to their class labels. This explicit alignment, driven by ground-truth semantic and spatial information guidance, strengthens the model's perceptual capabilities, thereby improving the accuracy of HD map construction from the BEV feature map.

In this paper, we propose a simple yet effective SMG module that enhances BEV representations by explicitly integrating ground-truth semantic and spatial feature guidance into the BEV encoding process. SMG is a plug-and-play module that can be seamlessly integrated with existing BEV-based methods to improve their accuracy in an end-to-end training manner. The module is applied exclusively during training, ensuring that it incurs no additional computational overhead at inference time. We present extensive experimental results and comparisons with several state-of-the-art (SOTA) methods on two widely used HD map construction datasets, namely nuScenes [1] and Argoverse2 [30]. The results demonstrate that incorporating the proposed SMG module consistently enhances HD map construction performance across these networks. We conduct a detailed ablation study, showing that using either a simple multilayer perceptron (MLP) or the more sophisticated CLIP text encoder [25] to extract semantic features consistently improves performance across multiple baseline models. In addition, we examine the impact of leveraging different types of ground-truth information and various alignment losses, demonstrating how these factors influence the performance of the networks.

2. Related Work

2.1. HD Map Construction

With the recent advancements in perspective-to-BEV (PV-to-BEV) techniques [3, 30], HD map construction has increasingly been approached as a segmentation task using surround-view images captured by vehicle-mounted cameras. Several works [4, 10, 14, 22] generate rasterized maps by performing semantic segmentation in the BEV space. To construct vectorized HD maps, HDMapNet [13] relies on grouping pixel-wise semantic segmentation outputs through

heuristic and time-consuming post-processing steps to produce vectorized representations. In contrast, VectorMapNet [18] introduces the first end-to-end solution, employing a two-stage coarse-to-fine pipeline with an autoregressive decoder that predicts points sequentially. However, the auto-regressive model of VectorMapNet leads to a long training time. MapTR [15] employs a one-stage Transformer framework built upon DETR [2], using permutation-equvalent point set modeling. The improved version, MapTRv2 [16], incorporates auxiliary dense segmentation supervision heads, decoupled self-attention in the decoder, and a one-to-many matching strategy, resulting in significant performance gains. MapVR [32] produces a vectorized map using differentiable rasterization, which enables instance-level segmentation supervision. StreamMapNet [31] is the first work to leverage temporal information from past frames to enable online HD map estimation from streaming data. Departing from point set representations, BeMapNet [24] adopts an instance-level approach using a piecewise Bézier head. Similarly, PivotNet [6] transforms point-level representations into instance-level representations through a point-to-line mask module. Different from prior works, rather than designing a new model from scratch, we introduce a flexible module that can be plugged into existing methods to enhance their performance and efficiency.

2.2. Bird's-Eye View Perception

BEV perception methods have been widely explored in 3D object detection and BEV segmentation tasks. It can be broadly classified into two categories [19]: bottom-up and top-down approaches. Bottom-up methods follow a forward process that transforms perspective-view features into 3D space. Lift-Splat-Shoot (LSS) [22] is representative of this approach, it lifts features from 2D to 3D space by predicting a categorical distribution over depth and a context vector. The outer product determines the feature at each point along the perspective ray, enabling a more accurate approximation of the true depth distribution. However, LSS-related methods suffer from high memory and computation costs due to explicit computation, storage, and preprocessing of large frustum features. BEVPool v2 [11] overcomes this by using precomputed voxel and frustum indices to directly access features during processing, eliminating the need for heavy frustum feature handling, which greatly reduces memory use and speeds up inference.

Top-down methods directly construct BEV queries, which then search for and attend to corresponding features in perspective images using a cross-attention mechanism. BEVFormer [14] leverages deformable attention mechanisms to enable interaction between dense BEV-plane queries and multi-view image features. It introduces a set of historical BEV queries and leverages temporal information through deformable attention between the current and historical queries. SimpleBEV [8] applies a variation of Inverse Perspective Mapping (IPM) [20] to sample features from 2D images onto predefined BEV anchor points. GKT [4] uses fixed-offset deformable attention on unfolded 2D kernel regions around reference points, enabling a fixed BEV-to-pixel mapping with a BEV-to-2D look-up table for fast inference.

2.3. Feature Alignment

Feature alignment is a key objective in many computer vision tasks, such as image classification, detection, and cross-modal retrieval. Techniques like contrastive learning (e.g., SimCLR [5]) have been widely adopted to enforce representation similarity by pulling positive pairs together and pushing negatives apart in the feature space. CLIP [25] extends this idea to align image and text embeddings across modalities. Previous works [7, 34] typically use two separate pathways to combine different types of data and align the features. Mean Squared Error (MSE) is another alignment method. In the field of online map segmentation using BEV representations, PCT [12] employs MSE loss to align features from the teacher and student models for unsupervised domain adaptation. Different from prior approaches that align features across modalities or between models, we align features from the ground-truth semantic and spatial information directly with the learned BEV features. This alignment acts as a training signal without adding parameters or affecting inference.

3. Semantic Map Guidance (SMG)

We present the technical details of our proposed method illustrated in Figure 2. The core innovation of our proposed model lies in its incorporation of ground-truth features to guide the generation of BEV representations. We first review the architecture of the previous methods, in Sec. 3.1. Next, we explore the details of the Semantic Map Guidance (SGM) Module in Sec. 3.2. The training loss function for our framework is presented in Sec. 3.3.

3.1. Preliminary

In many other BEV-based HD map construction models, the processing pipeline begins with multi-view camera images X_{views} which are first processed through an image backbone. This is followed by a BEV base model that integrates relevant image features into a unified top-down BEV feature map $X_{bev} \in \mathbb{R}^{H \times W \times C}$, where H, W, C denotes height, width, and number of feature channels of produced BEV feature map, respectively. The BEV base model leverages a series of self-attention and cross-attention mechanisms to integrate relevant image features and produce a top-down representation of the scene. Based on a transformer architecture, a set of decoder queries q_{dec} is initialized to ex-

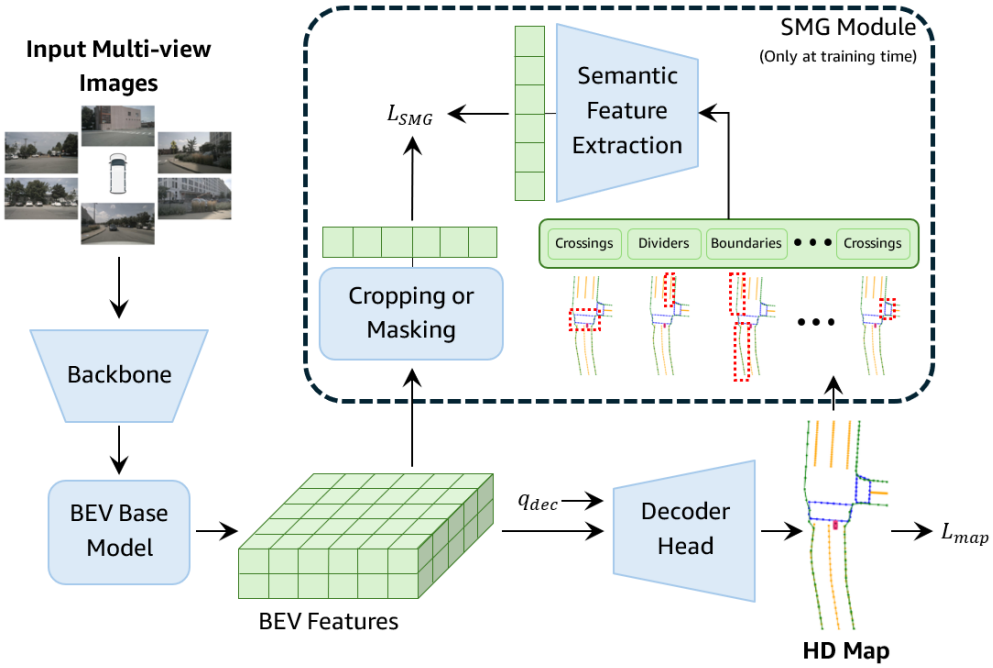

Figure 2. Proposed SMG module directly supervise the generated BEV representations by utilizing a semantic feature extractor to extract semantic features and a contrastive learning framework to enhance the quality of the BEV representations.

tract information from the BEV feature map. These queries are responsible for decoding spatial and semantic information specific to the map features (e.g., roads, pedestrian crossings) within the receptive field. The informed decoder queries are then passed to a map construction head to generate the predicted vectorized map, each query corresponds to a prediction for a specific map feature or an empty region. Afterwards, the final map construction loss is computed by applying a loss function between the predicted positions and classes of the queries and their corresponding ground-truth map features. Formally, this process can be expressed as:

$$X_{bev} = BEVBase(X_{views}), \tag{1}$$

$$q^{'}_{dec} = Dec(X_{bev}, q_{dec}), \tag{2}$$

$$q_{pred} = Head(q^{'}_{dec}, y), \tag{3}$$

$$L_{map} = L_{position} + L_{class}, \tag{4}$$

where $BEVBase$ and Dec denote the BEV base model and perception decoder, respectively, y denotes the ground-truth of the task. $L_{position}$ and L_{class} represent the point position regression loss and classification loss, respectively. If the used decoder isn't transformer-based, the q_{dec} input is removed from the above formulation.

3.2. Semantic Map Guidance (SMG) Module

In previous models, the lack of direct supervision on the BEV representation often results in inaccurate alignment of object classes and boundaries with the ground-truth which is in BEV perspective. This misalignment limits the model's ability to construct precise map features, leading to inherent constraints in the quality and accuracy of the generated vectorized map. To address this limitation, we propose the SMG module. The primary goal of SMG is to align the generated BEV representation with the features from ground-truth information, ensuring a precise organization of BEV elements according to their class labels, positions, and boundaries.

As illustrated in Fig. 2, in SMG, to enable the generated BEV features to incorporate ground-truth semantic information, we first feed semantic information c_i of the i^{th} instance on the BEV map into a semantic feature extractor SFE:

$$g^i = SFE(c_i), \tag{5}$$

where g^i with the same feature dimension C as the BEV feature, represents the encoded ground-truth semantic feature of the i^{th} instance on the BEV map.

To incorporate ground-truth spatial information and precisely align the semantic features derived from ground-truth labels with BEV features, we crop the corresponding regions in the produced BEV feature. This can be done using either bounding box coordinates or point positions. Point-based cropping offers fine-grained localization, whereas bounding boxes provide broader coverage. To balance precision and coverage, we adopt two bounding boxes for road boundaries and a single bounding box for the other two classes of dividers and crosswalks. Road boundaries

often exhibit irregular shapes; thus, two bounding boxes better capture their geometry while reducing irrelevant regions. In contrast, both crosswalks and lane dividers can be effectively enclosed within a single bounding box, since crosswalks are polygon-shaped and lane dividers are always straight lines. As illustrated in Fig. 3, we visualize the cropping process using a point, a single bounding box, and two bounding boxes to provide a clear and intuitive understanding of the approach. For crosswalks and dividers, the bounding box b_i is computed from the minimum and maximum x and y coordinates of all points. For road boundaries, the two bounding boxes b_{i1} and b_{i2} are obtained by sorting the point set, splitting it into two subsets, and calculating the min–max coordinates separately for each. Supporting experimental results for different cropping strategies are provided in Sec. 5.1.

After cropping the corresponding region in the BEV feature map, a pooling operation is then applied to the cropped tensor, which serves as the representation for the corresponding object, the process can be expressed below:

$$o_i = Pool(Crop(X_{bev}, b_i)). \tag{6}$$

To bring the learned BEV features and semantic embeddings closer together, we leverage contrastive learning to refine the relationships and spatial distances between elements in the BEV feature space, as expressed in the following formulation:

$$\mathcal{L}_{\text{SMG}} = -\frac{1}{2}\left(\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\text{sim}(g_i,o_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\text{sim}(g_i,o_j)}{\tau}\right)} + \sum_{i=1}^{N}\log\frac{\exp\left(\frac{\text{sim}(o_i,g_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\text{sim}(o_i,g_j)}{\tau}\right)}\right), \tag{7}$$

where $sim(g_i, o_i)$ denotes the similarity between the ground-truth feature g_i and the learned feature o_i for the i^{th} instance. This similarity is computed using cosine similarity, which measures the alignment between the two feature vectors in terms of their direction:

$$\text{sim}(g_i, o_i) = \frac{g_i \cdot o_i}{\|g_i\|\|o_i\|}. \tag{8}$$

After that process, the produced BEV features could learn the information from both ground-truth semantic and spatial cues.

3.3. Loss

The training loss formulation for our method consists of two components: the baseline map loss L_{map} and semantic supervision loss L_{SMG}:

$$L = \lambda_1 L_{map} + \lambda_2 L_{SMG}, \tag{9}$$

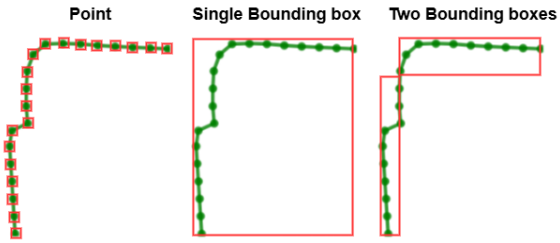

Figure 3. Comparison of cropping the produced BEV feature using point position, one bounding box and two bounding boxes.

where λ_1 and λ_2 are hyper-parameters to balance the two loss components. In this work, we set both parameters to 1. The semantic supervision, applied solely during training via the SMG module, introduces no additional parameters or computational overhead during inference. This ensures that the efficiency of the original model is preserved at the inference stage.

4. Experiments

4.1. Dataset and Metrics

Datasets. To evaluate our proposed approach, we conduct experiments on the widely used nuScenes dataset [1], which contains 28,130 training samples and 6,019 validation samples, corresponding to 700 and 150 driving scenes, respectively. Each scene comprises roughly 40 samples, each sample contains RGB images from 6 cameras of the same model, each with a horizontal field of view of about 70 degrees. Together, these cameras provide a full 360 degrees horizontal view around the ego-vehicle. For a fair comparison, all experimental settings and metrics are kept consistent with baseline models.

Additionally, we evaluate our method on the Argoverse2 dataset [30], which includes 1,000 driving sequences, each capturing 15 seconds of data. The dataset provides 20 Hz RGB images from seven high-resolution ring cameras (2048×1550), 10 Hz LiDAR sweeps, and a 3D vectorized map. The train, validation, and test splits comprise 700, 150, and 150 logs, respectively. Note that unlike StreamMapNet [31], which re-splits the official datasets, we retain the original splits used by MapTR [15] and MapTRv2 [16].

Following previous works [16, 31], we report results on the validation sets of both the Argoverse 2 and nuScenes datasets, focusing on the same three map elements as defined in nuScenes, lane dividers, pedestrian crosswalks, and road boundaries. For a fair comparison, all experimental settings and evaluation metrics are kept consistent with the baseline models, and all experiments are conducted on 8 NVIDIA A100 GPUs.

Metrics. We adopt the standard evaluation metrics com-

Method	BEV Base Model	BEV Size	SMG Input	SMG Encoder	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	$AP_{0.5m}$	$AP_{1.0m}$	$AP_{1.5m}$	mAP	FPS
MapTR-nano	GKT	200 × 100	-	-	49.9	39.6	48.2	-	-	-	45.9	35.0
MapTR	GKT	200 × 100	-	-	51.06	43.92	52.55	24.92	53.92	68.71	49.18	15.1
MapTR+SMG			One-hot	MLP	**53.23**	**47.99**	53.26	26.73	56.49	**71.25**	51.49	
MapTR+SMG			Text	CLIP-Text	53.15	47.67	**54.68**	**27.74**	**56.91**	70.86	**51.83**	
MapTR	BEVFormer	200 × 100	-	-	50.05	44.85	52.18	24.44	53.64	69.45	49.18	15.0
MapTR+SMG			One-hot	MLP	**52.48**	**47.39**	53.97	**25.5**	**57.01**	**71.33**	**51.28**	
MapTR+SMG			Text	CLIP-Text	51.73	43.91	**54.19**	25.11	54.65	70.07	49.95	
MapTR	BEV Pool v2	200 × 100	-	-	52.04	43.95	52.7	24.19	54.73	69.76	49.56	14.7
MapTR+SMG			One-hot	MLP	53.51	**48.3**	54.26	**27.43**	**57.71**	71.29	**52.14**	
MapTR+SMG			Text	CLIP-Text	**53.62**	47.62	**54.93**	26.92	57.52	**71.73**	52.06	
MapTR V2	BEVPool v2	200 × 100	-	-	60.19	59.23	61.64	37.89	66.21	76.95	60.35	14.1
MapTR V2+SMG			One-hot	MLP	**62.53**	**59.53**	**62.99**	**38.75**	**67.84**	**78.47**	**61.69**	
MapTR V2+SMG			Text	CLIP-Text	62.32	58.09	62.32	38.48	65.99	78.26	60.93	
StreamMapNet	BEVFormer	50 × 100	-	-	66.19	61.14	60.96	37.67	69.15	81.46	62.76	14.2
StreamMapNet+SMG			One-hot	MLP	**67.15**	**62.93**	**62.47**	**40.14**	**70.86**	81.55	**64.18**	
StreamMapNet+SMG			Text	CLIP-Text	66.56	62.7	61.68	38.93	70.23	**81.78**	63.65	

Table 2. Performance comparison with baseline methods on nuScenes Validation set. FPS values are measured on the same machine equipped with an RTX 3090. A dash ("–") indicates unavailable results.

monly used in prior works [15, 16, 31]. The perception area is set to [-15m, 15m] along the X-axis and [-30m, 30m] along the Y-axis. Average Precision (AP) is used to assess map construction quality, with Chamfer Distance D_{Chamfer} determining matches between predictions and ground-truth. We compute AP_τ under multiple D_{Chamfer} thresholds $\tau \in \{0.5, 1.0, 1.5\}$ and report: (1) per-class AP, evaluating each map element category (pedestrian crossings, lane dividers, and road boundaries); (2) per-threshold AP, capturing performance at each distance threshold; and (3) mean Average Precision (mAP), the average across all classes and thresholds for an overall score.

4.2. Performance on nuScenes

To demonstrate the generalizability of the proposed SMG module and its ability to enhance various networks, we integrated it into MapTR [15] using different BEV encoders. These include top-down methods such as GKT [4] and BEVFormer [14], the bottom-up method BEVPool v2 [11], as well as state-of-the-art approaches like MapTR v2 [16] and StreamMapNet [31]. The results is summarized in Tab. 2. As shown, the proposed SMG module consistently improves performance across all five models and map elements, including Average Precision (AP) for dividers, crossings, and boundaries, AP at different distance thresholds (0.5m, 1.0m, and 1.5m), as well as the overall mean Average Precision (mAP). For MapTR with BEVPool v2 as the encoder, our method using one-hot class labels and text descriptions as input outperforms the baseline by 2.58% and 2.5% mAP, respectively. For more recent methods, incorporating SMG yields improvements of 1.34% mAP for MapTR v2 and 1.42% for StreamMapNet.

The consistent improvements across various model configurations and architectures, whether incorporating temporal information or not, demonstrate the effectiveness and generalizability of injecting semantic information to guide the BEV encoding process in HD map construction. Notably, using text descriptions as SMG input and employing CLIP's text encoder to extract features improves all baselines across all classes, thresholds, and mAP. However, compared to the simpler MLP-based extractor, the more complex feature extractor does not yield additional gains in most cases, except when using MapTR with GKT as the BEV encoder. A possible reason is that the semantic input space is relatively simple, with only three classes. In this setting, the powerful representation capability of CLIP's encoder may be underutilized, offering limited advantage over lightweight alternatives like MLPs.

Figures 4 and 5 illustrate how SMG enhances online vectorized map construction. For comparison, we present the ground-truth map, the vectorized maps predicted by MapTR and MapTR v2, and the results obtained after incorporating our proposed SMG. Guided by semantic features, the outputs align more closely with the ground-truth. For instance, in the first two rows of Figure 4, SMG enables more accurate capture of map details compared to the original MapTR and MapTR v2, as highlighted by the red circles. In the last row of Fig. 4, the boundaries of pedestrian crossings are also more precisely delineated. In night scenes, as shown in Fig. 5, our method produces clearer map elements than the baseline models. Overall, SMG improves the alignment between predictions and ground-truth, demonstrating its effectiveness in enhancing the quality of BEV feature maps.

4.3. Performance on Argoverse 2

We also evaluate our framework on the Argoverse 2 dataset, which provides 3D vectorized maps with additional height information compared to the nuScenes dataset. As shown in Tab. 3, the proposed SMG consistently improves mAP and AP under all thresholds, as well as AP for most classes,

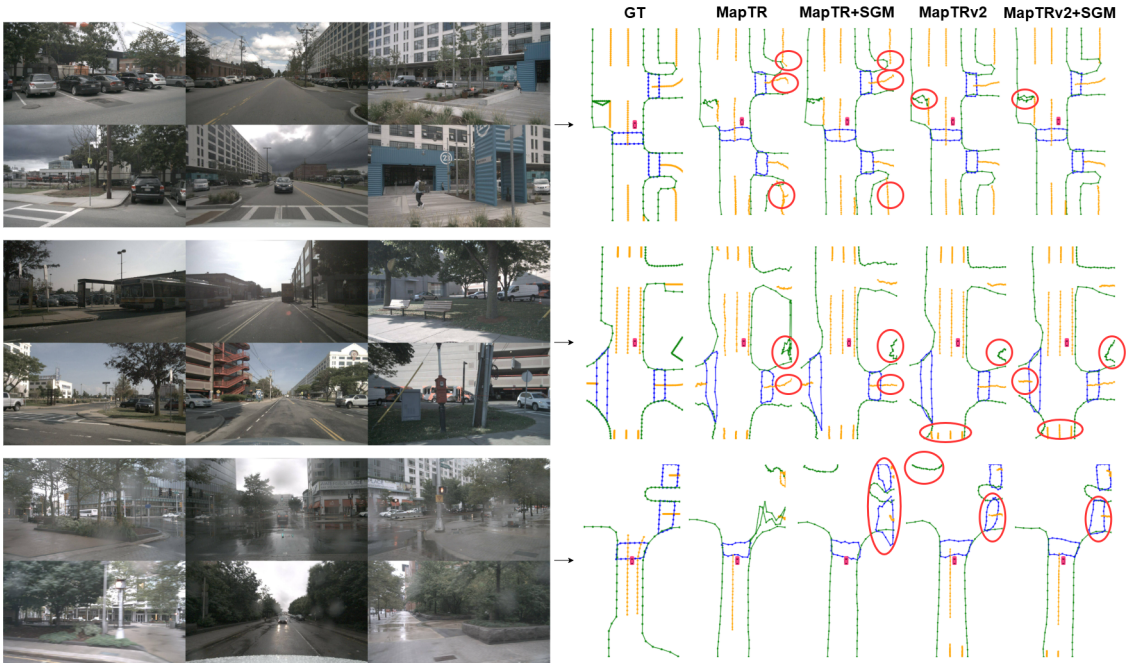

Figure 4. Visual comparison of MapTR and MapTR+SMG, as well as MapTRv2 and MapTRv2+SMG, on the nuScenes validation set for day scenes. The blue, yellow, and green lines represent pedestrian crossings, dividers, and road boundaries, respectively. Our method shows noticeably improved alignment with the ground-truth maps.

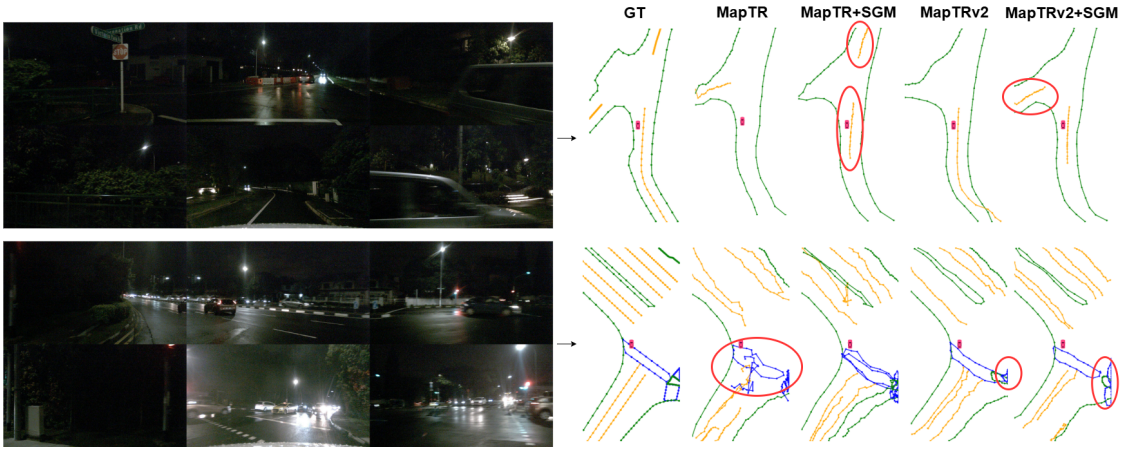

Figure 5. Visual comparison of MapTR and MapTR+SMG, as well as MapTRv2 and MapTRv2+SMG, on the nuScenes validation set for night scenes.

across all models. Specifically, it boosts the baseline performance by 1.88%, 1.60%, and 0.87% for MapTR, MapTR v2, and StreamMapNet, respectively.

5. Ablation Studies

We conducted ablation studies on the nuScenes dataset using MapTR (with GKT as the BEV encoder) as the baseline to evaluate the impact of each component in our proposed model.

5.1. The effectiveness of Cropping Strategy

As introduced in Sec.3.2, there are three methods to crop the corresponding region in the BEV feature: using point positions, using a bounding box for each instance, and using two bounding boxes for the road boundary with one bounding box for the other two map elements—crosswalk and divider. The comparison is summarized in Tab.4. It can be seen that cropping with bounding boxes slightly outperforms point-based cropping in terms of mAP and yields better results

Method	BEV Base Model	BEV Size	SMG Input	SMG Encoder	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	$AP_{0.5m}$	$AP_{1.0m}$	$AP_{1.5m}$	mAP
MapTR	GKT	200×100	-	-	63.16	54.09	60.53	31.85	66.11	78.82	59.26
MapTR+SMG			One-hot	MLP	63.53	55.81	**60.65**	33.06	66.91	80.02	60.00
MapTR+SMG			Text	CLIP-Text	**65.66**	**57.51**	60.26	**34.51**	**68.06**	**80.86**	**61.14**
MapTR V2	BEVPool v2	200×100	-	-	67.66	58.85	64.29	38.16	70.71	81.93	63.60
MapTR V2+SMG			One-hot	MLP	**69.62**	**61.1**	64.88	39.74	**72.55**	**83.3**	**65.20**
MapTR V2+SMG			Text	CLIP-Text	69.14	60.16	**65.19**	**39.94**	71.87	82.68	64.83
StreamMapNet	BEVFormer	50×100	-	-	54.95	**66.51**	55.07	42.76	61.81	71.97	58.84
StreamMapNet+SMG			One-hot	MLP	**56.39**	66.5	56.24	**44.38**	**62.27**	**72.48**	**59.71**
StreamMapNet+SMG			Text	CLIP-Text	55.4	66.09	**56.52**	43.98	61.8	72.23	59.34

Table 3. Performance comparison with baseline methods on Argoverse 2 Validation set

across most classes. This may be because bounding boxes aggregate features over a wider area, providing stronger spatial cues that help the model localize more effectively. Using two bounding boxes to split the road boundary while employing one bounding box for the divider and crosswalk significantly outperforms both the single bounding box approach and the point-based method, improving mAP by 1.26% and 1.39%, respectively.

Cropping Strategy	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	mAP
Point	51.26	46.34	52.69	50.10
Box	51.28	47.00	52.41	50.23
DualBox-bou	**53.23**	**47.99**	**53.26**	**51.49**

Table 4. Ablation study on different cropping methods

5.2. SMG input format and SMG encoder

We use either one-hot class labels or text descriptions as the semantic input to the semantic feature extractor (SFE). One-hot encoding represents each class as a binary vector with a single active position, allowing the model to clearly distinguish between classes without implying any ordinal relationship. To further explore the effectiveness of this input format, we enhance it by incorporating point position information for each map element. This additional input enriches the feature set and provides spatial cues to guide the decoder in predicting point positions. We evaluate both a simple MLP and a more expressive model, PointNet [23], as the feature extractor.

In addition to digital formats, we also experiment with using visual inputs. Specifically, we create three masked images—one per class—to clearly show the positions of each map element and employ UNet [26] as the feature extractor. The results of these experiments are summarized in Tab. 5. As shown, incorporating point position, whether using a simple MLP or a more complex PointNet for feature extraction—results in the two worst performances. A potential reason is that the model already captures spatial structure through BEV supervision with ground-truth labels. Introducing explicit point coordinates may create redundancy, leading the model to overfit specific spatial patterns rather than learning generalizable geometric or semantic features.

SMG Input	SMG Encoder	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	mAP
label	MLP	**53.23**	**47.99**	**53.26**	**51.49**
label + point position	MLP	49.24	43.52	51.39	48.05
label + point position	PointNet	49.73	43.9	52.27	48.64
masked class images	UNet	51.41	45.63	52.89	49.97

Table 5. Ablation study on SMG input format and SMG encoder

5.3. Feature Alignment

In our method, we employ contrastive learning to align the produced BEV features with the ground-truth BEV features. Additionally, we use mean squared error (MSE) loss for alignment. The results, shown in 6, indicate that contrastive loss provides better performance. However, compared to the baseline (MapTR), using MSE for alignment also yields satisfactory results.

Alignment loss	$AP_{div.}$	$AP_{ped.}$	$AP_{bou.}$	mAP
W/o alignment	51.06	43.92	52.55	49.18
MSE	51.39	45.53	**53.35**	50.09
Contrastive loss	**53.23**	**47.99**	53.26	**51.49**

Table 6. Ablation study on different alignment loss functions

6. Conclusion

In this study, we propose a Semantic Map Guidance Module (SMG) for online HD map construction using BEV representations. The module leverages ground-truth labels to guide BEV encoding and employs a contrastive loss to align semantic features with the corresponding BEV features. This enables the SMG to explicitly organize BEV elements based on class labels, enhancing their alignment with ground-truth structures. Extensive experiments on the nuScenes and Argoverse 2 datasets demonstrate that SMG accuracy gain is generalizable across different network architectures and consistently improves the performance of several state-of-the-art baselines. Notably, SMG is only applied during training and introduces no additional computational cost during inference, making it both effective and practical for real-world deployment.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 5

[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. 3

[3] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pages 550–567. Springer, 2022. 1, 2

[4] S Chen, T Cheng, X Wang, W Meng, Q Zhang, and W Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. arxiv 2022. *arXiv preprint arXiv:2206.04584*. 2, 3, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2, 3

[6] 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. 3

[7] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 3

[8] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? *arXiv preprint arXiv:2206.07959*, 2022. 3

[9] Xin He, Haiyun Guo, Kuan Zhu, Bingke Zhu, Xu Zhao, Jianwu Fang, and Jinqiao Wang. Monocular lane detection based on deep learning: A survey. *arXiv preprint arXiv:2411.16316*, 2024. 1

[10] 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. 2

[11] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 3, 6

[12] Haruya Ishikawa, Takumi Iida, Yoshinori Konishi, and Yoshimitsu Aoki. Pct: Perspective cue training framework for multi-camera bev segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13253–13260. IEEE, 2024. 3

[13] 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. 2

[14] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 3, 6

[15] 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, 3, 5, 6

[16] 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. *International Journal of Computer Vision*, 133(3):1352–1374, 2025. 3, 5, 6

[17] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. End-to-end lane shape prediction with transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3694–3702, 2021. 1

[18] 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, 3

[19] 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. 3

[20] Hanspeter A Mallot, Heinrich H Bülthoff, James J Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64(3):177–185, 1991. 3

[21] Chenbin Pan, Burhaneddin Yaman, Senem Velipasalar, and Liu Ren. Clip-bevformer: Enhancing multi-view image-based bev detector with ground truth flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15216–15225, 2024. 1

[22] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. 1, 2, 3

[23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 8

[24] 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. 1, 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 3

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 8

[27] 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

[28] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5135–5142. IEEE, 2020. 1

[29] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 913–922, 2021. 1

[30] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2, 5

[31] 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. 1, 3, 5, 6

[32] Gongjie Zhang, Jiahao Lin, Shuang Wu, Zhipeng Luo, Yang Xue, Shijian Lu, Zuoguan Wang, et al. Online map vectorization for autonomous driving: A rasterization perspective. *Advances in Neural Information Processing Systems*, 36:31865–31877, 2023. 3

[33] 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

[34] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 3