Leveraging SD Map to Assist the OpenLane Topology

Guang Li[1*] Jianwei Ren[2*§] Quanyun Zhou[1†] Anbin Xiong[1†] Kuiyuan Yang[1‡]

[1]XIAOMI EV [2]BUPT

lguang@live.cn, jvren42@gmail.com

{zhouquanyun, xionganbin, yangkuiyuan}@xiaomi.com

Abstract

OpenLane topology understanding aims to perceive various road elements in driving scenes and interpreting their topological relationships based on multi-view images, which has played a crucial role in scalable autonomous driving systems. Considering the basic road topology information is readily accessible in Standard Definition (SD) maps, we explore the efficacy of SD map in assisting the OpenLane topology problem and design a compact transformer-based architecture for SD map encoding and integration. Further, we propose a dynamic positional encoding scheme to improve the decoding performance, which exploits the intermediate lane points to refine the positional encoding for each lane attention layer in the detection head. The more precise location information addresses unpredictable changes in various driving scenes, and delivers more accurate localization results. The proposed method ranked 2^{nd} in the OpenLane-V2 UniScore (OLUS) on the final leaderboard of the OpenLane Topology Challenge 2024.

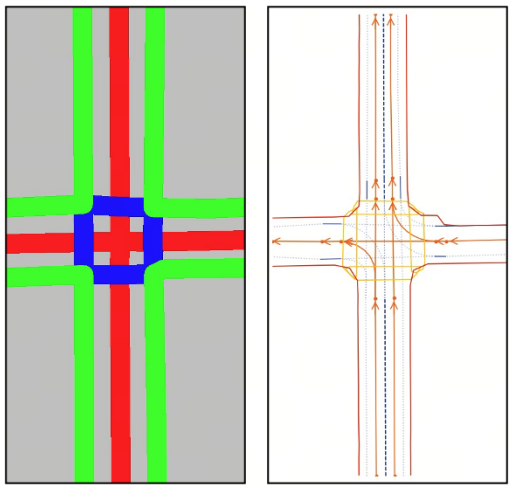

Figure 1. The left image is a sample SD map, and the right one is its corresponding HD map ground-truth. Though the SD map has relatively limited details, it provides a high-level description of the road in a coarse manner, which is important for distant or complex scenes.

1. Introduction

OpenLane Topology [3, 9] involves perceiving road elements such as lanes, traffic signs and traffic lights, and also understanding their relationships from multi-view images. This task is essential for enabling autonomous vehicles to navigate accurately and safely in complex driving environments. Specifically, five sub-tasks are considered in this problem, i.e., lane detection, area detection, traffic element detection, lane-lane topology, and lane-traffic topology prediction.

SD maps is cost-effective and widely available for both human and autonomous driving systems. Considering the basic road geometry, lane information, and connectivity inherent in SD maps, we attempt to explore its complementary information to multi-view images in solving the OpenLane topology problem. As shown in Figure 1, SD map provides basic road layout and geometry information, which is useful in guiding the construction of the High Definition (HD) map with more details. To this end, a SD map encoder is first designed to encode the vectorized elements in SD maps, where the encoding is additionally guided by the Bird Eye's View (BEV) features to provide rich spatial context information. Then the enhanced SD map tokens are further integrated in the lane detection head as an additional modality for better road layout awareness.

Besides, we also propose a dynamic positional encoding scheme [6] for the lane detection head to further improve the performance. Traditionally, position encoding is formulated as learnable tokens and remains fixed for different lane attention layer in the detection head. However, this static dilemma is inferior to provide precise location information and hinder the localization accuracy. We noted

* Equal Contribution † Tech Lead ‡ Corresponding Author

§ This work was completed during an internship at *XIAOMI EV*.

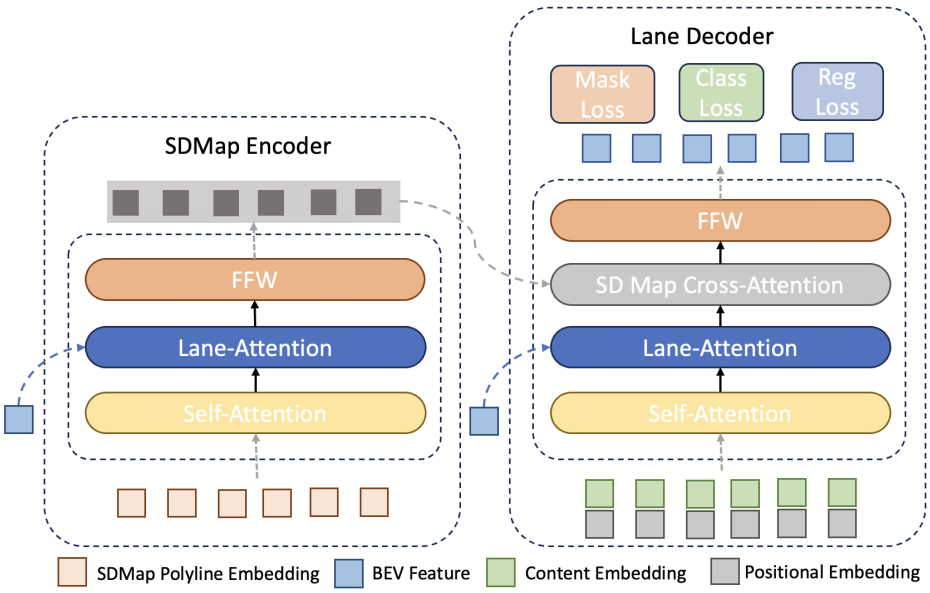

Figure 2. The architecture of our proposed SD map Encoder and Lane Decoder. The SD map Encoder is based on transformer with the cross-attention layer be replaced with the Lane-Attention layer to better preserve the structural information from the BEV feature map. The lane decoder is also built upon transformer, with an additional cross-attention layer to absorb the output of the SD map encoder for road layout information.

lane points from intermediate layers already cater for more explicit and precise location clues. Framed in the layer-by-layer architecture, the predicted lane points from each layer are naturally refined in a coarse-to-fine manner to provide more accurate location information to the underlying driving scene. Besides, lane points are changing in different scenes. Motivated by the two observations, we update the position encoding for each lane attention layer based on the intermediate outputs of reference points, which leads to more precise location awareness and improved detection accuracy.

In the OpenLane Topology Challenge 2024, our proposed architecture attains the 2^{nd} position on the leaderboard with a OLUS score of 63.9. Besides, our method outperforms all other competitors on Top_{ll} with a score of 47.43%.

2. Method

2.1. Model Architecture

Overall, our approach is divided into four consecutive components. Firstly, given the multi-view input images, the BEV features are constructed using BEVFormer [5]. Due to the length limit, illustration of the BEV feature extraction are omitted and please resort to [5] for more details. Secondly, based on the BEV features, we built a SD map encoder to extract SD-Map features with cross-modality interaction. Thirdly, we introduce a novel ensemble approach, which can further enhance the performance. Finally, we decouple the topological relationships from road element detection using a Multi-Layer Perceptron (MLP), which improves the accuracy of topology prediction. The SD map encoder, the lane decoder and the topology prediction are the main parts of our contribution, and we will elaborate on them in the following sections.

2.2. SD Map Encoder

Given an SD map, we aim to extract cross-modality features from the polylines with a modified transformer decoder[8], as illustrated in Figure. 2.

Given the SD map elements represented in polylines, we first encode them into polyline embeddings. Specifically, a fixed number of N points are evenly sampled from each of the M polylines, which are further normalized to the BEV range. The sampled points are then encoded into sinusoidal embeddings as:

$$E(p, 2j) = sin(\frac{p}{T^{\frac{2j}{d}}}),$$

$$E(p, 2j+1) = cos(\frac{p}{T^{\frac{2j}{d}}}),$$

where $p = (x_i, y_i)$ is the sampled point coordinate on the polyline, j is the index of dimension, d denotes the dimension and T is the temperature scale. A one-hot vector with dimension K is used to encode the lane type. Finally, the positional embedding of all the sampled points and the lane type embedding are concatenated, resulting to the polyline

embedding of a SD map with shape of $(N \cdot d + K)$. Before feeding the embeddings into the transformer encoder, we use a linear layer to align them with the model dimension d_m.

To enrich the SD map features with more detailed sensory information, we employ a transformer [8] to interactively fuse BEV features via the cross-attention mechanism. Specifically, we adopt Lane Attention proposed in [4] as an improved implementation of cross-attention layers, which is demonstrated to better preserve structured information compared to original version.

2.3. Lane Decoder

With enhanced SD map features and BEV features, the lane decoder is built to predict the coordinates of the road elements. The transformer decoder with the lane attention implementation is adopted to construct the lane decoder. To gain cross-modality alignment, an additional SD map cross attention layer is stacked after each lane attention layer to incorporate the SD map features.

In the Lane Decoder, the coordinates of the road elements we aim to detect are progressively adjusted and refined through each layer. However, the original positional encoding information of the decoder remains consistently fixed, which affects the accuracy of the detection outputs. To this end, we propose a dynamic positional encoding scheme for each lane attention layer in the lane decoder, and update the positional encoding based on the location information of lane points derived from the output of the preceding decoder layer. The mechanism is illustrated in Figure 3. Specifically, each lane points is encoded into d dimension using sinusoidal positional encoding. For a road element comprising N points, this results in a positional encoding of size $N \times d$. Subsequently, we employ a Linear layer to map this encoding to match with the model's dimensionality d_m. The resulting encoding is exploited as the new positional encoding to complete the lane attention operation in current decoder layer.

Beyond the detection of lane segments, the decoder is also tasked with pedestrian crossing identification and road boundary lines detection. In general, different road element types require different representational complexity. For example, a pedestrian crossing can be simply represented with a pair of points, while road boundaries necessitate a more number of points for accurate representation. As a result, it is inappropriate to condition the three different tasks on the same decoder. In our model training regime, we formulate the road element detection tasks in a multi-task fashion with the shared architectures but separate parameters.

2.4. Topology Prediction

Most existing approaches [4, 11, 12] infer the road topology together with road elements detection tasks in a multi-task framework. However, in a topology prediction task, the number of positive (associated elements) and negative (non-associated elements) samples are highly imbalanced, which makes it is hard to train and fine-tune in a multi-task framework. To better handle the imbalance problem in the topology prediction task, we propose to decouple it from the multi-task framework, and solve it in a subsequent stage based on the detection results from the lane decoders. We conjecture that the geometric relationship of the roads is adequate to derive topological relationships as long as the road detection results are sufficiently accurate.

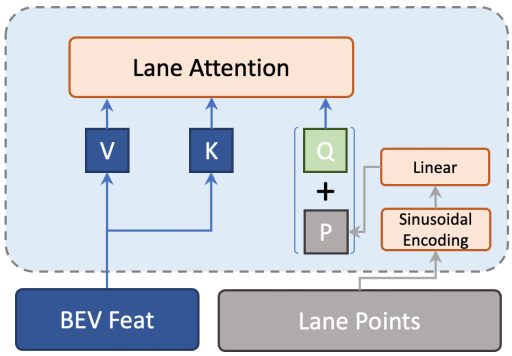

Figure 3. The dynamic positional encoding scheme which updates the positional encoding for each lane attention layer in the lane detection head based on predicted lane points of previous layer.

The multi-layer perceptron model similar to TopoMLP [12] is employed for lane-lane topology and lane-elements topology prediction with different trained parameters. The input to this MLP consists of two parts: the first part is the predictions from the lane decoder which caters for the geometric information between roads. To further compliment the geometric features, we additionally consider the distance between the endpoint of one lane to the starting point of another lane. The closer the starting and ending points of two lanes are, the more likely they are topologically linked.

In terms of the lane-elements topology prediction, we use the bounding box coordinates of traffic elements in the front-view camera and the camera's extrinsic parameters to encode the information of the traffic elements.

2.5. Model Ensemble

An OpenLane-friendly strategy is proposed to enhance the performance by ensembling outputs from distinct detection models. First, we select the results of the best-performing model on the validation set as the base proposals, and rank the others accordingly. Then, the results of the other candidate models are integrated into the base ones via a trust-based voting strategy. Specifically, we rely on the similarity of the upcoming proposals in the candidate models with the base ones and accordingly assign a confidence score to rank

ID	n_{points}	$Task$	w/ SD	w/ DP	DET_l	DET_a	AP_{ped}	AP_{bd}
1	10	L	×	×	30.02	-	-	-
2	10	L	×	✓	30.78	-	-	-
3	10	L	✓	✓	**32.39**	-	-	-
4	2	P	×	×	-	-	32.34	-
5	10	P	×	×	-	-	31.97	-
6	20	B	×	×	-	-	-	24.64
7	30	B	×	×	-	-	-	20.26
8	10/2/20	LPB	×	×	27.48	29.99	32.16	27.82
9	10/2/20	LPB	✓	✓	29.77	**34.26**	**37.43**	**31.09**

Table 1. Ablation study on the OpenLane-V2 [9] validation set with the same backbone of ResNet-50, all the models are trained for 24 epochs with a batch size of 96. L, P and B denotes the task of lane segment, pedestrian crossing detection and road boundary detection, respectively. w/SD means using SD map. w/DP means using Dynamic Positional Encoding module.

the proposals. In the affirmative case [1], the base model accepts all new proposals (the proposals that are dis-similar to any of base proposals) from the candidate models, while the confidence of these proposals is reduced by a decay factor. Conversely, in the consensus case [1], proposals from the candidate and base models are quite similar. Given the original confidence score p_b of the base proposal, its new confidence score after ensembling will be enhanced to $(1 + p_c^2) \times p_b$, where p_c is the confidence of the candidate proposal. In this situation, we discard the candidate proposals, keep those base proposals, and increase their confidence by multiplying with a scaler larger than 1. With this strategy, we can effectively reduce the number of false negatives while boosting the confidence of true positives.

3. Experiments

3.1. Implementation Details

Following LaneSegNet [4], we adopt the same data processing pipeline and BEV feature dimensions. Each backbone undergoes an initial training phase of 70 epochs on the three tasks jointly, followed by 10 epochs dedicated to specific tasks. Our model is trained across 24 GPUs, with a batch size of 4 per GPU. The AdamW optimizer starts with a learning rate of 1e-3, which decays by a factor of 1/5 at milestones 0.7 and 0.9. For finetuning, the learning rate is adjusted to 3e-4. For the topology prediction task, we use a fixed learning rate of 1e-4, and training the model on distortional ground truth lane and traffic elements for 10 epochs. In addition, increase the scale of the input images can significantly enhance the expression capability of BEV features, so we enlarge the default scale of 0.33 to 0.5 for our submitted models.

When handling road boundary and crosswalks, to resolve ambiguity, we standardized the traversal direction of the elements. Consequently, during training, we do not need to consider the diversity of representations.

ID	$Backbone$	$Pretrain$	$Data$	DET_l	DET_a
1	Swin-L	ImageNet-1K	trainval	42.34	45.30
2	VoVNetV2-99	ADE20K	trainval	38.82	42.72
3	InternImage-b	ADE20K	trainval	45.73	-
4	Ensemble	-	-	**49.97**	**49.80**

Table 2. Performance of different backbones evaluated on the OpenLane-V2 test set.

3.2. Experimental Results

SD-Map and Dynamic Positional Encoding. Comparing results in Row 1-3 and Row 8-9 of Table 1, using SD Map and Dynamic Positional Encoding consistently enhances performance across all three tasks, demonstrating the effectiveness of these two proposed strategies for better OpenLane Topology Understanding.

Number of points. Comparing the results in Row 4-7 of Table 1, employing a more concise representation with fewer points is beneficial with improved performances in both $ped\_crossing$ and $road\_boundary$ tasks.

Joint training. Comparing the results of Row 6&8 in Table 1, we find that using the multi-task joint training significantly improves the boundary detection performance. Additionally, the model trained on three detection tasks serves as a good pre-trained model for training downstream tasks separately.

Model Ensemble. To further enhance the model's capability and effectiveness, we utilized larger backbones as feature extractors. For our competition submissions, we employed models such as Swin-L [7], InternImage-b [10] and VoVNetV2-99 [2]. And we trained them using the same strategy as above. Table 2 examines the effects of different backbones on the performances. It can be observed that despite the smaller size of the Internimage-b [10] model, its performance is better due to the use of appropriate pretraining. Therefore, in the subsequent model ensembles, we used its results as the base and gradually integrated the results of Swin-L and VoVNetV2-99 via the proposed trust-based voting scheme. As observed, the ensembled model achieves the best performance, validating the effectiveness of the proposed ensemble strategy.

4. ACKNOWLEDGMENTS

We would like to express our sincere appreciation to all the interns, who have already completed their internships with the company, including Jianghai Shuai and Zexuan Cheng, for their significant contributions to this project. Their enthusiasm and tireless work have greatly enhanced our research and development endeavors.

References

[1] Ángela Casado-García and Jónathan Heras. Ensemble methods for object detection. In *ECAI 2020*, pages 2688–2695. IOS Press, 2020. 4

[2] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 4

[3] Tianyu Li, Li Chen, Xiangwei Geng, Huijie Wang, Yang Li, Zhenbo Liu, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, et al. Topology reasoning for driving scenes. *arXiv preprint arXiv:2304.05277*, 2023. 1

[4] 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. 3, 4

[5] 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. 2

[6] 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. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[9] Huijie Wang, Zhenbo Liu, Yang Li, Tianyu Li, Li Chen, Chonghao Sima, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, et al. Road genome: A topology reasoning benchmark for scene understanding in autonomous driving. *arXiv preprint arXiv:2304.10440*, 2023. 1, 4

[10] W Wang, J Dai, Z Chen, Z Huang, Z Li, X Zhu, X Hu, T Lu, L Lu, H Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. arxiv. *arXiv preprint arXiv:2211.05778*, 2022. 4

[11] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. TopoMLP: An simple yet strong pipeline for driving topology reasoning. *arXiv preprint arXiv:2310.06753*, 2023. 3

[12] Dongming Wu, Fan Jia, Jiahao Chang, Zhuoling Li, Jianjian Sun, Chunrui Han, Shuailin Li, Yingfei Liu, Zheng Ge, and Tiancai Wang. The 1st-place solution for cvpr 2023 openlane topology in autonomous driving challenge. *arXiv preprint arXiv:2306.09590*, 2023. 3