Scene Perception and Reasoning with SD Map in Aligned Feature Space for Mapless Driving

Rui Ding[1*] Xiuye Rui[2] Chenchen Wang[2] Bowen Ma[1] An He[1] Jianjun Wang[1]
[1]Bosch Automotive Products (Suzhou) Co., Ltd.
[2]Bosch China Innovation & Software Development Campus
6221905018@stu.jiangnan.edu.cn, jianjun.wang2@cn.bosch.com

Abstract

In this report, we present our solution for the CVPR 2024 Autonomous Driving Challenge: Mapless Driving. Image information is highly sensitive to environmental changes, which may limit performances of perception models utilizing images as the sole input source. Therefore, we propose a multimodal, multitask perception model that leverages a standard-definition (SD) map, aiming at significantly enhancing the detection performance of lane-related elements. Upon detecting lanes and traffic elements, the model aligns these intermediate results within the feature space. This alignment enables the model to accurately deduce the topological relationships between the lanes and the traffic elements. An OULS score of 0.5164 is achieved in our final submission.

1. Introduction

The vehicle's ability to perceive the environment is crucial in autonomous driving, as it directly determines the accuracy, robustness, and reliability of inputs to downstream modules. In static perception tasks, the system needs to detect traffic elements, e.g., traffic lights and signs, and lane-related elements, e.g., lane centerlines, dividers, and curbs. Currently, models like TopoNet[3] and LaneSegNet[4] effectively detect the lane and traffic elements using multi-view images captured by onboard cameras of the ego vehicle. the performances of these models may be limited by the narrow field of view in images which lacks long-distance lane information. Additionally, the models' detection capabilities could be deteriorated by other factors including the complex and dynamic road environment, numerous traffic participants, changes in weather or lighting, and onboard cameras being physical obscured. Fortunately, standard-definition (SD) maps can provide static prior information to aid the models in understanding road structures and improving detection performances. In contrast to high-precision (HD) maps, SD maps are cost-effective, update less frequently, and cover a broader area, making them suitable for large-scale deployment. Therefore, integrating SD maps with perception models to enhance perception performance is the first research focus of this report.

In addition to perception, it is vital for models to reason topological relationships among detected elements. For example, models should clearly infer connections between detected traffic lights and their corresponding lanes, as each traffic light controls one or more specific lanes at complex intersections. However, relying solely on the detected lanes and traffic elements, reasoning such relationships was proved ineffective[9]. This ineffectiveness may stem from the mismatch of the frame of references of the two detection tasks, i.e. the lane detection tasks typically outputs the 3D coordinates of lane lines in the ego vehicle's frame of reference, whereas the traffic element detection tasks provide the 2D coordinates of traffic elements in the image's frame of reference. To address this challenge, reasoning topological relationships using aligned features from both detection tasks is another research emphasis of this work.

In this report, we propose a multimodal, multitask perception model that integrates SD maps to elevate the performance of lane-related element detection. Our model aligns lane detection and traffic element detection results in the feature space to reason their relationships effectively. Our main contributions are summarized as follows:

- We propose a novel multimodal, multitask perception model that leverages SD maps to enhance lane-related element detection.
- We introduce a novel alignment mechanism to reason topological relationships between lanes and traffic elements.
- We evaluate our model on the OpenLaneV2 dataset and achieve an OULS score of 0.5164.

*This work is done during the internship at Bosch.

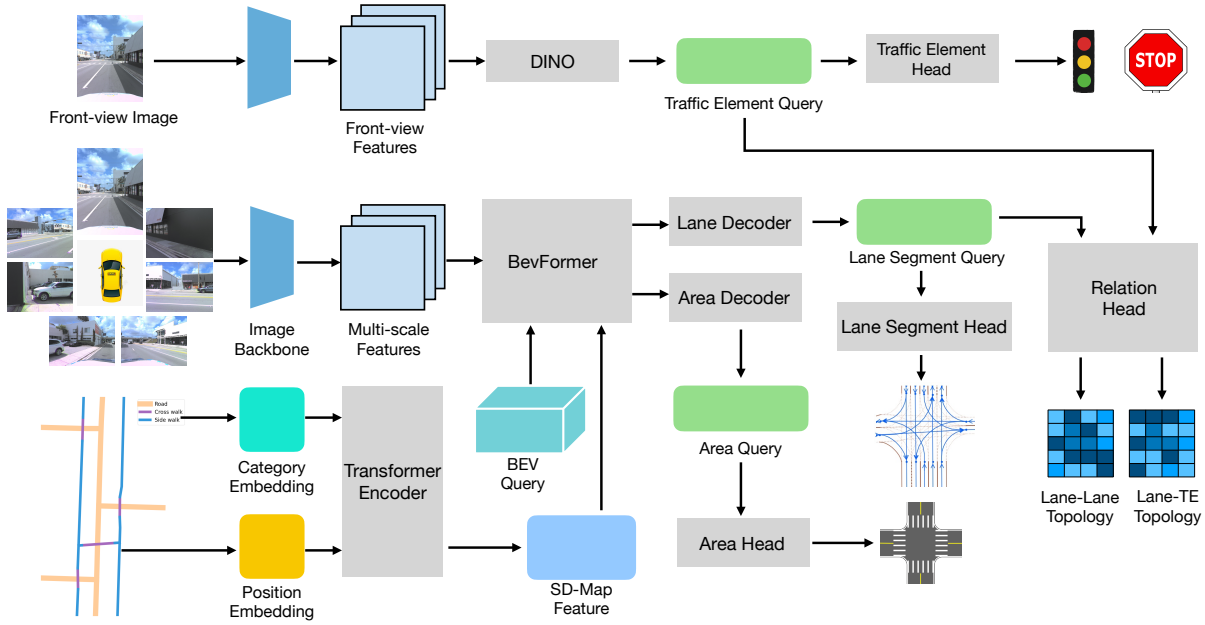

Figure 1. The framework of our model.

2. Methodology

The main framework of our model is shown in Figure 1. In this section, we will introduce the details of our model.

2.1. SD Map Feature Extraction

SD maps are comprised of a collection of polylines, with each polyline representing a specific road element, such as roads, sidewalks, or crosswalks. Given that polylines vary in length, direct processing by the model is infeasible. Thus, we segment each polyline into a set of equidistant point sequences, with each point sequence $p = \{x_i, y_i\}_{i=1}^{N}$ having a fixed length L. We use the sine-cosine positional encoding to map each point sequence p to a fixed-dimensional embedding v, where $v \in \mathbb{R}^d$. Additionally, to distinguish different categories of polylines, a category-specific embedding $c \in \mathbb{R}^{|c|}$ is appended to each polyline. We concatenate the positional embedding and category embedding to obtain the SD map features M, where $M \in \mathbb{R}^{n \times (d+|c|)}$. To further capture the global information of the SD map, we use an MLP to process M and a following standard 6-layer Transformer encoder to extract features, obtaining the meticulously processed features M_{map}, where $M_{map} \in \mathbb{R}^{n \times d}$.

2.2. Map-Aware BEV Feature Extraction

To project multi-view images into a unified Bird's Eye View (BEV) space, we adopt the methodology proposed by BEVFormer[5]. We commence by using Swin Transformer[7] as the backbone to extract image features, denoted as F_{img}. Subsequently, we initialize the BEV query Q_{bev} and make it interact with F_{img} through a spatial cross-attention mechanism to derive the BEV features from the multi-view images. Building on this framework, we incorporate a map cross-attention layer following the spatial cross-attention. This additional layer enables the BEV query to interact with the SD map features M_{map}, thereby generating a map-aware BEV feature that enhances the contextual understanding of the scene.

2.3. Lane Detection

Based on the map-aware BEV feature obtained in Section 2.2, we employ the lane detection methodology proposed by LaneSegNet[4]. Initially, we prepare B lane queries Q_{lane}, where $Q_{lane} \in \mathbb{R}^{B \times d}$. Utilizing the lane attention mechanism from LaneSegNet, these queries interact with the map-aware BEV features. This interaction enriches the queries with prior information on road structures. Subsequently, we apply classification and regression heads to accurately determine the lane categories and positions.

2.4. Traffic Element Detection

For traffic element detection, the objective is to compute the 2D coordinates within the image coordinate system, focusing primarily on elements visible in the front view. We apply the Transformer-based object detection algorithm DINO[2] for this purpose. The resulting traffic element query, $Q_{te} \in \mathbb{R}^{K \times d}$, contains detailed information of both the category and the 2D position of each traffic element.

Method	DET_l	DET_te	DET_a	TOP_ll	TOP_lt	OLUS
LaneSegNet	0.3084	0.3609	0.2097	0.2564	0.2120	0.3692
LaneSegNet+SD	0.3798	0.3828	0.2676	0.3227	0.2398	0.4176
LaneSegNet+FSA	0.3103	0.3626	0.2061	0.2582	0.2688	0.3811

Table 1. Results on OpenLaneV2 validation dataset.

Method	Backbone	DET_l	DET_te	DET_a	TOP_ll	TOP_lt	OLUS
LaneSegNet	ResNet50	0.2826	0.4578	0.2177	0.2321	0.2432	0.3866
Ours	ResNet50	0.3296	0.6491	0.2679	0.2870	0.3099	0.4678
Ours	SwinL	0.3874	0.6848	0.2964	0.3172	0.3485	0.5044
Ours*	SwinL	0.3874	0.7410	0.2964	0.3172	0.3529	0.5164

Table 2. Results on OpenLaneV2 test dataset. * indicates the result of DET_te is obtained by yolov8.

We then apply classification and regression heads to accurately determine these attributes. Compared to methods like YOLO, the comprehensive data embedded within Q_{te} offers significant advantages in subsequent tasks of reasoning about relationships between elements. Although Transformer-based approaches generally show less efficacy in detecting small objects compared to YOLO[1], the enhanced capability for relational reasoning justifies our preference for this method.

2.5. Area Detection

Leveraging the map-aware BEV features detailed in Section 2.2, we implement the area detection methodology introduced by MapTR[6]. Initially, we establish R region queries, Q_{area}, where $Q_{area} \in \mathbb{R}^{R\times d}$. Utilizing the map decoder from MapTR, we extract area-specific features from the map-aware BEV features. Subsequent application of classification and regression heads allows us to accurately determine both the category and position of each area.

2.6. Lane-Lane Relation Reasoning

Given that all lane queries share the same feature space, we can directly utilize the intermediate lane detection results Q_{lane} for reasoning lane-lane relationships. To achieve this, we concatenate the lane queries pairwise to form the lane pair query, Q_{lane_lane}, where $Q_{lane\_lane} \in \mathbb{R}^{B\times B\times 2d}$. An MLP is employed to transform Q_{lane_lane} into a scalar matrixT_{lane_lane}, where $T_{lane\_lane} \in \mathbb{R}^{B\times B}$. This matrix effectively encapsulates the relationships between pairs of lanes.

2.7. Lane-Traffic Element Relationship Reasoning

Unlike the relationship reasoning among lanes, the lane query Q_{lane} and the traffic element query Q_{te} exist in distinct feature spaces. This separation results in ineffective outcomes when attempting direct interaction[9]. To handle this, we first employ the regression head mentioned in Section2.3 to determine the lane positions and project these positions into the image coordinate system, thereby obtaining the 2D coordinates of lanes. These coordinates are then transformed into a fixed-dimensional embedding V_{lane} through sine-cosine positional encoding, where $V_{lane} \in \mathbb{R}^d$. Subsequently, V_{lane} is processed through an MLP and added with the original lane query to generate an enhanced lane query Q^*_{lane}. Following a similar approach to lane-lane reasoning, we concatenate Q^*_{lane} and Q_{te} to form the lane-traffic element pair query Q_{lane_te}, where $Q_{lane\_te} \in \mathbb{R}^{B\times K\times 2d}$. An MLP is then used to transform Q_{lane_te} into a scalar matrix T_{lane_te}, which characterizes the relationships between lanes and traffic elements.

3. Expriment

3.1. Dataset

The OpenLaneV2 is the first large-scale dataset for research on topological relationship reasoning tasks in autonomous driving scenarios. This dataset contains a large amount annotated data of lane, traffic element, lane-lane relationship, and lane-traffic element relationship, as well as corresponding multi-view images and SD map. The OpenLaneV2 dataset used in this competition contains 22477 training samples, 4806 validation samples, and 4816 test samples. Each sample contains 7 surround-view images and the corresponding SD map around the location. The evaluation metric uses the OpenLane-V2 UniScore (OLUS) metric, which comprehensively considers the performance of five tasks: lane detection, traffic element detection, area detection, lane-lane relationship reasoning, and lane-traffic element relationship reasoning, specifically[8]:

$$\text{OLUS} = \frac{1}{5}(DET_l + DET_{te} + DET_a + f(TOP_{ll}) + f(TOP_{lt})) \tag{1}$$

where f represents the square root function.

3.2. Implementation Details

To maximize training efficiency, we standardize all input images to a resolution of 512x714 pixels. We utilize ResNet50 and Swin Transformer as our image backbones to extract multi-scale image features effectively. For optimization, we employ the AdamW optimizer with an initial learning rate of 2e-4, which is adjusted throughout the training period using the cosine annealing strategy. The configurations for the model include 200 lane queries (Q_{lane}), 100 traffic element queries (Q_{te}), and 900 region queries (Q_{area}). The model is trained on six NVIDIA A100 GPUs, maintaining a batch size of one to ensure optimal processing and memory management.

3.3. Results

To verify the effectiveness of our proposed method, we conduct experiments on the OpenLaneV2 validation dataset. The results are detailed in Table 1. In the table, 'LaneSegNet+SD' denotes the integration of SD map information with LaneSegNet, while 'LaneSegNet+FSA' signifies the implementation of feature space alignment during the reasoning of relationships between lanes and traffic elements. The experimental outcomes substantiate the efficacy of our proposed module.

Following these results, additional tests were carried out on the OpenLaneV2 test dataset, with the results presented in Table 2.

4. Conclusion

In this report, we propose an effective method to integrate SD maps with perception models and a feature space alignment module to enhance the performance in lane-related element detection and relationship reasoning tasks. Finally, we achieve 0.5164 on the test set.

Acknowledgment

This work is financially supported by Robert Bosch under Advanced Driving Assist Solution WIN Project.

References

[1] 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*, page 213–229, 2020. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[3] Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo, and Hongyang Li. Graph-based 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. In *ICLR*, 2024. 1, 2

[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. *arXiv preprint arXiv:2203.17270*, 2022. 2

[6] 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. In *International Conference on Learning Representations*, 2023. 2

[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. 2

[8] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. In *NeurIPS*, 2023. 3

[9] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: An simple yet strong pipeline for driving topology reasoning. *ICLR*, 2024. 1, 2