LGmap: Local-to-Global Mapping Network for Online Long-Range Vectorized HD Map Construction

Kuang Wu * Sulei Nian * Can Shen Chuan Yang † Zhanbin Li ‡
Langge Technology
{Kuang.Wu, Sulei.Nian, Can.Shen1, Chuan.Yang3, Zhanbin.Li}@geely.com

Abstract

This report introduces the first-place winning solution for the Autonomous Grand Challenge 2024 - Mapless Driving [1]. In this report, we introduce a novel online mapping pipeline LGmap, which adept at long-range temporal model. Firstly, we propose symmetric view transformation(SVT), a hybrid view transformation module. Our approach overcomes the limitations of forward sparse feature representation and utilizing depth perception and SD prior information. Secondly, we propose hierarchical temporal fusion(HTF) module. It employs temporal information from local to global, which empowers the construction of long-range HD map with high stability. Lastly, we propose a novel ped-crossing resampling. The simplified ped crossing representation accelerates the instance attention based decoder convergence performance. Our method achieves 0.66 UniScore in the Mapless Driving OpenLaneV2 test set.

1. Introduction

The High-Definition (HD) map is designed for high-precision autonomous driving. It contains instance-level vectorized representation such as pedestrian crossing, lane divider, road boundaries, etc. The rich semantic information of road topology and traffic rules is important for the navigation of autonomous driving. The Mapless Driving Track [2] aims to dynamically construct a local HD map from the images of the surrounding camera on board and the SD map. In this work, we present a multi-stage framework, which decouples the 2D / 3D elements detection and topology prediction tasks.

Our method focuses mainly on three aspects to handle the competition.

1. Fusion from close to distant. We propose an innovative approach that incorporate both forward projection and backward projection strategies together with SD-map fusion and depth supervision.
2. Fusion from local to global. We present an novel online mapping pipeline adept at both short-range and long-range, which integrates both streaming strategy and stacking strategy.
3. Ped crossing resampling. We simplify the ped crossing to 4 corners, and then uniformly sample 6 points on each edge.

2. Method

This section introduces the details of our method. We first introduce the main pipeline of the LGmap architecture, as shown in Fig. 1. Then the area components and lane segment components are presented. Furthermore, we introduce the traffic elements. Finally, we describe the attention-based heads for topology reasoning.

2.1. Pipeline

2.1.1 Encoder

There are mainly two types of view transformation, forward projection and backward projection. Lift-Splat-Shoot (LSS)[4] takes advantage of the depth distribution to model the uncertainty of each pixel's depth. But the drawbacks of forward projection is discrete and sparse BEV representation. BEVFormer [5] projects 3D points back onto 2D images. As a backward projection, one limitation of BEVFormer is false correlation between 3D and 2D space due to occlusion. To address these issues, we introduce a symmetric view transformation. The depth-map of each camera is generated from synchronized lidar point cloud. The LSS utilize depth supervision only at the training phase. Given the SD map of the scene, we evenly sample along each of the polylines for a fixed number of points. With sinusoidal embedding, BEVFormer apply cross-attention between the SD map feature representation with features from vision inputs on each encoder layer. In order to fuse BEV representations, we use the channel-attention-based fusion module.

*Equal Contribution †Tech Lead ‡Corresponding Authors

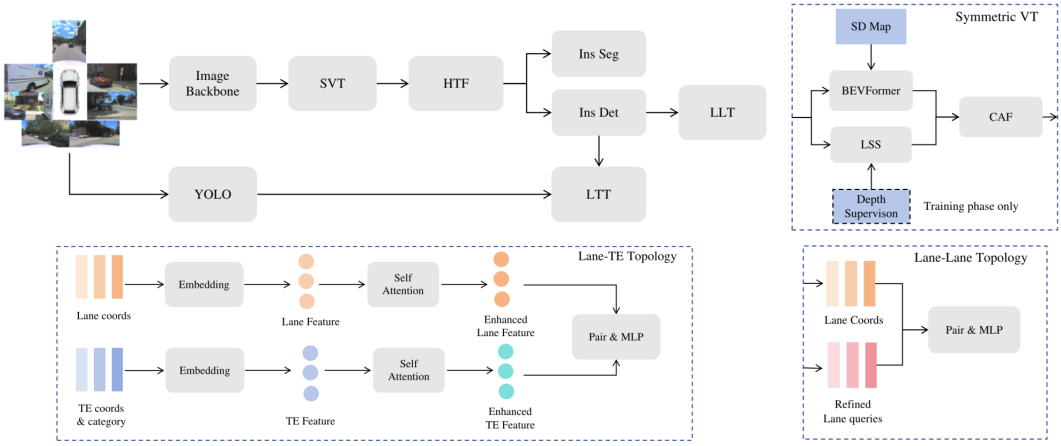

Figure 1. The overall model architecture of LGmap. The entire model is consists of mainly six components: a image backbone equipped with SVT(Symmetric View Transformation), a hierarchical temporal fusion(HTF) module, a unified instance detection and segmentation predictor, a traffic elements detector(YOLO [3]), a Lane-Lane Topology(LLT) and a Lane-TE Topology(LTT).

2.1.2 Decoder

In order to handle different map elements with distinct shape priors, we extend the instance-wise detection decoder with additional segmentation tasks. The unified transformer-based decoder for instance detection and segmentation benefits from both pixel-level classification task and region-level regression task. Additional segmentation branches accelerate the convergence performance of the instance-wise feature embedding.

2.1.3 Temporal fusion

The streaming strategy facilitates longer temporal association as the propagated hidden states encode all historical information. But a temporal fuser such as convGRU [6] may still face the problem of forgetting. The stacking strategy may integrate features from specific previous frames, offering flexibility in fusion of long-range information. The computational cost is linearly related to the number of fused frames. We propose a novel hierarchical temporal fusion (HTF). The hierarchical temporal fusion fully leverage local fusion capability of streaming strategy and long-range fusion capability of stacking strategy. And it minimizes memory and latency costs compared to the stacking strategy. Here we present two variants of HTF, streaming-streaming strategy and streaming-stacking strategy, as shown in Fig. 2. For streaming-stacking strategy, we random select N frames from the latest M previous frames for the stacking mode layer during the training phase. And select N frames by a certain distance strides during testing phase.

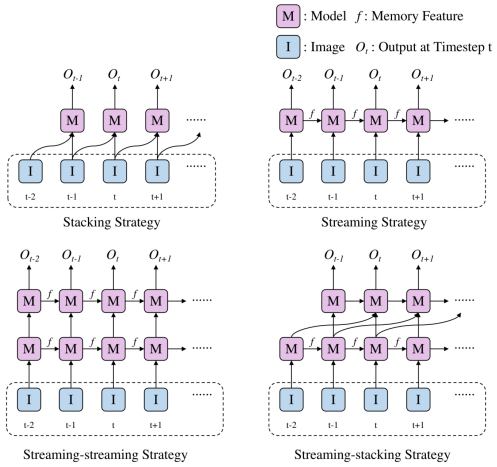

Figure 2. Stacking strategy and streaming strategy are same as StreamMapNet's [7] summary. In order to demonstrate the effectiveness of long-range stacking for streaming-stacking strategy in figure, the stacking previous frame interval parameter is set to 2. Stacking strategy only fuses one previous frame in this figure, and actually it may fuse more than one frame.

2.1.4 Loss functions

Firstly, we adopt classification loss, point2point loss and edge direction loss same as MapTR [8]. Secondly, we adopt image segmentation auxiliary dense prediction loss and depth prediction loss same as MapTRv2 [9]. Thirdly, we adopt BEV instance segmentation loss. Lastly, we adopt geometric 3D loss. Unlike the geometric loss of GeMap [10], which ignores the Z-axis, we extend the euclidean loss dimension from 2d to 3d.

Figure 3. The ped crossing form of MapTR, MachMap and LGmap.

2.2. Area

Inspired by Machmap [11], we simplified the ped crossing by four corners. Then we unified the four corners into the MapTR form of N points. The main difference is that MapTR uses 20 evenly sampled points, MachMap uses 4 points, and we use 6 points evenly sampled on each edge, as shown in Fig. 3.

Our ped-crossing representation keeps four corners as key points, which are essential shape priors. What's more, the permutations of ped crossing are simpler than MapTR. Compared to MapTR's 40 equivalent permutations of one 20 points polygon, LGmap only requires 8. Instead of point-wise permutations, we only use corner-wise permutations. Lastly, preserving corners is beneficial for instance query embedding.

2.3. Lane segments

Based on the centerline output of the regression branch, an offset branch is introduced to predict the offset to left and right lane boundaries, and two classification branches are introduced to predict the attribute of lane boundary, with reference to LaneSegNet [2].

2.4. Traffic elements

We utilize YOLOv8 as a base 2D detector, and we utilize YOLOv9 [3] additionally for model ensemble. Based on the OpenlaneV2 dataset, we propose a series of data augmentation excluding HSV and horizontal flipping, since these tricks may lead to confusion of traffic lights and the direction of traffic signs. The distribution of categories in dataset is highly imbalanced, some categories differing by an order of magnitude. Moreover, pseudo-labels, which are generated on the test set, improve the results. We adopt test-time augmentation(TTA) with the scale range between 0.7-1.4 to improve both the recall of small-objects and large-objects.

2.5. Lane-Lane topology

We use the TopoMLP method [12]. Firstly, we pass the centerline coordinates to MLP and add them to the refined query features. Finally, we apply MLP to perform topology classification.

2.6. Lane-Traffic topology

We use the coordinates of centerlines, and coordinates, categories from traffic element bboxes. We train topology model using ground truth data of lane segments and traffic elements, since the feature embedding are not used. By decoupling with the upstream detection model, the training and prediction process of topology becomes more convenient. Due to the complexity of intersections, we use self-attention to facilitate information exchange among elements and obtain relative relationships.

BEVFormer	LSS	Data-aug	mAP%
✓			40.36
	✓		32.57
✓	✓		40.89
✓	✓	✓	43.75

Table 1. Ablation study of SVT on the Openlanev2 val set.

3. Experiments

3.1. Implementation details

We build our system based on MapTRv2 codebase [9]. Training setup. We adopt two data augmentation methods, image data augmentation and BEV data augmentation, e.g. random rotating, scaling, cropping and flipping. For ablation study, we use the ResNet50 [13] pretrained on ImageNet dataset. And we use ViT-L [14] as the scaling up image backbone. We pretrain the ViT model on nuScenes dataset with vectorized map construction task. For training large-scale models, we use a batch size of 16 on 16 A800 GPUs, AdamW [15] optimizer with a learning rate of 6e-4. The layer-wise learning rate decay is 0.9. Partial freeze block number of ViT is 3. The resolution of input images are 1536×1536. And the image features from the backbone are downsampled with a stride of 16. The depth net predicts depth from 1m to 56m. The BEV feature-map resolution is 100×200. We train the model by two stages. Single-frame mode for 48 epochs and streaming-stacking mode for 36 epochs. During the temporal fusion mode, we change the partial freeze block number of ViT to 12. And turn off both image and BEV data augmentations.

3.2. Ablation Study

3.2.1 SymmetricVT

We examine the efficacy of SVT component through ablation studies, utilizing the OpenlaneV2 dataset [1]. Starting with BEVFormer [5] and LSS [4] as baseline, the best score is 40.36% on the validation set, as shown in Table 1. Compared to the best baseline, the integration of BEVFormer and LSS increase 0.5% mAP. After adding image data-augmentation and BEV data-augmentation, the model performance has improved to 43.75%.

Temporal fusion strategy	mAP%
None	52.93
Streaming	56.61
Streaming-streaming	53.49
Streaming-stacking	57.13

Table 2. Ablation study of HTF on the Openlanev2 val set.

Method	DET-a%
ins-pt attention	33.6
ins attention	34.05
ins attention + ped crossing resampling	35.42

Table 3. Ablation study of ped crossing resampling on the Openlanev2 val set. The ins-pt attention is short for hierarchical attention used in MapTR [8] model.

3.2.2 Temporal fusion

We build a single-frame baseline model by training 72 epochs with ResNet50 checkpoints. And then all experiments finetune the baseline model by 12 epochs. We use single-frame mode to finetune the baseline model, model can reach a score of 52.93% mAP, as shown in Table 2. For the streaming strategy, we use one convGRU [6] as dense fusion encoder. It has a performance improvement of 3.7%. And for the streaming-streaming strategy, two layers of convGRU are used instead of one. Unfortunately, the performance increase only 0.56% compared to single-frame. For the streaming-stacking mode, we select 4 frames out of latest 10 frames for the layer of stacking mode during training phase, and a certain distance strides of 5, 10, 15, 20 meters during testing phase. the performance reaches 57.13% mAP.

3.2.3 Ped crossing resampling

We use the hierarchical attention based decoder same as MapTR as baseline. The model performance reaches score of 33.6% DET-a, as shown in Table 3. Then we change the decoder to instance attention. The model performance increase 0.45%. Finally, we utilize ped crossing resampling to improve the performance to 35.42%

3.2.4 Traffic elements

We utilize COCO pretrained model and finetune for 40 epochs as our 2D detector baseline. The dataset is resampled by a ratio from 5 to 20 times. The entire model is optimized by AdamW with a learning rate of 0.04 and resolution of 1568×2048. And then we generate pseudo labels by threshold of 0.3. YOLOv8-x with data augmetation can reach a score of 79.42% on DET-l, as shown in Table 4. Applying resampling has a performance improvement of 0.64%. TTA further improves 1.0% score. We utilize pseudo label to improve the performance to 81.81%. Finally, the model ensemble of YOLOv8 and YOLOv9 [3] improve the performance to 82.4%.

Method	DET-t%
YOLOv8+data-aug	79.42
+Resampling	80.06
+TTA	81.07
+Pseudo label learning	81.81
+YOLOv9 ensemble	82.40

Table 4. Ablation study of traffic elements on the Openlanev2 test set.

Model A	Model B	Model C	DET-l%	TOP-ll%	TOP-lt%
✓			48.67	40.04	48.27
	✓		46.5	36.82	47.36
		✓	42.33	35.01	44.53
✓	✓		49.8	43.0	50.97
✓	✓	✓	50.74	46.32	53.59

Table 5. Ablation study of lane segments model ensemble on the Openlanev2 test set.

3.2.5 Lane segments

We train three versions of models, using different backbones (ViT [14], InternImage-XL [16]) with different input image resolution scales (0.5, 0.75, 1). During the ensemble process, we utilize an ensemble strategy that incorporating predictions with low similarity. Initially, the models are sorted by their evaluation scores, the best model is the base model, and the other two models are subsequently integrated as proposal models. From the Table 5, it can be seen that the more models ensembled, the more remarkable performance improved.

4. Conclusion

In this work, we rethink the pipeline of 2D / 3D elements detection and topology reasoning of mapless driving. Firstly, we employ a symmetric view transformation(SVT) to combine forward projection and backward projection to form complementary advantages. Secondly, we introduce the hierarchical temporal fusion(HTF) to integrate temporal features from local-to-global stably. Moreover, we improve ped crossing representation by a novel resampling method. Finally, LGmap is the 1st-place solution on the Mapless Driving track, which achieves 0.66 UniScore.

References

[1] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[2] HU Xuyang, GAO Shangbing, WANG Changchun, Hu Liwei, and Li Shaofan. Lanesegnet: an efficient lane line detection method. *Nanjing Xinxi Gongcheng Daxue Xuebao*, 14(5):551–558, 2022. 1, 3

[3] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024. 2, 3, 4

[4] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 1, 3

[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. 1, 3

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2, 4

[7] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024. 2

[8] 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. 2, 4

[9] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*, 2023. 2, 3

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

[11] Limeng Qiao, Yongchao Zheng, Peng Zhang, Wenjie Ding, Xi Qiu, Xing Wei, and Chi Zhang. Machmap: End-to-end vectorized solution for compact hd-map construction. *arXiv preprint arXiv:2306.10301*, 2023. 3

[12] 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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 3, 4

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[16] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 4