Dynamic Graph Topology: 5th Solution for CVPR 2023 Autonomous Driving Challenge: OpenLane Topology

Jiawei Zhao[1] Xuede Li[1] Junfeng Luo[1]
[1]Meituan Vision AI Department
{zhaojiawei12,lixuede,luojunfeng}@meituan.com

Abstract

In this report, we introduce the technical details of our solution for CVPR 2023 Autonomous Driving Challenge: OpenLane Topology. For topology relation modeling in road scenes, it is very challenging to build relations between lanes and traffic elements in an end2end manner. To address this issue, we propose a strong baseline and an efficient dynamic topology modeling head with dynamic graph correlations to improve the performance of the released baseline solution, especially for the lcte relations. Finally, we achieve 0.39 on the test set and win the 5th prize.

1. Introduction

Lane detection and traffic element detection are crucial for reliable autonomous driving perception systems. In recent years, deep learning has achieved remarkable success in detecting lane lines [6] and traffic elements [14]. Go beyond conventional lane line detection as segmentation. This task requires recognizing lanes as an abstraction of the scene-centerline, and building the topology between lanes and traffic elements. Such a topology is to facilitate planning and routing. However, how to build topology between lanes and traffic elements in an end2end manner is less explored.

Toward this, we build a baseline with a multi-view centerline detection method (*e.g.*, STSU [1], mapTR [9], etc.). Then a traffic element detection head is added with a DETR-like decoder. As the official baseline solution builds the topology with multiple MLP layers, the implicit learning might neglect the characteristics of each image, leading to negative optimization for topology with explicit co-occurrences. To solve this problem, we propose a dynamic graph network for building relations between detected lanes and traffic elements. As in Fig 1. a), the simple topology head directly concatenates two embedding of lanes or traffic elements, following 3-layer MLP to learn the relations in a binary classification manner. However, in this direct learning manner, the characteristic of each image are less taken into consideration. To explicitly construct correlations between lanes and traffic elements, we resort to graph networks to model this relationship as in Fig 1. b). We propose an efficient relation head with dynamic relation graph to model the joint relations between two embeddings with the learnable correlations matrix.

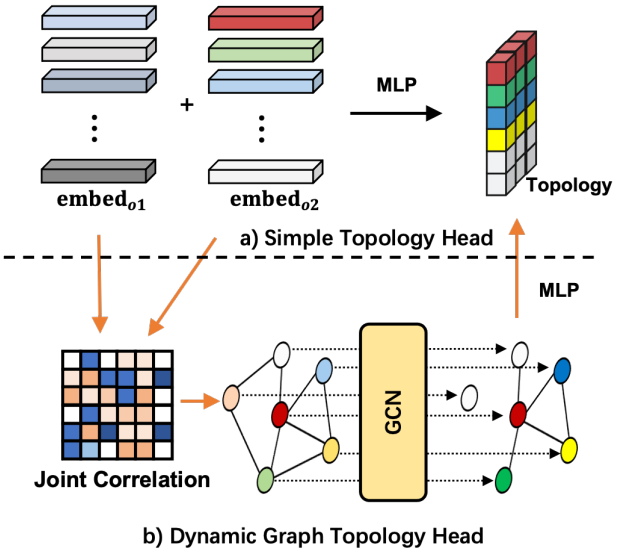

Figure 1. Motivation of the proposed dynamic relations. a) The simple topology head directly concatenates two embeddings and learns the relations. b) The dynamic relation graph head builds dynamic correlations of two embeddings for a better road structure understanding.

2. Related Work

DETR-series Approachs. Different from CNNs and RNNs, transformer is recently proposed to extract intrinsic features with self-attention mechanism [12]. Transformer has demonstrated its success in natural language processing tasks [13, 4]. As the pioneer work, Vaswani *et al.* [13] first propose the vanilla Transformer architecture, which is

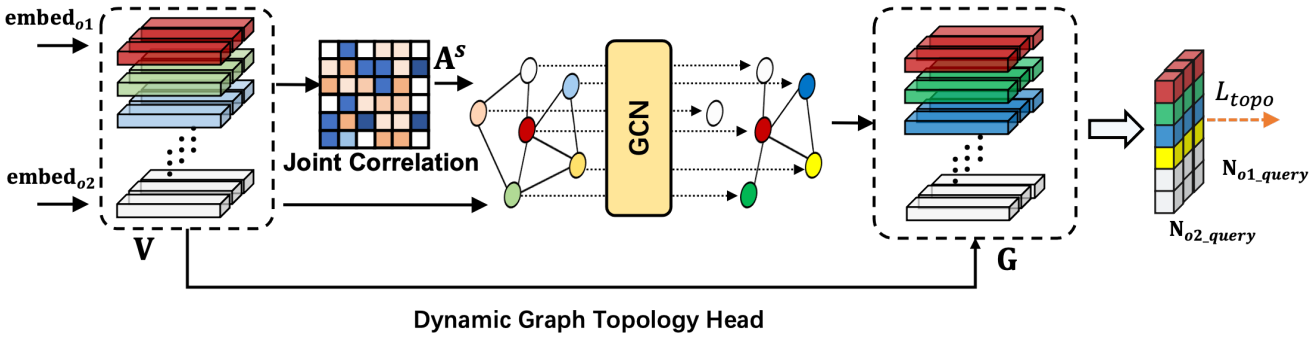

Figure 2. The detailed architecture of our proposed dynamic graph topology head, which consists of two essential modules: the joint correlation matrix to learn dynamic correlations of two embeddings and the graph neural network to model the joint relationship between two embeddings with the learnable correlation matrix.

based on self-attention mechanisms for machine translation. Transformer has not only obtained great breakthrough in NLP tasks, but also shows huge potential in Computer Vision (CV) tasks [5, 2, 16, 15, 8]. For example, recently, Carion *et al.* [2] design a fully end-to-end object DEtection TRansformer (DETR), which shows impressive performance on object detection. Zhu *et al.* [16] introduce a deformable attention module to solve the defects of DETR, *e.g.*, poor performance on small objects.

3. Approach

Strong Baseline. We propose a stronger baseline to obtain a more robust representation and better performance. Firstly, we replace the backbone from resnet50 to swin-base pre-trained on ImageNet. Then we scale the BEVFormer encoder layers from 3 to 6 to learn better BEV features. Moreover, we utilize more centerline queries to locate more accurate lane predictions. Finally, we train the train set and val set together for more robust performance.

Dynamic Graph Topology Head. To learn a better representation of topology relations in an end2end manner. Especially to explicitly construct correlations between lanes and traffic elements, we resort to graph networks to model this relationship. Graph neural networks propagate messages between adjacent nodes based on correlation matrix. We first update embeddings with 3 MLP layers with dropout and then align two embedded feature dimensions with their query numbers. As in Fig. 3, we build the joint correlation matrix $\mathbf{A}^s$ from two embeddings in topology modeling head, *i.e.* updated traffic elements embedding $\mathbf{V}_t$ and updated centerline embedding $\mathbf{V}_l$ in a learnable manner:

$$\mathbf{A}^s = \texttt{sigmoid}(\varphi_c(\texttt{concat}(\varphi_t(\mathbf{V}_l), \mathbf{V}_t))), \quad (1)$$

where $\varphi_{\{c,t\}}$ denote the learnable dimension transformation operation for concat feature and traffic elements embedding respectively.

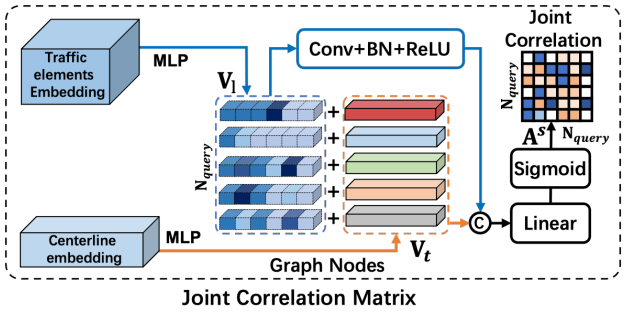

Figure 3. Motivation of the proposed dynamic relations. a) The simple topology head directly concatenates two embeddings and learns the relations. b) The dynamic relation graph head builds dynamic correlations of two embeddings for a better road structure understanding.

Obtaining graph nodes $\mathbf{V} = \texttt{concat}(\mathbf{V}_l, \mathbf{V}_t)$ and correlation matrix $\mathbf{A}^s$, we further model the joint topology relations between these two embeddings based on correlation matrix using Kipf *et al.*'s [7] Graph Convolutional Networks, which can be formulated as:

$$\mathbf{G} = \delta(\mathbf{A}^s \mathbf{V} \mathbf{W}_G) + \mathbf{V}, \quad (2)$$

where $\mathbf{G}$ denotes the updated dynamic topology relation graph, $\mathbf{W}_G$ is the learnable graph weights. $\delta(\cdot)$ denotes the LeakyReLU [11] activation function.

3.1. Learning Objective

The traffic elements detection head is supervised by L_{te}, which decomposed into a classification loss L_{cls} (Focal Loss), a regression loss L_{reg}, and an IoU loss L_{iou} (GIoU):

$$L_{te} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} + \lambda_{iou} L_{iou}$$

where $\lambda_{cls}, \lambda_{reg}, \lambda_{iou}$ is set to $1.0, 2.5, 1.0$ respectively.

The centerline detection head is supervised by L_{lc}, which decomposed into a classification loss L_{cls} (Focal Loss) and a regression loss L_{reg}:

$$L_{lc} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg}$$

where $\lambda_{cls}, \lambda_{reg}$ is set to $1.5, 0.0075$ respectively.

The topology head is supervised by L_{topo}, which decomposed into a classification loss L_{cls} (Focal Loss):

$$L_{topo} = \lambda_{lclc} L_{lclc} + \lambda_{lcte} L_{lcte}$$

where both $\lambda_{lclc}, \lambda_{lcte}$ are set to 10.

4. Experiments

4.1. Datasets and Evaluation Metrics

OpenLane-V2. The OpenLane-V2 dataset is the perception and reasoning benchmark for scene structure in autonomous driving. Given multi-view images covering the whole panoramic field of view, participants are required to deliver not only perception results of lanes and traffic elements but also topology relationships among lanes and between lanes and traffic elements simultaneously.

Evaluation Metrics. To evaluate performances on different aspects of the task, several metrics are adopted: DET_l for mAP on directed lane centerlines, DET_t for mAP on traffic elements, TOP_{ll} for mAP on topology among lane centerlines, TOP_{lt} for mAP on topology between lane centerlines and traffic elements.

4.2. Implementation Details

We adopt swin-base [10] pre-trained on ImageNet [3] as our backbone. A deformable DETR encoder is employed without temporal feature, while a BEVFormer encoder without temporal information is adopted to construct the BEV feature. We adopt the AdamW optimizer and a consine annealing schedule with an initial learning rate of 2e-4. Our proposed method is trained for 24 epochs with a batch size of 8 with 8 NVIDIA Tesla A100 GPUs. We set the BEVFormer encoder layer as 6, the lane query number as 300, the traffic element query number as 100. The resolution of input images is 2048 ×1550, except for the front-view image which is in the size of 1550×2048 and is cropped into 1550 × 1550. Our implementation is based on the open-source object detection toolbox MMDetection3d.

4.3. Performance Analysis

Ablation Studies. To evaluate the effectiveness of our proposed strong baseline and dynamic graph topology head, we reconstruct our model with different ablation factors in Tab. 1. Compared with the baseline, our proposed strong baseline could effectively improve all the metrics, especially for the detection scores due to the stronger feature representation. Furthermore, applying the proposed dynamic graph topology head could significantly improve the topology scores due to the better relation learning manner between embeddings. Specifically, the TOP_{ll} doubled and the TOP_{lt} also improved to 0.21 with a clear margin.

Table 1. Ablation study for different components on the test set. The strong baseline and our proposed dynamic graph head improve the performance significantly compared with the baseline.

Method	DET_l	DET_t	TOP_{ll}	TOP_{lt}	**OLS**
baseline_large(r50)	0.11	0.64	0.01	0.08	**0.28**
+swinB+val+lc300	0.14	0.70	0.02	0.15	**0.35**
+Dynamic graph	0.18	0.70	0.04	0.21	**0.39**

5. Conclusion

In this paper, we propose an effective dynamic graph topology head for end2end openlane topology modeling and provide a strong baseline for better performance. Finally, we achieve 0.39 on the test set and win the 5th prize.

References

[1] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15661–15670, 2021. 1

[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. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Songbin Li Jigang Tang and Peng Liu. A review of lane detection methods based on deep learning. In *Pattern Recognition*, 2020. 1

[7] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[8] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 16478–16488, 2021. 2

[9] 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

[10] Yutong Lin Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Liu, Ze and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[11] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013. 2

[12] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016. 1

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 1

[14] Ma Weixin Shang Yanmei Xiong Changzhen, W. Cong. A traffic sign detection algorithm based on deep convolutional neural network. In *IEEE International Conference on Signal and Image Processing*, 2016. 1

[15] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 2

[16] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2