

LaneGraph2Seq: Lane Topology Extraction with Language Model via Vertex-Edge Encoding and Connectivity Enhancement

Renyuan Peng¹, Xinyue Cai², Hang Xu^{2†}, Jiachen Lu¹, Feng Wen², Wei Zhang², Li Zhang^{1†}

¹School of Data Science, Fudan University

²Huawei Noah's Ark Lab

chromexbjxh@gmail.com, lizhangfd@fudan.edu.cn

Abstract

Understanding road structures is crucial for autonomous driving. Intricate road structures are often depicted using lane graphs, which include centerline curves and connections forming a Directed Acyclic Graph (DAG). Accurate extraction of lane graphs relies on precisely estimating vertex and edge information within the DAG. Recent research highlights Transformer-based language models' impressive sequence prediction abilities, making them effective for learning graph representations when graph data are encoded as sequences. However, existing studies focus mainly on modeling vertices explicitly, leaving edge information simply embedded in the network. Consequently, these approaches fall short in the task of lane graph extraction. To address this, we introduce *LaneGraph2Seq*, a novel approach for lane graph extraction. It leverages a language model with vertex-edge encoding and connectivity enhancement. Our serialization strategy includes a vertex-centric depth-first traversal and a concise edge-based partition sequence. Additionally, we use classifier-free guidance combined with nucleus sampling to improve lane connectivity. We validate our method on prominent datasets, nuScenes and Argoverse 2, showcasing consistent and compelling results. Our LaneGraph2Seq approach demonstrates superior performance compared to state-of-the-art techniques in lane graph extraction. Code is available at <https://github.com/fudan-zvg/RoadNet>

Introduction

Road structure understanding serves as a foundational pillar that empowers autonomous vehicles to navigate safely, efficiently, and intelligently within complex real-world scenarios (Cui et al. 2019; Chen et al. 2020; Hong, Sapp, and Philbin 2019; Espinoza et al. 2022). A road structure is suitable to be represented by a Directed Acyclic Graph (DAG) due to its characteristics of one-way traffic flow, acyclic layout, complex lane topologies, and meaningful spatial relationships. The attributes of a DAG align with the inherent features of roads, making it a potent and intuitive representation for modeling and analyzing road structures. Hence, the complex road network is frequently represented as a directed acyclic lane graph, which includes the *centerline curves* and their *connection relationships* (Can et al. 2021,

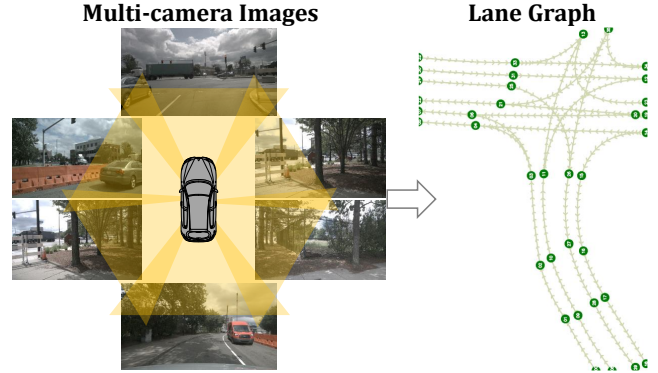


Figure 1: The extraction of a lane graph, which captures the centerline curves and their intricate connectivity relationships on the road, is crucial for the perception system of autonomous driving. Our proposed *LaneGraph2Seq* framework adeptly addresses this challenge by utilizing multi-camera images as input.

2022; Büchner et al. 2023). Nevertheless, extracting a lane graph using onboard sensors is inevitable to be a challenging task due to the intricate nature of the road structure.

The prevailing frameworks for lane graph extraction often draw inspiration from models like STSU (Can et al. 2021) and TPLR (Can et al. 2022). These frameworks model centerline curves as vertices and represent the interconnections between centerlines as edges. In the process of lane graph extraction, they employed a DETR-like module for detecting the centerline's shape and position. Subsequently, an MLP is employed to approximate the connectivity between any pair of centerlines. This two-stage framework predicts the vertices and edges of the lane graph with a relatively independent approach. While utilizing visual features, this method falls short in terms of overall graph feature learning.

Graph Neural Network (GNN) and its variants are widely adopted to aggregate features of vertices and extract information from graph data (Velickovic et al. 2018; Hamilton, Ying, and Leskovec 2017; Kipf and Welling 2017). Recently, Graphormer (Ying et al. 2021) has demonstrated that by adeptly encoding the structural information of a graph into the model, the standard Transformer architecture can surpass GNNs and achieve remarkable outcomes across a wide spectrum of graph representation learning tasks. How-

[†]Corresponding authors are Hang Xu and Li Zhang.

ever, Graphormer solely focuses on explicitly encoding the vertices of the graph into sequences. In the context of lane graph extraction, the task necessitates the prediction of both vertices and edges. This crucial distinction renders existing Transformer-based graph representation learning methods, such as Graphormer, unsuitable for the specific demands of lane graph extraction.

To enhance the acquisition of graph features and facilitate lane graph extraction, we present **LaneGraph2Seq**—a novel approach utilizing a Transformer-based language model with vertex-edge encoding and connectivity enhancement. Precisely, our approach employs the language model with Transformer architecture to grasp the intricacies of the lane graph. By representing the graph’s vertices and edges as a sequence of discrete tokens, we encapsulate the complete information of a lane graph within it. Consequently, this challenge, graph prediction, is transmuted into a sequence prediction task. Leveraging bird’s-eye-view (BEV) encoder and language model with Transformer architecture, we achieve the dual objectives of acquiring graph features from onboard sensors and predicting graph sequences. Furthermore, to mitigate the occurrence of premature termination or continuous loops in sequence prediction, enhancing the diversity of language model prediction samples proves to be an effective strategy. Hence, we employ classifier-free guidance (Ho and Salimans 2021) and nucleus sampling (Holtzman et al. 2019) to augment the connectivity conditions and increase the sampling diversity. This mechanism enables the network to strike a balance between accuracy and completeness in edge prediction.

The **contributions** of this work are summarized as follows: (i) We introduce an innovative framework for lane graph extraction, known as **LaneGraph2Seq**, which casts the task as a sequence-to-sequence prediction challenge. Specifically, we present a vertex-edge encoding approach for sequence construction and employ a language model to simultaneously analyze visual features and extract graphical attributes. (ii) During the inference stage, we employ a method that combines classifier-free guidance with nucleus sampling to enhance the diversity of sampling. This approach contributes to improved accuracy in edge prediction and a reduction in the false negative rate by enhancing lane graph connectivity conditions. (iii) Extensive experiments conducted on two large-scale datasets (nuScenes (Caesar et al. 2020) and Argoverse 2 (Wilson et al. 2021)) demonstrate that our approach attains state-of-the-art performance in lane graph extraction (Figure 1).

Related Work

Bird’s-Eye-View (BEV) Semantics Learning Recently, there’s been growing interest in transforming monocular or multi-view images from ego car cameras into bird’s-eye-view (BEV) representations. To learn BEV feature from onboard cameras, OFT (Roddick, Kendall, and Cipolla 2019), LSS (Phillion and Fidler 2020) and FIERY (Hu et al. 2021) predict 3D spatial features from front view images through real or predicted depth information, and then obtain BEV features from 3D spatial features. (Saha et al. 2022; Roddick and Cipolla 2020; Li et al. 2022; Lu et al. 2022) lever-

age a Transformer to implicitly learn 3D spatial information and get features in the bird’s eye view. To capture the lane graph represented by BEV, we employ the LSS technique works (Phillion and Fidler 2020; Huang et al. 2021) to extract BEV features from onboard cameras.

Lane Graph Extraction To extract the lane graph online, STSU (Can et al. 2021) first identifies centerlines using a Transformer-based approach on image features. Subsequently, it forecasts centerline associations using an MLP layer, culminating in a final merging step to construct the lane graph. Based on STSU, TPLR (Can et al. 2022) introduces a minimal cycle to eliminate ambiguity in its connectivity representation. Furthermore, LaneGAP (Liao et al. 2023) adopts an innovative pathwise approach to modeling the lane graph, effectively preserving the lane’s continuity.

While the previously mentioned methods concentrate on local visual attributes, they overlook the lane graph’s inherent features. In contrast, our approach involves encoding the lane graph into sequences and leveraging a language model for predictions. This methodology not only facilitates the acquisition of visual features but also comprehends the graph’s distinctive characteristics.

Language Modeling While initially conceived for natural languages, language modeling has demonstrated its aptitude in modeling a range of sequential data tasks (Sutskever, Vinyals, and Le 2014; Raffel et al. 2020; He et al. 2016). Motivated by these achievements in natural language processing, contemporary efforts within the realm of computer vision and vision-language have also begun to delve into the utilization of language modeling for various tasks (Chen et al. 2021, 2022; Wang et al. 2022; Liu et al. 2023).

These instances of success inspire us that language models excel not only in comprehending linguistic logic but also exhibit proficiency in visual localization. This demonstrates that employing sequence prediction methods enables the language model to effectively predict both road topology (centerline connection) and the shape of centerline curves.

Graph Represent Learning Graph Neural Networks (GNNs) (Scarselli et al. 2008; Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017; Velickovic et al. 2018) play a pivotal role in aggregating information and extracting insights from graph-structured data. This methodology has also been employed in recent endeavors to perceive autonomous driving scenarios (Weng, Yuan, and Kitani 2021; Weng et al. 2020). Notably, LaneGCN (Liang et al. 2020) constructs a lane graph from HD maps, while TopoNet (Li et al. 2023) employs a GNN model to accomplish road network extraction and detect traffic elements. Recent research (Ying et al. 2021) has further demonstrated that the Transformer architecture can surpass GNNs in a broad spectrum of graph-level prediction tasks when the graph information is pertinently embedded.

Inspired by these findings, we adopt a vertex-edge encoding method to serialize the lane graph and subsequently extract lane graph using a Transformer structure.

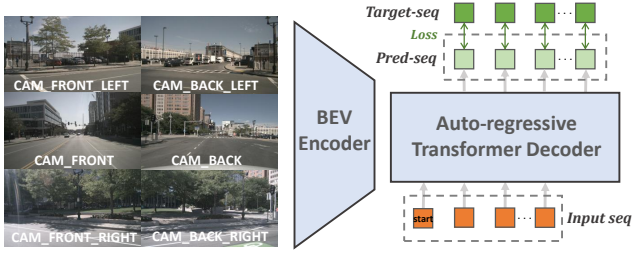


Figure 2: Our LaneGraph2Seq employs a BEV-encoder to transition features from the front view image to the bird’s-eye view plane. Subsequently, a Transformer decoder generates tokens of the target sequence in sequence, guided by prior tokens and the encoded BEV feature.

Method

Lane Graph Representation The lane graph comprises centerlines and their associated junction points (Can et al. 2021, 2022). Because traffic flows in a single direction on each lane, the graph can be represented as a Directed Acyclic Graph (DAG), *i.e.*, $G = (V, E)$ where the vertex set V is the set of all junction points and the edge set E is the set of all centerlines. Each vertex $v = (v_x, v_y) \in V$ holds the coordinates of junction points. Each edge $e = (e_s, e_m, e_t) \in E$ is characterized by three Bezier control points: the source point, the midpoint, and the target point. The shape of the curve can be determined by the coordinates of the three control points in this specific order.

Overall Structure of the Sequence As shown in Figure 3, the sequence of a lane graph comprises two components: the *vertex sequence* and the *edge sequence*.

The vertex sequence consists of vertices coordinates arranged in a specific order. Since lane graphs encompass varying vertex quantities, the resulting sequences will exhibit distinct lengths. In order to demarcate the conclusion of a vertex sequence, we introduce an $\langle \text{EOV} \rangle$ (End of Vertex Sequence) token.

As mentioned before, an edge can be represented as a set of the source vertex, the target vertex, and the Bezier middle control point. For each vertex in DAG, We denote the target vertex using the sequential index of its child nodes within the vertex sequence. The source vertex of each edge is determined by arranging the edge in the order of its source vertex in the vertex sequence.

So with three parameters ($[\text{Index}(\text{childnode}), e_{mx}, e_{my}]$) and the position of this triple, the curve shape and direction of an edge can be determined. Due to the variable nature of the out-degree for each vertex in the DAG, the number of edges extended by each vertex may vary. Consequently, we employ a $\langle \text{Split} \rangle$ token to ascertain the end of the edge subsequence corresponding to a given vertex as the parent node. And at the end of the entire edge sequence, we set the $\langle \text{EOE} \rangle$ (End of Edge Sequence) token to determine the termination.

Besides, we use $\langle \text{Start} \rangle$ token to indicate the beginning of the whole sequence. The $\langle \text{N/A} \rangle$ token is utilized to separately pad the vertex sequence and edge sequence,

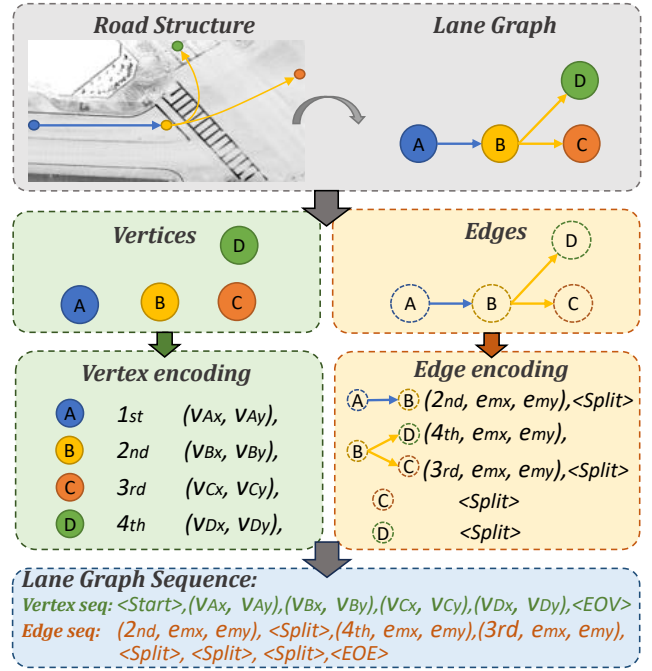


Figure 3: This depiction outlines the procedure for constructing a sequence that represents the actual road structure. The upper part illustrates the abstraction of the real road into a Directed Acyclic Graph (DAG), while the middle section showcases the detailed process of encoding vertices and edges. The lower part exhibits the resulting vertex sequence and edge sequence after vertex-edge encoding, which is then combined to form the comprehensive lane graph sequence.

which guarantees that both of them maintain a consistent length.

Serialization While the order of edges is linked to the sequence of vertices, the exact positioning of the vertices remains uncertain. Hence, serializing the vertices into a deterministic sequence becomes essential for constructing a lane graph. In contrast to detection and segmentation tasks (Chen et al. 2021; Liu et al. 2023), there are reasons to believe that the traversal order of the lane graph impacts the eventual performance. We explored several deterministic ordering strategies, including depth-first traversal, breadth-first traversal, sorting by coordinate value magnitude, and random sorting. Figure 4 depicts the sequence construction process using distinct ordering strategies. We assume that utilizing a depth-first traversal order is better suited for enabling the auto-regressive Transformer network to autonomously explore and reveal the entirety of the lane graph.

Discretization While the indices and tokens with special significance ($\langle \text{Start} \rangle$, $\langle \text{EOV} \rangle$, $\langle \text{EOE} \rangle$, $\langle \text{Split} \rangle$, $\langle \text{N/A} \rangle$) are represented as discrete tokens, the coordinates of the vertices and Bezier middle control points are not treated discretely. Following (Chen et al. 2021), We discretize the continuous coordinate values by partitioning the image into varying numbers of bins. The greater the number of bins, the smaller the quantization error. We use a

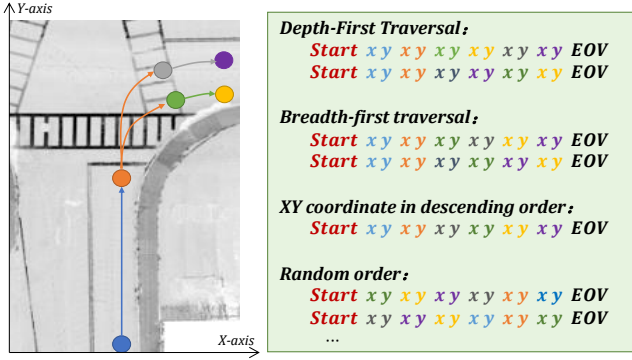


Figure 4: Examples of serialization order. *Right* provides possible example vertex sequences of different sorting methods. *Start* denotes the `<start>` token, while *EOV* signifies the `<EOV>` token. The differently colored *xy* entries indicate the *x* and *y* coordinate values of points corresponding to the colors on the *left*. As edge order aligns with point order, we exclusively present the vertex sequence.

shared fixed vocabulary for all tokens. It’s worth noting that, for the network to learn distinct token classes, these classes are distributed across varying ranges within the vocabulary. For instance, if vertex coordinates fall within the range of $[1, num_bins]$, and an offset constant would be added to the Bezier control point coordinates, shifting their values into the range of $[num_bins, num_bins * 2]$. Furthermore, tokens with special significance fall outside of these ranges.

Architecture Just like in various other downstream tasks within autonomous driving, our approach involves initially utilizing a BEV-encoder to transform the features from the front view images onto the bird’s-eye view plane. Subsequently, we employ a decoder like (Chen et al. 2021), which generates tokens of the target sequence sequentially based on preceding tokens and the encoded BEV feature. The overall architecture is illustrated in the Figure 2.

Objective Given the ground-truth sequence denoted as y with a length of L , and the predicted sequence denoted as \hat{y} , the model’s objective is to minimize the maximum likelihood loss, which can be expressed as:

$$\max \sum_{i=1}^L w_i \log P(\hat{y}_i | y_{<i}, \mathcal{F}) \quad (1)$$

Here, \mathcal{F} is the BEV feature, y_i represents the i^{th} token of y , $y_{<i}$ signifies all tokens preceding y_i , and w_i stands for the class weight.

In practical implementation, to prevent premature termination predictions, we reduce the weights of `<EOV>`, `<Split>`, and `<EOE>` tokens accordingly. The `<N/A>` token, employed for padding, does not contribute to the back-propagation of loss.

Inference with Connectivity Enhancement The `<EOS>` and `<EOE>` tokens empower the model to determine when to conclude the generation of the vertex sequence and the edge sequence. Similarly, the `<Split>` token enables the

model to ascertain when to halt the generation of a sub edge sequence pertaining to a vertex. In practice we find that the model tends to finish without predicting all edges.

Classifier-free guidance entails guiding an unconditioned sample towards a conditioned counterpart (Gafni et al. 2022). This approach directs a model’s generation process without the need for explicit classifiers or labels. It finds application in various domains such as text generation, image synthesis, and sequence prediction, offering increased flexibility and creativity in outputs while avoiding limitations of traditional classification methods.

The vertex sequence encompasses the coordinates of all junction points, encapsulating the lane graph’s connectivity information and providing a strong prior for generating the edge sequence. So we take the vertex sequence as condition to generate the edge sequence. Continuing the approach from previous research (Gafni et al. 2022), in the inference phase, we generate two simultaneous token streams: a conditioned token stream dependent on the vertex sequence, and an unconditioned token stream linked to a mask token. For the Transformer decoder, we apply classifier-free guidance on logit scores:

$$logits_{cond} = T(t_e | (t_v, \mathcal{F})) \quad (2)$$

$$logits_{uncond} = T(t_e | (<mask>, \mathcal{F})) \quad (3)$$

$$logits_{cf} = logits_{uncond} + \alpha_c (logits_{cond} - logits_{uncond}) \quad (4)$$

where T represents the Transformer decoder, t_e and t_v represent the token of edge sequence and vertex sequence respectively, \mathcal{F} stands for the BEV feature, and α_c is the scale factor.

Through the implementation of classifier-free guidance and utilizing the vertex sequence as a condition, the model increases the diversity of samples on the basis of enhanced connectivity during edge sequence prediction. This results in more comprehensive and accurate edge predictions.

Sequence to Lane Graph Initially, we identify the first `<EOV>` and `<EOE>` tokens in the sequence, extracting the vertex sequence and edge sequence without padding. By detecting `<Split>` tokens, we divide the edge sequence into subsequences. Each subsequence signifies edges, with the node located at the corresponding position serving as the parent node. Hence, with all the vertices and edges in place, the lane graph is reconstructed through the sequence.

Experiments

Dataset

We conducted benchmarking on two challenging datasets: **nuScenes** (Caesar et al. 2020) The comprises 1000 sequences, officially divided into train, validation, and test sets with 700, 150, and 150 scenes, respectively. Each sequence is captured at a 2Hz frame rate, offering RGB images from 6 surrounding cameras, covering a 360-degree horizontal field of view around the ego-vehicle.

Argoverse 2 (Wilson et al. 2021) is even more extensive, featuring 1000 sequences collected from 6 different cities. It is randomly partitioned into the train, validation, and test

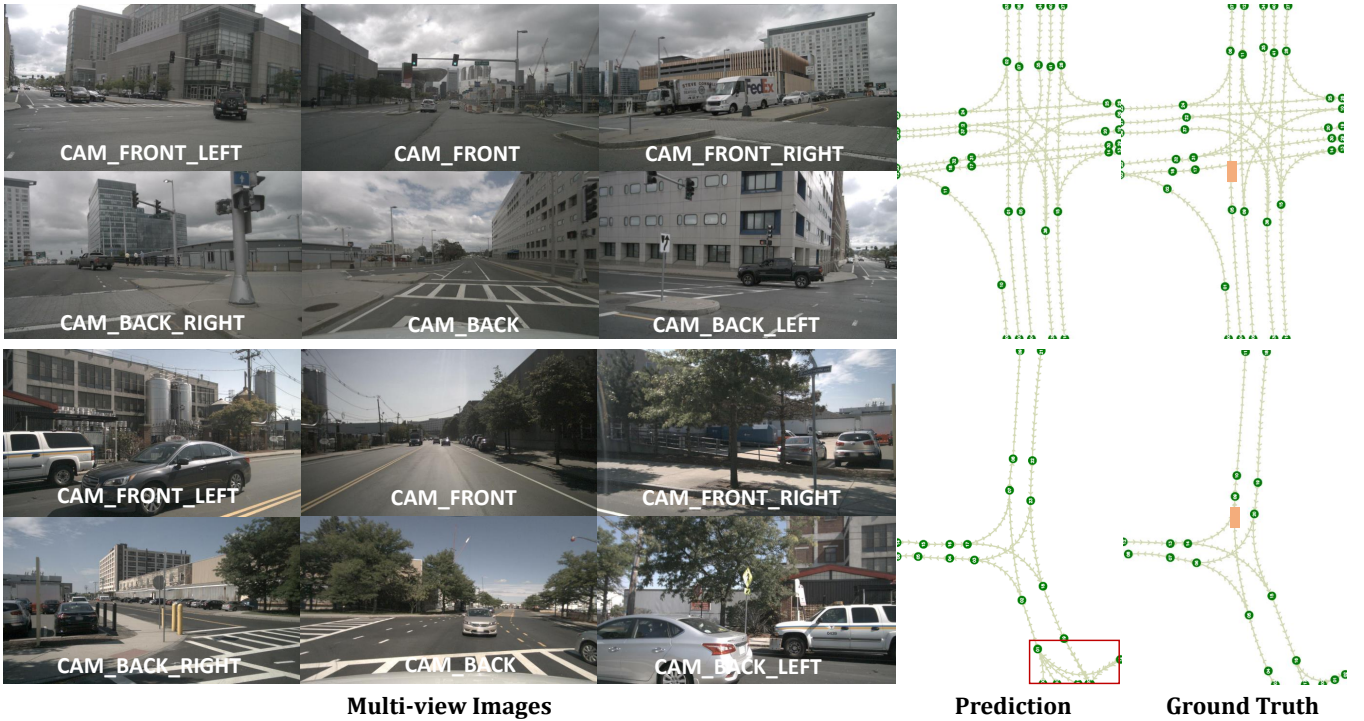


Figure 5: Our qualitative results on nuScenes (Caesar et al. 2020) validation set. Evidently, our approach demonstrates an impressive ability to attain highly accurate predictions, with only a slight error in the red-boxed section.

splits, with ratios of 700, 150, and 150 sequences respectively. Argoverse 2 provides images from 7 surrounding cameras and offers a significantly larger volume of data, with four times the amount available in the nuScenes.

Both datasets offer lane graph information represented as lane centerlines. In our online lane graph construction scenario, the target BEV range extends from -48 to 48m along the X-axis and -32 to 32m along the Y-axis, with points sampled at intervals of 0.5m. We conduct training and evaluation using the specified official training and validation sets for both datasets with all available surrounding-view images.

Implementation Details

Our BEV encoder, following the approach of LSS (Phillion and Fidler 2020), employs either the ResNet50 (He et al. 2016) or VovNetV2 (Lee and Park 2020) backbone to transform input RGB images into BEV features. For improved initialization, we pretrain the BEV encoder on the task of centerline segmentation.

For sequence construction, we pad the vertex sequence to 200 and the edge sequence to 400, resulting in a total lane sequence length of 600. The Transformer decoder layers are set to 6. We trained the network for 300 epochs on 8 NVIDIA V100 GPUs using the AdamW optimizer, with an initial learning rate of 2×10^{-4} and a batch size of 2×8 . During the training process, we apply random flip, random rotation, and random scaling on BEV feature similar to (Huang et al. 2021; Lu et al. 2022). In the inference process, we configure the α_c parameter of the classifier-free guidance to 4.

Metrics

To assess the accuracy of lane graph extraction, we employed the identical evaluation metrics as those utilized in STSU (Can et al. 2021):

Precision-Recall (M-P, M-R, M-F ratio) measures how closely the estimated centerlines fit the matched ground truth (GT) centerlines and how accurately the subgraph is captured. This metric focuses on matching estimated centerlines to GT ones, without penalizing unmatched true centerlines.

Detection ratio (Detect) quantifies the portion of distinct ground truth centerlines matched by at least one estimated line, addressing the precision-recall metric’s limitation by considering missed centerlines. High precision-recall and low detection ratio suggest proximity between estimated and matched GT lines, yet numerous GT centerlines go undetected.

Connectivity (C-P, C-R, C-F ratio) evaluates the association and connectivity between the estimated centerlines using a precision-recall-based approach. This metric focuses on the edge of the lane graph and measures whether the estimated centerlines have a similar connection relationship with the ground truth.

Comparison with the State-of-the-Art

We compare our proposed model with previous state-of-the-art methods for lane graph extraction on the nuScenes dataset. Detailed outcomes of this comparison are presented in Table 1. Impressively, our model outperforms all prior including STSU (Can et al. 2021) and TPLR (Can et al. 2022) approaches across every evaluation metric. Notably,

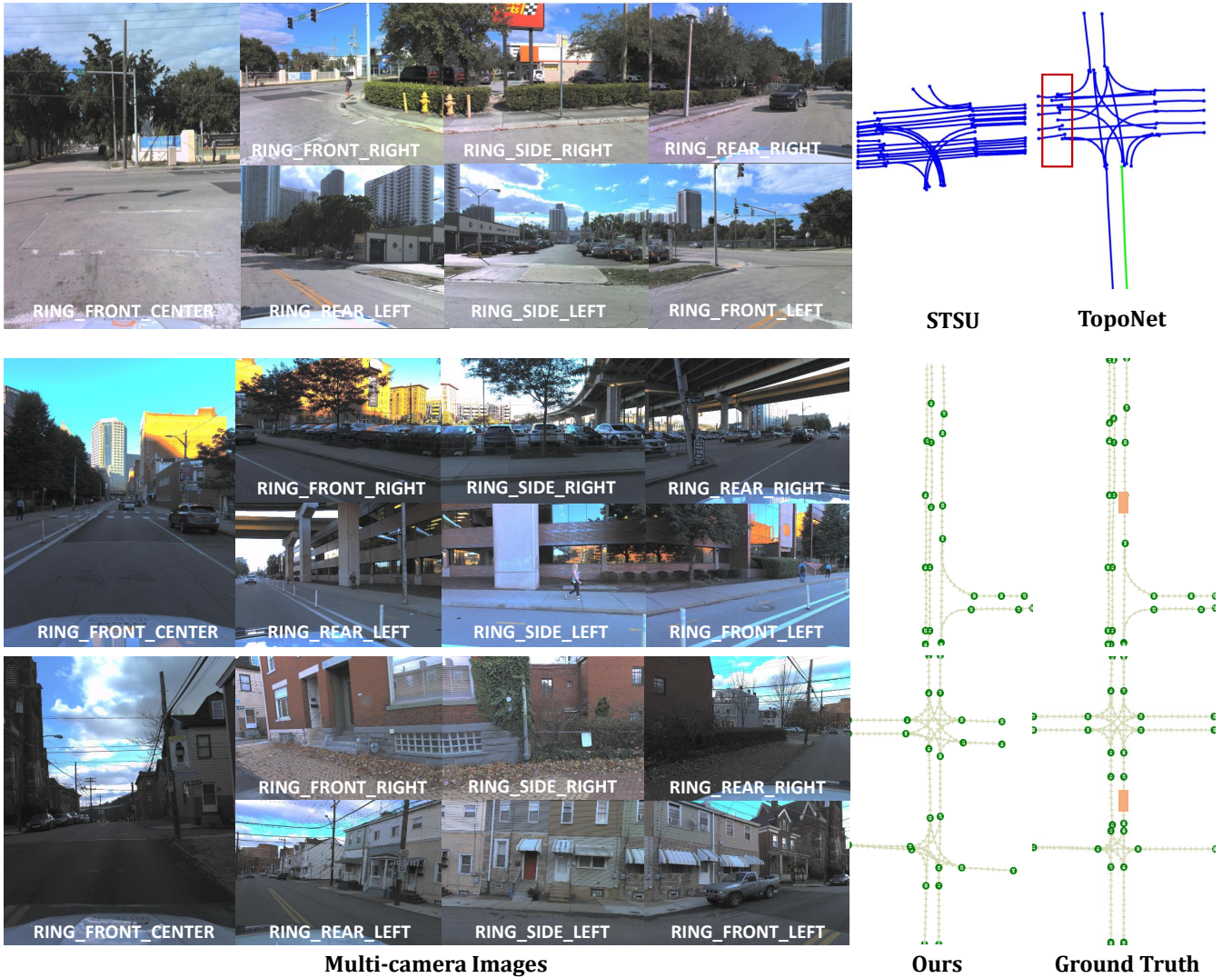


Figure 6: Qualitative outcomes on the Argoverse 2 (Wilson et al. 2021) validation set. The first row (Li et al. 2023) shows the visualization result of STSU (Can et al. 2021) and TopoNet (Li et al. 2023). The results indicate that STSU inaccurately predicts the road layout, and TopoNet displays minor deficiencies in the vicinity of junction points. The results show that our approach effectively predicts the overall layout of the road structure with accuracy, especially with continuous and clear junction points.

Precision-Recall metrics show an improved match between estimated centerlines and ground truth. This notable improvement results from innovative language modeling techniques, including vertex-edge encoding and connectivity enhancement.

Ablation Studies

We perform a comprehensive set of ablation studies on the validation set of nuScenes (Caesar et al. 2020). All the results presented in this section were obtained utilizing six surrounding-view images as input.

Connectivity Enhancement Building upon the insights from the ablation study on the nuScenes dataset concerning the parameter α_c , the results reveal a non-linear trend: performance first escalates and subsequently diminishes, reaching its zenith when $\alpha_c = 4$. At this value, metrics like M-F,

Detect, and C-F all register at their optimum, underscoring the effectiveness of the chosen approach. Such a trend suggests an inherent trade-off between diversity and accuracy. Pushing α_c towards its peak value ensures that the model does not easily overlook pivotal points in the graph, striking a balance between breadth and precision of detection. The specific experimental data are shown in the Figure 7.

Serialization Order In the ablation study detailed in Table 2, conducted on the nuScenes dataset using VoVNetV2 as the image backbone, the depth-first traversal (DFS) method stands out as the most effective serialization strategy. Conversely, the $Coord_{xy}$ method, which sorts vertices based on ascending xy coordinate values, records the weakest performance. This data underscores the importance of traversal-based serialization, like DFS, in capturing spatial relationships between vertices, a crucial factor for superior

Methods	Dataset	M-P	M-R	M-F	Detect	C-P	C-R	C-F
PINET (Ko et al. 2021)	nuScenes	54.1	45.6	49.5	19.2	-	-	-
Poly (Acuna et al. 2018)	nuScenes	54.7	51.2	52.9	40.5	58.4	16.3	25.5
STSU (Can et al. 2021)	nuScenes	60.7	54.7	57.5	60.6	60.5	52.2	56.0
TPLR (Can et al. 2022)	nuScenes	-	-	58.2	60.2	-	-	55.3
LaneGraph2Seq	nuScenes	64.6	63.7	64.1	64.5	69.4	58.0	63.2
LaneGraph2Seq [†]	nuScenes	68.1	79.3	73.3	68.7	75.2	61.4	67.6
LaneGraph2Seq	AV2	62.6	60.9	61.7	62.6	60.1	57.3	60.1
LaneGraph2Seq [†]	AV2	65.6	70.7	68.0	65.9	66.5	59.8	63.9

Table 1: Comparison of state of the art and our method on nuScenes dataset and Argoverse 2 dataset. ResNet-50 (He et al. 2016) is applied as image backbone by default. “[†]” use VoVNetV2 (Lee and Park 2020) pretrained on extra data as backbone. M-P, M-R, M-F stand for mean precision/recall/F1-score. Detect stands for Detection ratio metrics. C-P, C-R, C-F stand for connectivity precision/recall/F1-score.

Order	M-F	Detect	C-F
<i>DFS</i>	73.3	68.7	67.6
<i>BFS</i>	72.1	67.2	67.2
<i>Coord_{xy}</i>	69.8	61.7	61.2
<i>Random</i>	70.2	61.3	62.1

Table 2: Ablation study on serialization order with six cameras as input on nuScenes. VoVNetV2 is applied as image backbone by default. *DFS* and *BFS* denote depth-first traversal and breadth-first traversal, while *Coord_{xy}* indicates sorting vertices based on their xy coordinate values in ascending order, and *Random* is a random arrangement.

Layers	M-F	Detect	C-F
3	70.8	63.5	63.6
4	72.0	65.2	64.2
6	73.3	68.7	67.6
8	74.0	69.3	69.1

Table 3: Ablation study on number of Transformer layers with six cameras as input on nuScenes dataset. VoVNetV2 is applied as image backbone by default.

lane graph detection.

Transformer Layers In Table 3, an ablation study focusing on the number of Transformer layers is conducted on the nuScenes dataset. Among the evaluated configurations, the model with 8 Transformer layers achieves the highest scores, suggesting it is the optimal choice. The consistent increase in performance with the rising number of layers shows the importance of the large language model (LLM) in the task. The progressive enhancement with additional layers evidences the crucial role of depth in capturing intricate patterns and relationships for lane graph extraction.

Qualitative Results

In Figure 5, we present visualizations based on the nuScenes dataset, while Figure 6 displays visualizations from the Argoverse2 dataset. The depicted outcomes exhibit a remarkable resemblance between our predictions and the ground truth of the overall road layout, underscoring the efficacy of our method in conducting lane graph extraction.

From Figure 6, we can find that STSU (Can et al. 2021) exhibits inaccuracies in predicting the road layout, and To-

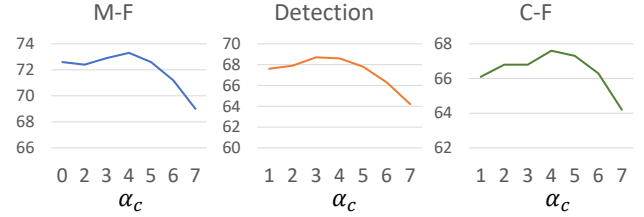


Figure 7: Altering parameter α_c in classifier-free guidance yields distinct M-F, Detection, and C-F outcomes. When α_c is set to 1, it equates to inference without connectivity enhancement.

poNet (Li et al. 2023) displays minor deficiencies in the vicinity of junction points. Our approach effectively predicts the overall layout of the road structure with continuous and clear junction points.

Conclusions

In this work, we presented **LaneGraph2Seq**, an advanced framework that harnesses the capabilities of Transformer-based language models for the task of lane graph extraction. By adopting a novel vertex-edge encoding mechanism coupled with a depth-first traversal and an edge-based partition sequence, our approach adeptly captures the intricate topologies and geometries inherent in lane graphs. Further refinement during the inference stage, through classifier-free guidance combined with nucleus sampling, elevates the accuracy of predictions while minimizing edge false-negative rates. Our comprehensive evaluations on the nuScenes and Argoverse 2 datasets underscore the superiority of our method. Looking ahead, leveraging large-scale pretraining from the HD map for the large language model offers a promising direction for enhancing the robustness and generalizability of our model in diverse and challenging road scenarios.

Acknowledgments

This work was supported in part by STI2030-Major Projects (Grant No. 2021ZD0200204), National Natural Science Foundation of China (Grant No. 62106050 and 62376060), Natural Science Foundation of Shanghai (Grant No. 22ZR1407500) and USyd-Fudan BISA Flagship Research Program.

References

- Acuna, D.; Ling, H.; Kar, A.; and Fidler, S. 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*.
- Büchner, M.; Zürn, J.; Todoran, I.-G.; Valada, A.; and Burgard, W. 2023. Learning and aggregating lane graphs for urban automated driving. In *CVPR*.
- Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.
- Can, Y. B.; Liniger, A.; Paudel, D. P.; and Van Gool, L. 2021. Structured bird’s-eye-view traffic scene understanding from onboard images. In *CVPR*.
- Can, Y. B.; Liniger, A.; Paudel, D. P.; and Van Gool, L. 2022. Topology preserving local road network estimation from single onboard camera image. In *CVPR*.
- Chen, D.; Zhou, B.; Koltun, V.; and Krähenbühl, P. 2020. Learning by cheating. In *CORL*.
- Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; and Hinton, G. 2021. Pix2seq: A language modeling framework for object detection. In *ICLR*.
- Chen, T.; Saxena, S.; Li, L.; Lin, T.-Y.; Fleet, D. J.; and Hinton, G. E. 2022. A unified sequence interface for vision tasks. In *NeurIPS*.
- Cui, H.; Radosavljevic, V.; Chou, F.-C.; Lin, T.-H.; Nguyen, T.; Huang, T.-K.; Schneider, J.; and Djuric, N. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *ICRA*.
- Espinoza, J. L. V.; Liniger, A.; Schwarting, W.; Rus, D.; and Van Gool, L. 2022. Deep interactive motion prediction and planning: Playing games with motion prediction models. In *LADC*.
- Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*.
- Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NeurIPS*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- Ho, J.; and Salimans, T. 2021. Classifier-free diffusion guidance. In *NeurIPS*.
- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The Curious Case of Neural Text Degeneration. In *ICLR*.
- Hong, J.; Sapp, B.; and Philbin, J. 2019. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *CVPR*.
- Hu, A.; Murez, Z.; Mohan, N.; Dudas, S.; Hawke, J.; Badrinarayanan, V.; Cipolla, R.; and Kendall, A. 2021. FIERY: future instance prediction in bird’s-eye view from surround monocular cameras. In *ICCV*.
- Huang, J.; Huang, G.; Zhu, Z.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint*.
- Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.
- Ko, Y.; Lee, Y.; Azam, S.; Munir, F.; Jeon, M.; and Pedrycz, W. 2021. Key points estimation and point instance segmentation approach for lane detection. *IEEE Transactions on Intelligent Transportation Systems*.
- Lee, Y.; and Park, J. 2020. Centermask: Real-time anchor-free instance segmentation. In *CVPR*.
- Li, T.; Chen, L.; Geng, X.; Wang, H.; Li, Y.; Liu, Z.; Jiang, S.; Wang, Y.; Xu, H.; Xu, C.; et al. 2023. Topology reasoning for driving scenes. *arXiv preprint*.
- Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*.
- Liang, M.; Yang, B.; Hu, R.; Chen, Y.; Liao, R.; Feng, S.; and Urtasun, R. 2020. Learning lane graph representations for motion forecasting. In *ECCV*.
- Liao, B.; Chen, S.; Jiang, B.; Cheng, T.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. *arXiv preprint*.
- Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023. PolyFormer: Referring image segmentation as sequential polygon generation. In *CVPR*.
- Lu, J.; Zhou, Z.; Zhu, X.; Xu, H.; and Zhang, L. 2022. Learning ego 3d representation as ray tracing. In *ECCV*.
- Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.
- Roddick, T.; and Cipolla, R. 2020. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*.
- Roddick, T.; Kendall, A.; and Cipolla, R. 2019. Orthographic feature transform for monocular 3d object detection. In *BMVC*.
- Saha, A.; Mendez, O.; Russell, C.; and Bowden, R. 2022. Translating images into maps. In *ICRA*.
- Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.
- Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2018. Graph attention networks. In *ICLR*.
- Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

- Weng, X.; Wang, Y.; Man, Y.; and Kitani, K. M. 2020. Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning. In *CVPR*.
- Weng, X.; Yuan, Y.; and Kitani, K. 2021. PTP: Parallelized tracking and prediction with graph neural networks and diversity sampling. *IEEE Robotics and Automation Letters*.
- Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; et al. 2021. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*.
- Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T. 2021. Do transformers really perform bad for graph representation? In *NeurIPS*.