Topo2Seq: Enhanced Topology Reasoning via Topology Sequence Learning

Yiming Yang[1,2], Yueru Luo[1,2], Bingkun He[3], Erlong Li[4], Zhipeng Cao[4], Chao Zheng[4], Shuqi Mei[4], Zhen Li[2,1] *

[1]FNii-Shenzhen, Shenzhen, China
[2]SSE, CUHK-Shenzhen, Shenzhen, China
[3]SCSE, Wuhan University, Wuhan, China
[4]T Lab, Tencent, Beijing, China
{yimingyang@link., lizhen@}cuhk.edu.cn

Abstract

Extracting lane topology from perspective views (PV) is crucial for planning and control in autonomous driving. This approach extracts potential drivable trajectories for self-driving vehicles without relying on high-definition (HD) maps. However, the unordered nature and weak long-range perception of the DETR-like framework can result in misaligned segment endpoints and limited topological prediction capabilities. Inspired by the learning of contextual relationships in language models, the connectivity relations in roads can be characterized as explicit topology sequences. In this paper, we introduce Topo2Seq, a novel approach for enhancing topology reasoning via topology sequences learning. The core concept of Topo2Seq is a randomized order prompt-to-sequence learning between lane segment decoder and topology sequence decoder. The dual-decoder branches simultaneously learn the lane topology sequences extracted from the Directed Acyclic Graph (DAG) and the lane graph containing geometric information. Randomized order prompt-to-sequence learning extracts unordered key points from the lane graph predicted by the lane segment decoder, which are then fed into the prompt design of the topology sequence decoder to reconstruct an ordered and complete lane graph. In this way, the lane segment decoder learns powerful long-range perception and accurate topological reasoning from the topology sequence decoder. Notably, topology sequence decoder is only introduced during training and does not affect the inference efficiency. Experimental evaluations on the OpenLane-V2 dataset demonstrate the state-of-the-art performance of Topo2Seq in topology reasoning.

Introduction

In recent years, lane topology reasoning in autonomous driving has gained increasing attention (Li et al. 2023b, 2024; Ma et al. 2024). This is because autonomous driving has traditionally relied on offline HD maps to provide path information. However, road conditions can be uncertain and challenging, and outdated offline HD maps can be disastrous for autonomous vehicles (Liao et al. 2023a). Solely relying on these maps is insufficient to meet the higher demands of advanced autonomous driving, such as L4 and L5 levels.

To address these issues, autonomous vehicles need to perform lane topology reasoning, which involves perceiving the surrounding road flow in real-time from panoramic images and extracting both the geometric positions of the centerlines and their topological relationships. Therefore, lane topology reasoning is crucial for trajectory prediction and planning in end-to-end autonomous driving (Can et al. 2021; Wu et al. 2023; Li et al. 2023a).

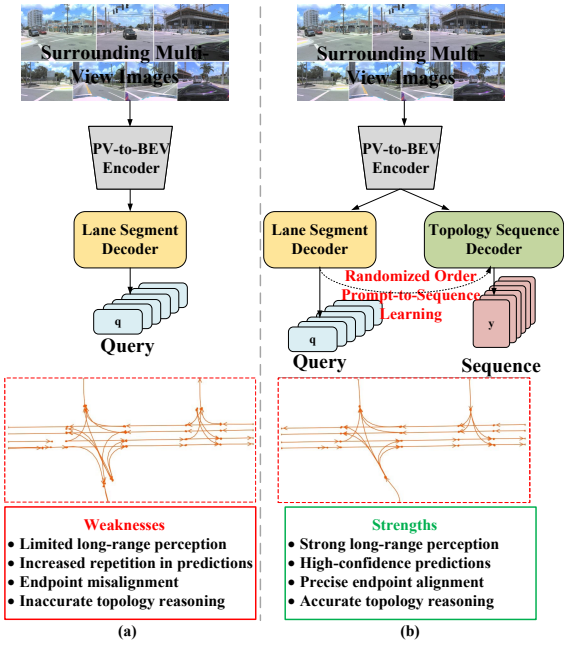

Figure 1: Comparison between previous methods (a) and Topo2Seq (b). Due to the limited sampling positions for each query in deformable-DETR and the unordered detection characteristics, existing methods exhibit several weaknesses. (b) Topo2Seq employs a prompt-to-sequence learning strategy, which enhances lane segment perception and topology reasoning through topology sequence learning.

Early works focused on the characterization of map elements. The semantic map learning approaches assign labels to each pixel to mark map elements (Li et al. 2022a; Liu et al. 2023a,b). However, inconsistent semantic predictions and insufficient instance perception can cause confusion in the planning of autonomous driving systems. More-

*Corresponding author.

over, pixel-by-pixel post-processing often fails in topology extraction and causes time-consuming problems (Liao et al. 2023a). As an upgraded alternative, vectorized map learning approaches employ learnable queries to extract map elements by vectorized lines and polygons in DETR-like framework. These methods offer high detection accuracy and fast inference speeds (Liu et al. 2023b; Liao et al. 2022). Nevertheless, they lack comprehensive trajectory detection in map learning, such as lane direction and connectivity. To address these issues, recent studies on lane topology reasoning have transformed the centerline topology into lane graphs (Li et al. 2023b,a). These end-to-end networks are designed to predict both the line segments, characterized by ordered points, and the topological relationships, represented by an adjacency matrix. However, these methods do not explicitly model the relationships between lanes, instead relying on MLPs to determine the connection probabilities between queries. Due to weak long-range perception and unordered detection characteristics in DETR-like framework (Carion et al. 2020), simple MLPs struggles to effectively learn the connectivity between lanes. As a result, existing methods meet several weaknesses as illustrated from Fig.1 (a). In language models, sequence learning can capture contextual relationships in long texts while maintaining the correct order (Vaswani 2017; Devlin et al. 2018). Inspired by language models, representing lane graph as sequences can explicitly capture the geometric positions and topological relationships of lane. However, in sequence-to-sequence approaches (Peng et al. 2024; Lu et al. 2023), the auto-regressive model depends on prior predictions to generate subsequent outputs, resulting in considerable inefficiencies (approximately 0.1 FPS) (Lu et al. 2023) due to the need for repeated inference.

In this paper, we introduce **Topo2Seq**, a novel approach that enhances topology reasoning via topology sequence learning. Topo2Seq utilizes a dual-decoder architecture, comprising a lane segment decoder and a topology sequence decoder. The topology sequence decoder predicts lane topology sequences extracted from a Directed Acyclic Graph (DAG), while the lane segment decoder extracts lane graphs containing geometric information. Randomized order prompt-to-sequence learning is then employed to extract unordered key points from the lane graph predicted by the lane segment decoder. These key points are input into the prompt design of the topology sequence decoder, enabling the reconstruction of an ordered and complete lane graph. In this way, the lane segment decoder gains powerful long-range perception and accurate topological reasoning from the topology sequence decoder through a shared encoder as illustrated from Fig.1 (b). Notably, the topology sequence decoder is only introduced during training and does not impact inference efficiency.

The contributions of this paper can be summarized as follows:

- We present Topo2Seq, a novel framework with dual-decoder training for enhancing topology reasoning by leveraging topology sequence learning.
- We explicitly model lane graph as sequences to capture long-range geometric positions and topological relationships of lanes.
- We introduce a randomized order prompt-to-sequence learning mechanism, which enables the lane segment decoder to gain robust long-range perception and accurate topology reasoning capabilities from the topology sequence decoder.
- Extensive experiments conducted on the multi-view topology reasoning benchmark OpenLane-V2 (Wang et al. 2024) demonstrate the state-of-the-art performance of Topo2Seq in topology reasoning.

Related Work

Online Map Learning Recent advancements in online map learning focus on detecting map elements using onboard sensors to construct local high-definition (HD) maps. Traditionally, this task has been approached as a pixel-level semantic segmentation problem (Liu et al. 2023a,c). To mitigate the time-consuming post-processing and shape ambiguity issues (Li et al. 2022a), recent approaches have shifted towards learning vectorized representations of map elements. For instance, VectorMapNet (Liu et al. 2023b) introduces a vectorized HD map learning framework that predicts a sparse set of polylines from a bird's-eye view. The MapTR series (Liao et al. 2022, 2023b) leverage hierarchical query embedding to more effectively learn geometrical shapes at both point and shape levels. To further enhance positional embedding in queries, MapQR (Liu et al. 2024) introduces a scatter-and-gather query mechanism that encodes shared content across different positions. However, current online map learning methods still fall short in providing detailed lane information, such as lane direction and topological structure. While these methods are effective at detecting map elements, they are not equipped to deliver the detailed trajectories needed for planning in end-to-end autonomous driving.

Topology Reasoning Topology Reasoning primarily focuses on centerline perception and connectivity relations. STSU (Can et al. 2021) is the first end-to-end framework to detect centerlines and objects using learnable queries from a BEV perspective. The centerline queries are processed by detection, control, and association heads to generate a lane graph, which is widely followed in most subsequent works. TopoNet (Li et al. 2023a) represents lane connectivity as a lane graph and designs a scene graph neural network to refine the position and shape of lanes. LaneSegNet (Li et al. 2023b) introduces lane attention and identical initialization to enhance long-range perception. RoadPainter (Ma et al. 2024) generates centerline masks to better refine points with large curvature. To fully leverage higher recall 2D results, Topo2D (Li et al. 2024) updates 3D lane queries using 2D lane priors. TopoMLP (Wu et al. 2023) enhances topology results by incorporating lane point coordinates as positional embeddings. Most of these methods are built on a DETR-like framework (Zhu et al. 2020). However, the unordered nature and weak long-range perception of the DETR-like framework limit topology reasoning. In this paper, we address these challenges by introducing sequence-to-sequence learning.

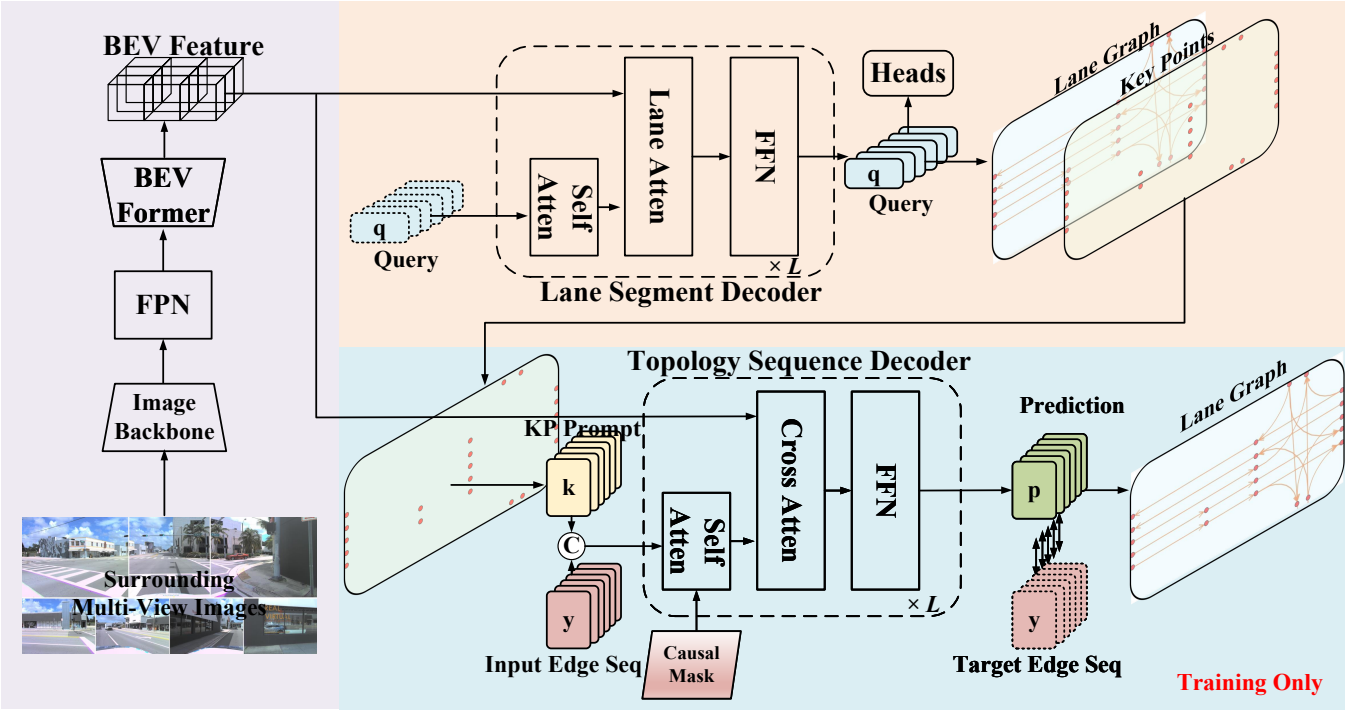

Figure 2: The framework of Topo2Seq. Topo2Seq is composed of three main components. First, the surrounding multi-view images are processed by the image backbone, FPN, and BEVFormer to generate bird's-eye view (BEV) features. The lane segment decoder then predicts the lane graph. The key points from this predicted lane graph are fed into the topology sequence decoder to construct key point prompts, which are subsequently concatenated with edge sequences. The topology sequence decoder infers the relationships between discrete key points and reconstructs them into a coherent lane graph. By doing so, the topology sequence decoder enhances the BEV features with improved long-range dependencies and contextual integration, thereby aiding the lane segment decoder in topology reasoning.

Visual Sequence-to-Sequence Learning Visual sequence-to-sequence learning relies on pixel-based observations, transforming downstream task objectives into a language-like format, i.e., sequences. Pix2Seq (Chen et al. 2021) quantizes and serializes detected objects into sequences of discrete tokens, and by dequantizing the output sequences from the auto-regressive transformer, the detected bounding boxes are obtained. Pix2Seqv2 (Chen et al. 2022) introduces task-specific prompts at the beginning of sequences, enabling training for multiple vision tasks, such as detection, segmentation, keypoint detection, and captioning. However, the aforementioned tasks do not emphasize the order and relationships between instances, which are crucial for topology reasoning. To enable lane graph extraction in a sequence-to-sequence manner, RoadNet (Lu et al. 2023) integrates landmarks, curves, and topology into a unified sequence representation. LaneGraph2Seq (Peng et al. 2024) transforms lane graphs into a combination of vertex and edge sequences. Nevertheless, sequence-to-sequence learning, which relies on auto-regressive transformers, tends to be slow in inference. To fully leverage the strengths of sequence-to-sequence learning for long-range modeling and relationship extraction, we incorporate it into the training phase to enhance feature extraction, rather than using it directly for lane graph inference.

Method

Problem Formulation

Given multi-view images $\mathcal{I}$ captured by a vehicle's surround-view cameras, the goal of Topo2Seq is to perceive centerlines and their topology. Centerlines are represented as a list of ordered 3D points $L = \left[k^0, \cdots, k^{n-1}\right]$, where n is set to 10 and each 3D point is denoted as $k_i = (x, y, z) \in \mathbb{R}^3$. The topology is described by an adjacency matrix $A \in \mathbb{R}^{m \times m}$, where m is the number of detected centerlines. In this matrix, $A_{ij} = 1$ means the endpoint of lane L_i coincides with the starting point of lane L_j. However, it is generally challenging to achieve precise alignment of the starting and end points of connected lane predicted by the network.

Overview

As depicted in Fig. 2, Topo2Seq takes multi-view images as input. The image backbone (He et al. 2016), FPN (Lin et al. 2017a), and BEVFormer (Li et al. 2022b) are utilized to encode these multi-view images into a BEV feature $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$. The lane segment decoder updates learnable query Q^l through self-attention and cross-attention with the BEV feature $\mathcal{F}$. The updated queries are passed to prediction heads for lane segmentation. The key points from the

starting and end points of predicted lane graph are extracted as key point prompts y^K. Ground Truth (GT) lane graph is translated into edge sequences y^E. Edge sequences are concatenated with key point prompts y^K and are fed into topology sequence decoder, where they interact with BEV features $\mathcal{F}$. The prediction sequences are supervised by the target edge sequences and can be dequantized into a lane graph. During testing, the topology sequence decoder does not perform inference. Instead, the lane segment decoder predicts the centerlines and the adjacency matrix.

Lane segment decoder We denote a set of instance-level queries as $\{Q_j^l\}_{j=1}^{N_L}$, where N_L is the preset number of queries, which is usually greater than the number of centerlines in the lane graph. The queries are feed into lane segment decoder to obtain the updated queries:

$$\hat{Q}^l = \textbf{LaneDec}(\mathcal{F}, Q^l) \tag{1}$$

where **LaneDec** denotes Lane segment decoder. Within each lane segment decoder layer, the lane queries are sequentially updated through a self-attention module, a lane attention module (Li et al. 2023b), and a feed-forward network.

Heads We employ MLPs to generate 3D coordinates of lane L and topology A. The topology between lanes are predicted by:

$$\hat{Q}^l_{emb_1}, \hat{Q}^l_{emb_2} = MLP(\hat{Q}^l), MLP(\hat{Q}^l) \tag{2}$$

$$A = Sigmod(MLP(Concat(\hat{Q}^l_{emb_1}, \hat{Q}^l_{emb_2}))) \tag{3}$$

where the MLPs are independent of each other. To provide a more detailed representation of the lane graph, we can predict not only the topology but also the offsets of the left and right lane boundaries, the types of these boundaries, and pedestrian crosswalks.

Topology sequence decoder We follow (Chen et al. 2021) to build topology sequence decoder. Each decoder layer includes a self-attention module, a cross attention module, and a feed-forward network. Auto-regressive property is maintained by causal mask in self-attention module. The whole structure brings several advantages in extracting and refining the BEV features: (1) **Enhanced feature refinement:** The model can selectively focus on relevant areas of the BEV features based on sequence. This targeted attention helps refine BEV features by emphasizing regions critical for accurately reconstructing the lane graph or understanding the scene. (2) **Improved long-range dependencies:**The topology sequence decoder enhances the capture of long-range dependencies between distant key points. This injects the contextual information between key points into the lane segment decoder, enabling it to predict more aligned lane segment endpoints. (3) **Contextual integration:** By focusing on specific key point prompts, the model can reduce the impact of irrelevant or redundant information in the BEV features. This results in more efficient feature extraction and potentially reduces noise in the final predictions. The output from training the topology sequence decoder can be represented as:

$$\hat{y}^E = \textbf{TopoSeqDec}(\mathcal{F}, Concat(y^K, y^E)) \tag{4}$$

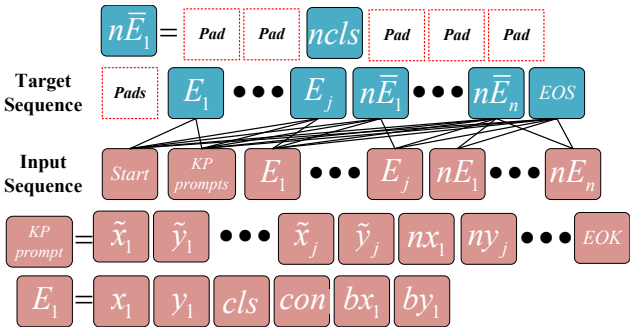

Figure 3: The illustration of input and target sequences.

where **TopoSeqDec** represents the topology sequence decoder, and $\hat{y}^E$ denotes the predicted edge sequence.

Structure of the sequences Following RoadNet (Lu et al. 2023), we transform the Directed Acyclic Graph (DAG) into edge sequences. Each key point in a lane serves as either a starting point or an endpoint, and each edge can be represented by six integers:

$$y^E = [int(x_i), int(y_i), cls, con, int(bx_i), int(by_i)] \tag{5}$$

where the first two integers $int(x_i), int(y_i)$ represent the discretized coordinates of the key point. cls indicates the category of key point, which can be Ancestor, Lineal, Offshoot, or Clone. The con denotes the connection of the key point. If cls is either Ancestor or Lineal, then con is set to 0. Otherwise, con is set to the index of the parent key point. Since a cubic Bezier curve can effectively represent the trajectory of a lane between two key points, the last two integers $int(bx_i), int(by_i)$ indicate the intermediate control points of the Bezier curve. To determine a unique order of key points, we select the location at the right front in the BEV perspective as the starting point and use Depth-First Search to perform the sorting.

During training, we construct two types of sequences as illustrated in Fig. 3: the input sequence and the target sequence used for supervision. The input sequence starts with the $< Start >$ token, followed by the key point prompts, then the GT edges y^E, and the remaining length is filled with noise edges y^{nE} (Chen et al. 2021). Key point prompts y^K include key points for all predicted edges as well as noise edges. Notably, the key points for all predicted edges are unordered and do not correspond to the order of the coordinates in the edge sequence. Finally, the key point prompts conclude with the $< EOK >$ token. In the target sequence, the positions of key point prompts are filled with $< pad >$ tokens, followed by the ground truth edges and noise edges, and ending with $< EOS >$. To help the topology sequence decoder identify which edges are noise edges, the supervised noise edges are marked with the noise class $< ncls >$ at their category positions, while other positions are filled with $< pad >$ tokens. The $< pad >$ tokens are excluded from loss calculation.

Randomized order prompt-to-sequence learning The lanes predicted by the lane segment decoder often have misaligned endpoints. Representing two lanes requires four

Method	Backbone	Epochs	mAP ↑	AP_{ls} ↑	AP_{ped} ↑	TOP_{lsls} ↑
TopoNet (Li et al. 2023a)	ResNet-50	24	23.0	23.9	22.0	-
MapTR (Liao et al. 2022)	ResNet-50	24	27.0	25.9	28.1	-
MapTRv2 (Liao et al. 2023b)	ResNet-50	24	28.5	26.6	30.4	-
LaneSegNet (Li et al. 2023b)	ResNet-50	24	33.4	31.9	34.9	25.4
LaneSegNet (Li et al. 2023b)	ResNet-50	48	36.4	34.9	37.9	27.3
Topo2Seq (ours)	ResNet-50	24	33.6	33.7	33.5	26.9
Topo2Seq (ours)	ResNet-50	48	**37.7**	**36.9**	**38.5**	**29.9**

Table 1: Comparison with the state-of-the-arts on OpenLane-V2 benchmark on lane segment. mAP (%), AP_{ls} (%), AP_{ped} (%), and TOP_{lsls} (%) are reported.

endpoints, which may exhibit geometric inconsistencies between the endpoints of different lanes. In contrast, the edge sequence uses only three points to represent two adjacent lane lines with perfectly aligned endpoints, improving trajectory comprehension for autonomous driving. To leverage the long-range understanding and sequential relationship capabilities of sequence-to-sequence learning, we facilitate interaction between the lane segment decoder and the sequence topology decoder at the key point prompts.

Based on the predictions from the lane segment decoder, we rank the predicted lanes by confidence from highest to lowest and filter out any duplicate key points in each predicted lane using predicted adjacency matrix:

$$y^K \leftarrow \begin{cases} k_j^{n-1} & A_{ij} = 1 \\ k_j^0, k_j^{n-1} & A_{ij} = 0 \end{cases} \tag{6}$$

where the coordinates of key points are discretized. The object of randomized order prompt-to-sequence learning can be expressed as:

$$max \sum_{i=1}^{L} w_i log P(\hat{y}_i^E | , Concat(y^K, y_{<i}^E), \mathcal{F}) \tag{7}$$

where w_i denotes the class weight, $y_{<i}^E$ indicates all tokens before y_i^E, and $\hat{y}^E$ is predicted target sequence. The input key point prompts are unordered relative to the edge sequence, enabling the sequence topology decoder to guide the network in inferring relationships between discrete key points. In this way, the network infers the correct associations among unordered key point prompts, compelling it to focus on long-range relationships. Additionally, this process encourages the network to refine the positions of high-confidence key points, reducing duplicate predictions and aligning endpoints in BEV domain. By enhancing interaction between two decoders, this approach indirectly addresses the limitations in capturing long-range relationships via sequence-to-sequence learning.

Loss function The overall loss function in Topo2Seq is defined as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_{cls} + \alpha_3 \mathcal{L}_{seg} + \alpha_4 \mathcal{L}_{lt} + \alpha_5 \mathcal{L}_{top} + \alpha_6 \mathcal{L}_{seq}$$

Where $\mathcal{L}_1$ represents a L1 loss. $\mathcal{L}_{cls}$ denotes a focal loss (Lin et al. 2017b) for lane classification. $\mathcal{L}_{seg}$ includes a cross-entropy loss and a dice loss. $\mathcal{L}_{lt}$ represents a cross-entropy loss for classifying the left and right lane types (e.g., non-visible, solid, dashed). $\mathcal{L}_{top}$ is a focal loss used to supervise the relationship information between the predicted adjacency matrix A and the GT adjacency matrix $\tilde{A}$. $\mathcal{L}_{seq}$ indicates a maximum likelihood loss that supervises the topology sequence decoder in predicting tokens. The weights for each loss are denoted by α_1, α_2, α_3, α_4, α_5, and α_6.

Experiments

Dataset

We evaluate our Topo2Seq model on the OpenLane-V2 dataset (Wang et al. 2024), a recently released open-source dataset specifically designed to focus on topology reasoning in autonomous driving. OpenLane-V2 is derived from Argoverse2 (Wilson et al. 2023) and nuScenes (Caesar et al. 2020) datasets. The data spans various global locations and includes challenging scenarios such as daytime and nighttime, sunny and rainy conditions, as well as urban and suburban environments. In this paper, we primarily evaluate Topo2Seq on the $subset_A$, which consists of 7 surrounding images per frame. The training set includes approximately 27,000 frames, and the validation set contains around 4,800 frames.

Implementation Details

We employ ResNet-50 pre-trained on ImageNet (Deng et al. 2009) as our image backbone. FPN (Lin et al. 2017a) is used to obtain multi-scale features. The BEVFormer (Li et al. 2022b) encodes the multi-scale features of the surround view image into BEV features. Consistent with recent works, the BEV perception range is set to cover the x-axis from $[-50.0m, +50.0m]$ and the y-axis from $[-25.0m, +25.0m]$. The BEV grid is configured to be 200×100. The configuration of the Lane Segment Decoder follows that of LaneSegNet. It consists of 6 layers with the number of queries set to 200. The model achieves a frame rate of 14.7 FPS. Both the input and target sequences consist of 802 tokens, comprising a 201-token key point prompt and a 601-token edge sequence. Due to resource limitations, we train our network on 4 NVIDIA A100 GPUs with a total batch size of 4. We ensure that each sample underwent the same number of iterations with recent works. The initial learning rate is 2×10^{-4} with a cosine annealing schedule during training. AdamW (Kingma and Ba 2015) is adopted as optimizer. The values of α_1, α_2, α_3, α_4, α_5, and α_6 are

Method	Backbone	Epochs	OLS* ↑	DET_l ↑	TOP_{ll} ↑
VectorMapNet (Liu et al. 2023b)	ResNet-50	24	13.8	11.1	2.7
STSU (Can et al. 2021)	ResNet-50	24	14.9	12.7	2.9
MapTR (Liao et al. 2022)	ResNet-50	24	21.0	17.7	5.9
TopoNet (Li et al. 2023a)	ResNet-50	24	30.8	28.6	10.9
Topo2D (Li et al. 2024)	ResNet-50	24	38.2	29.1	26.2
TopoMLP (Wu et al. 2023)	ResNet-50	24	37.4	28.3	21.7
TopoMLP* (Wu et al. 2023)	Swin-B	48	33.5	32.5	11.9
RoadPainter* (Ma et al. 2024)	ResNet-50	24	29.4	30.7	7.9
LaneSegNet (Li et al. 2023b)	ResNet-50	24	40.7	31.1	25.3
LaneSegNet (Li et al. 2023b)	ResNet-50	48	43.3	34.3	27.3
Topo2Seq (ours)	ResNet-50	24	42.7	33.5	27.0
Topo2Seq (ours)	ResNet-50	48	**45.8**	**36.7**	**30.0**

Table 2: Comparison with the state-of-the-arts on OpenLane-V2 benchmark on centerline perception. OLS* (%), DET_l (%), TOP_{ll} (%), and TOP_{lsls} (%) are reported. The OLS* is calculated between DET_l and TOP_{ll}. The centerlines from LaneSegNet and Topo2Seq are extracted from the lane segment results. $*$ denotes metrics from the TOP_{ll} v 1.0.0 version extracted from the referenced paper.

Index	OP	RP	RPL	mAP ↑	AP_{ls} ↑	AP_{ped} ↑	TOP_{lsls} ↑
1	✓			32.5	31.5	33.5	25.1
2		✓		34.9	33.4	36.3	27.8
3			✓	33.8	32.3	35.4	26.5
4		✓	✓	37.7	36.9	38.5	29.9

Table 3: Ablation study on the OpenLane-V2 benchmark: OP, RP, and RPL refer to the ordered GT key points prompts, randomized order GT key points prompts, and randomized order prompt-to-sequence learning, respectively.

set to 0.025, 1.5, 3.0, 0.1, 5.0, and 1.0, respectively. We evaluate three training strategies: (1) Training the network for 24 epochs to achieve stable output using randomly ordered GT key points as key point prompts, followed by 24 epochs focusing on the interaction between the two decoders. (2) Conducting 24 epochs of training solely on the interaction between the two decoders. (3) Training for 12 epochs to achieve stable output using randomly ordered GT key points as key point prompts, followed by 12 epochs of decoder interaction.

Metrics

We evaluate Topo2Seq on two types of tasks: lane segment perception and centerline perception. For lane segment perception, we use mAP, AP_{ls}, AP_{ped}, and TOP_{lsls}. For centerline perception, we use DET_l, TOP_{ll}, and a redefined OLS between DET_l and TOP_{ll}. It is noted that in the OpenLane-V2 benchmark, centerline and lane segment labels are misaligned, and lane segment labels are more challenging to detect due to their higher number of lane pieces.

Main Results

We first compare our Topo2Seq with state-of-the-art methods on OpenLane-V2 benchmark on lane segment. The results on the OpenLane-V2 $subset_A$ are shown in Tab. 1. The results of the state-of-the-art methods are obtained primarily from their respective papers. When trained for 24 epochs (12 epochs for achieving stable output followed by 12 epochs of decoder interaction), Topo2Seq outperforms LaneSegNet by 1.8% in AP_{ls} and 1.5% in TOP_{lsls}. With a two-stage training process over a total of 48 epochs using ResNet-50, Topo2Seq achieves a 37.7% mAP and 29.9% TOP_{lsls}. Under the same configuration, Topo2Seq surpasses LaneSegNet by 2.0% in AP_{ls} and 2.6% in TOP_{lsls}.

The results of centerline perception on the OpenLane-V2 $subset_A$ are displayed in Tab. 2. With the same 24 epochs of training, Topo2Seq outperforms LaneSegNet by 2.0% in OLS*, 2.4% in DET_l, and 1.7% in TOP_{ll}. Compared with TopoMLP and LaneSegNet with the same 48 training epochs, Topo2Seq exceeds TopoMLP by 4.2% improvement in DET_l and performs better than Lanesegnet by 2.5% in OLS*, 2.4% in DET_l, and 2.7% in TOP_{ll}. These results indicate that introducing an additional sequence decoder interaction during training allows the network to achieve considerable improvements in topology reasoning.

Ablation Studies

We have studied each important design in Topo2Seq. The ablation studies are shown in Tab. 3. When introducing ordered GT key point prompts into sequence learning, the network is only capable of learning the trajectories between key points, without being forced to infer the relationships between them. This explains why the results in index 2 outperform those in index 1, with a 2.7% increase in TOP_{lsls}. Comparing the results from index 2 and index 3, it can be seen that due to the inaccuracies and instability in the outputs of the lane segment decoder, interacting with the

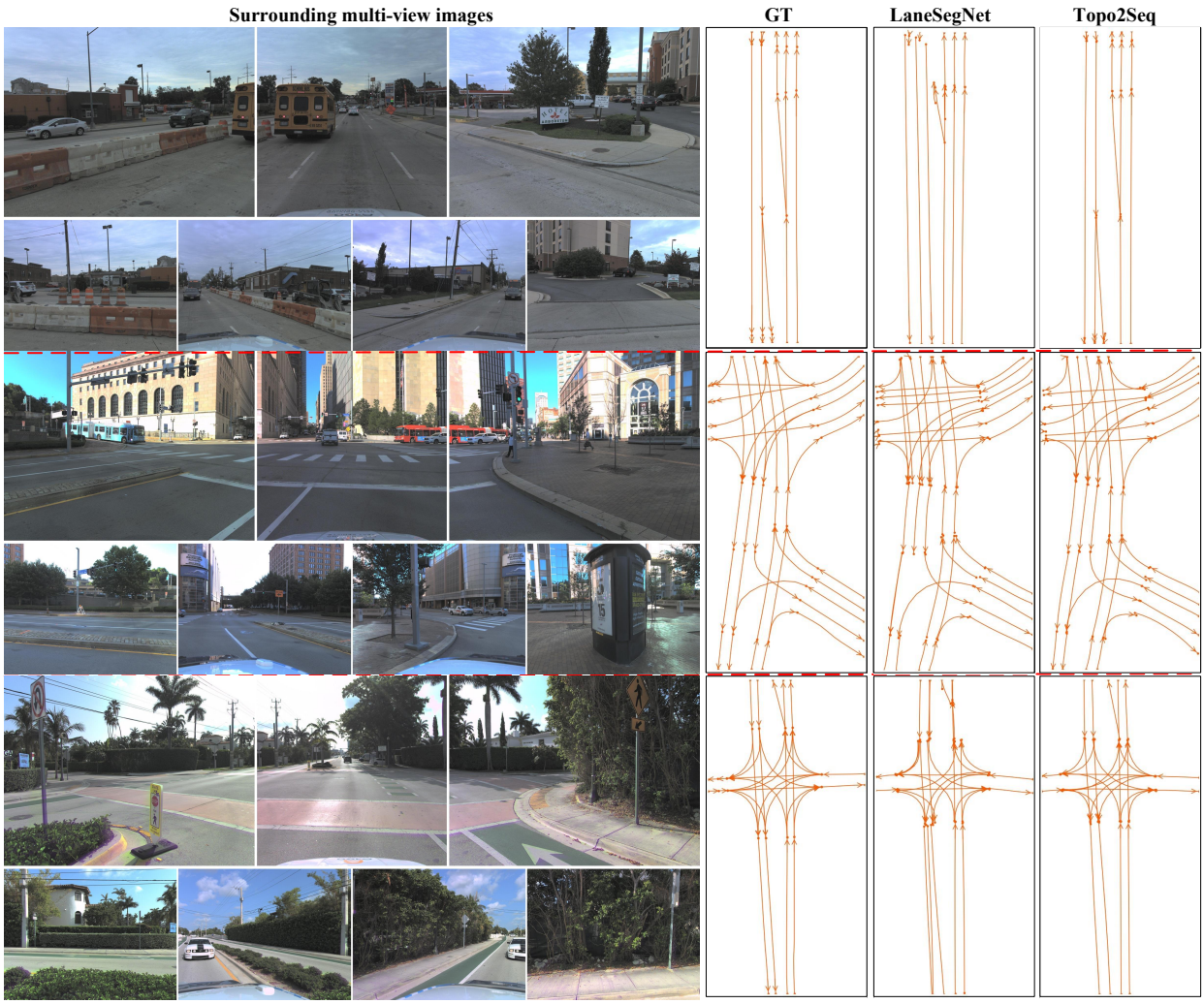

Figure 4: Qualitative results. Compared to LaneSegNet, Topo2Seq, with the assistance of sequence learning, produces higher-quality lane graphs.

topology sequence decoder too early results in worse performance than using randomly ordered GT key points as key point prompts. However, compared to the results in index 1, this approach still leads to a slight improvement in topology reasoning. From the results in index 2 and index 4, it can be seen that when the predicted key points from the lane segment decoder are introduced into the key point prompts for an additional 24 epochs of interaction training between the two decoders, the mAP improves by 2.8%, and the TOP increases by 2.1%. This result indicates that sequence learning can further enhance the extraction of BEV features in the regions of interest for the lane segment decoder, specifically strengthening long-range perception and topology reasoning.

Qualitative Results

As shown in Fig. 4, we visualize the lane graphs generated by LaneSegNet and Topo2Seq. In comparison, Topo2Seq generates higher-quality lane graphs with aligned endpoints, more reliable long-range perception, and accurate topological relationships. This is attributed to the advantages inherited from the interaction with the sequence topology.

Conclusions

We present Topo2Seq, a topology reasoning method enhanced by topology sequence learning. Drawing inspiration from language models, we address the limitations in long-range perception and relationship modeling inherent in DETR-based topology reasoning frameworks through sequence-to-sequence learning. By incorporating randomized order prompt-to-sequence learning, we enhance the interaction between the topology sequence decoder and the lane segment decoder. This approach enables Topo2Seq to generate lane graphs with more accurately aligned endpoints and precise topology. Experimental results on the OpenLane-V2 dataset show that Topo2Seq achieves state-of-the-art performance in topology reasoning.

Acknowledgments

This work was supported by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, CHUK-Shenzhen, by the NSFC 61931024&12326610, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), and by Tencent & Huawei Open Fund.

References

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15661–15670.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; and Hinton, G. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.

Chen, T.; Saxena, S.; Li, L.; Lin, T.-Y.; Fleet, D. J.; and Hinton, G. E. 2022. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35: 31333–31346.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Li, H.; Huang, Z.; Wang, Z.; Rong, W.; Wang, N.; and Liu, S. 2024. Enhancing 3D Lane Detection and Topology Reasoning with 2D Lane Priors. *arXiv preprint arXiv:2406.03105*.

Li, Q.; Wang, Y.; Wang, Y.; and Zhao, H. 2022a. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, 4628–4634. IEEE.

Li, T.; Chen, L.; Wang, H.; Li, Y.; Yang, J.; Geng, X.; Jiang, S.; Wang, Y.; Xu, H.; Xu, C.; et al. 2023a. Graph-based topology reasoning for driving scenes. *arXiv preprint arXiv:2304.05277*.

Li, T.; Jia, P.; Wang, B.; Chen, L.; Jiang, K.; Yan, J.; and Li, H. 2023b. Lanesegnet: Map learning with lane segment perception for autonomous driving. *arXiv preprint arXiv:2312.16108*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liao, B.; Chen, S.; Jiang, B.; Cheng, T.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023a. Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. *arXiv preprint arXiv:2303.08815*.

Liao, B.; Chen, S.; Wang, X.; Cheng, T.; Zhang, Q.; Liu, W.; and Huang, C. 2022. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*.

Liao, B.; Chen, S.; Zhang, Y.; Jiang, B.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023b. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023a. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Liu, Y.; Yuan, T.; Wang, Y.; Wang, Y.; and Zhao, H. 2023b. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, 22352–22369. PMLR.

Liu, Z.; Chen, S.; Guo, X.; Wang, X.; Cheng, T.; Zhu, H.; Zhang, Q.; Liu, W.; and Zhang, Y. 2023c. Vision-based uneven bev representation learning with polar rasterization and surface estimation. In *Conference on Robot Learning*, 437–446. PMLR.

Liu, Z.; Zhang, X.; Liu, G.; Zhao, J.; and Xu, N. 2024. Leveraging Enhanced Queries of Point Sets for Vectorized Map Construction. *arXiv preprint arXiv:2402.17430*.

Lu, J.; Peng, R.; Cai, X.; Xu, H.; Li, H.; Wen, F.; Zhang, W.; and Zhang, L. 2023. Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, 23–33.

Ma, Z.; Liang, S.; Wen, Y.; Lu, W.; and Wan, G. 2024. RoadPainter: Points Are Ideal Navigators for Topology transformER. *arXiv preprint arXiv:2407.15349*.

Peng, R.; Cai, X.; Xu, H.; Lu, J.; Wen, F.; Zhang, W.; and Zhang, L. 2024. LaneGraph2Seq: Lane Topology Extraction with Language Model via Vertex-Edge Encoding and Connectivity Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4497–4505.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, H.; Li, T.; Li, Y.; Chen, L.; Sima, C.; Liu, Z.; Wang, B.; Jia, P.; Wang, Y.; Jiang, S.; et al. 2024. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36.

Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; et al. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*.

Wu, D.; Chang, J.; Jia, F.; Liu, Y.; Wang, T.; and Shen, J. 2023. Topomlp: An simple yet strong pipeline for driving topology reasoning. *arXiv preprint arXiv:2310.06753*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.