

Graph-based Topology Reasoning for Driving Scenes

Tianyu Li*, Li Chen*†, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo and Hongyang Li†

Abstract—Understanding the road genome is essential to realize autonomous driving. This highly intelligent problem contains two aspects - the connection relationship of lanes, and the assignment relationship between lanes and traffic elements, where a comprehensive topology reasoning method is vacant. On one hand, previous map learning techniques struggle in deriving lane connectivity with segmentation or laneline paradigms; or prior lane topology-oriented approaches focus on centerline detection and neglect the interaction modeling. On the other hand, the traffic element to lane assignment problem is limited in the image domain, leaving how to construct the correspondence from two views an unexplored challenge. To address these issues, we present **TopoNet**, the first end-to-end framework capable of abstracting traffic knowledge beyond conventional perception tasks. To capture the driving scene topology, we introduce three key designs: (1) an embedding module to incorporate semantic knowledge from 2D elements into a unified feature space; (2) a curated scene graph neural network to model relationships and enable feature interaction inside the network; (3) instead of transmitting messages arbitrarily, a scene knowledge graph is devised to differentiate prior knowledge from various types of the road genome. We evaluate TopoNet on the challenging scene understanding benchmark, OpenLane-V2, where our approach outperforms all previous works by a great margin on all perceptual and topological metrics. The code is released at <https://github.com/OpenDriveLab/TopoNet>

Index Terms—Autonomous Driving, Lane Perception, Traffic Element Recognition, Topology Reasoning, Graph Neural Network.

1 INTRODUCTION

IMAGINE that an autonomous vehicle is navigating towards a complex intersection and planning to go straight: it is wondering which one of the lanes in front to drive into and which traffic signal to follow. This high-level intellectual problem requires the agent not only to perceive lane position accurately, but also to understand the topology relationship from sensor inputs. Specifically, the driving scene topology includes: (1) the **lane topology graph** comprising centerlines as well as their connectivity, (2) and the **assignment relationship** between lanes and traffic elements (e.g., traffic lights, traffic boards, and road markers). As illustrated in Fig. 1, they altogether build a topological structure that provides explicit navigation signals for downstream tasks such as motion prediction and planning [1, 2].

Conventional autonomous driving datasets [3, 4] include lane topology implicitly in the High-Definition (HD) map, which is designed for mapping but not being learned by neural networks. Various formulations are proposed to serve as substitutes to HD maps, such as 2D and 3D laneline detection [5, 6, 7, 8, 9], bird’s-eye-view (BEV) segmentation [10, 11, 12, 13], and BEV map vectorization [14, 15, 16]. To derive lane connectivity, a “tabula-rasa” resolution is to directly average two neighboring lanelines to get centerlines and then connect them into a lane graph, based on the instance-wise lanelines in vectorization-based methods.

• * equal contribution. † project lead.
 • Primary contact: hy@opendrivelab.com
 • T. Li, L. Chen, H. Wang, Y. Li, J. Yang and H. Li are with OpenDriveLab, Shanghai AI Lab, China. H. Xu and X. Xu are with Huawei. T. Li is also affiliated with Fudan University. L. Chen and P. Luo are with The University of Hong Kong. J. Yan and H. Li are with Shanghai Jiao Tong University, China.

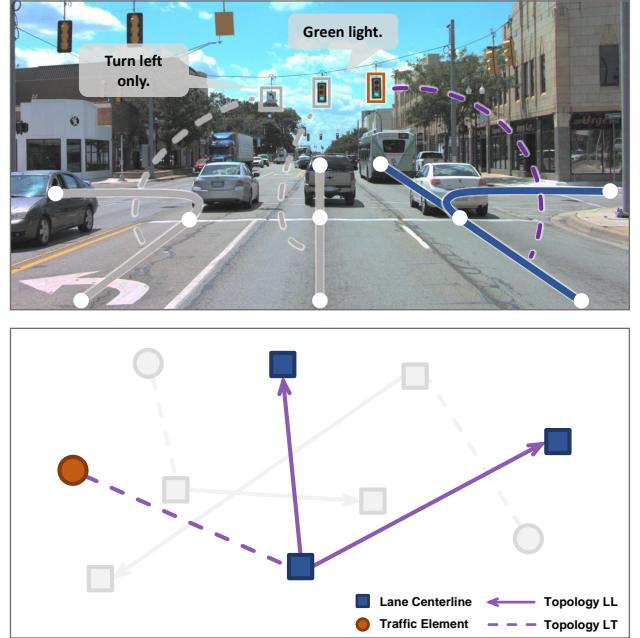


Fig. 1: **Topology relationship of driving scenes.** While driving into an intersection, the self-driving vehicle has to reason about the correct lane and traffic information for downstream navigation. We advocate, and present TopoNet, to directly achieve topology understanding on the heterogeneous graph. “Topology LL” and “Topology LT” represent the relationship among lane centerlines and the relationship between lane centerlines and traffic elements respectively.

Yet, it demands complicated hand-crafted rules and heavy burdens on post-processing. Another approach is to supervise the original perception frameworks with relationship labels. Recent studies, STSU [17] and TPLR [18], employ a Transformer-based architecture to predict lane instances and an additional MLP to output their connectivity. But still, they suffer from overfitting to surrounding background characteristics or difficulty in finding useful information without explicit relationship modeling inside the network.

Moreover, the relationship assignment problem between traffic elements and lanes from sensor inputs remains mostly unexplored. Previous work [19] tried to associate the ground truth of lanelines and traffic lights in the image domain (perspective view, PV). However, integrating traffic elements and lanes in a heterogenous graph (Fig. 1) is a different story. A reason is that traffic elements are described as bounding boxes in PV, while lanes are characterized as curves in 3D or BEV space. Meanwhile, spatial locations remain less important for traffic elements as their semantic meanings are essential, but positional clues of lanes are crucial for autonomous driving vehicles.

To address these issues, we present a Topology Reasoning Network (**TopoNet**), which predicts the driving scene topology in an end-to-end manner. As an attempt to reason about scene topology in a single network, TopoNet comprises two branches with a shared feature extractor, for traffic elements and centerlines respectively. Motivated by the Transformer-based detection algorithms [20, 21], we employ instance queries to extract local features via the deformable attention mechanism, which restricts the attention region and accelerates convergence. Since the clues for locating a specific centerline instance could be encoded in its neighbors and corresponding traffic elements, a Scene Graph Neural Network (SGNN) is devised to transmit messages among instance-level embeddings. Furthermore, we propose a scene knowledge graph to capture prior topological knowledge from entities of different types. Specifically, a series of GNNs are developed based on categories of traffic elements and the centerline connectivity relationship (i.e., predecessor, ego, successor). Updated queries are ultimately decoded as the perception results and driving scene topology. With the proposed designs, we substantiate TopoNet on the large-scale topology reasoning benchmark, OpenLane-V2 [22]. TopoNet outperforms state-of-the-art approaches by 15-84% for centerline perception, and achieves times of performance in terms of the challenging topology reasoning task. Ablations are conducted to verify the effectiveness of our framework.

To sum up, our key contributions are as follows:

- We present TopoNet, which unifies and enhances heterogeneous feature learning via graph-based design. The superiority of TopoNet is demonstrated on the challenging topology reasoning benchmark with a wide margin on perceptual and topological metrics.
- Through the elaborated graph-based modules, the position and shape of lanes are refined by neighboring centerlines and assigned traffic elements. We conduct exhaustive experiments to validate the message-passing mechanism.
- Drawing inspiration from the proposed network and our philosophy, a collection of innovative methods

emerged in the CVPR 2023 Autonomous Driving Challenge¹. The TopoNet and works from participants have made joint contributions to shape the community.

2 RELATED WORK

2.1 Lane Graph Learning

Lane Graph Learning has received abundant attention due to its pivotal role in autonomous driving. Prior works investigate generating road graphs [23, 24] or spatially denser lane graphs [25, 26, 27, 28] from aerial images. However, roads in aerial images are often occluded by trees and buildings, leading to inaccurate results. Recently, there has been a growing focus on producing lane graphs directly from vehicle-mounted sensor data. STSU [17] proposes a DETR-like neural network to detect centerlines and then derive their connectivity by a successive MLP module. Based on STSU, Can et al. [18] introduce additional minimal cycle queries to ensure proper order of overlapping lines. CenterLineDet [13] regards centerlines as vertices and designs a graph-updating model trained by imitation learning. It is also worth noticing that Tesla proposes the “language of lanes” to represent the lane graph as a sentence [29]. The attention-based model recursively predicts lane tokens and their connectivity. In this work, we focus on explicitly modeling the centerline connectivity inside the network to enhance feature learning and indulging traffic elements in constructing the full driving scene graph.

2.2 HD Map Perception

With the trending popularity of BEV perception [16, 30, 31, 32, 33, 34], recent works focus on learning HD Maps with segmentation and vectorized methods. Map segmentation aims at predicting the semantic meaning of each BEV grid, such as lanelines, pedestrian crossings, and drivable areas. These works differentiate from each other mainly in the perspective view to BEV transform module, i.e., IPM-based [35, 36], depth-based [37, 38], or Transformer-based [31, 39]. Though dense segmentation provides pixel-level information, it cannot touch down the complex relationship of overlapping elements. Li et al. [12] handles the problem by grouping and vectorizing the segmented map with complicated post-processings. VectorMapNet [14] proposes to directly represent each map element as a sequence of points, which uses coarse key points to decode laneline locations sequentially. MapTR [15] further explores a unified permutation-based modeling approach for the sequence of points to eliminate the modeling ambiguity and improve performance and efficiency. In fact, since vectorization also enriches the direction information for lanelines, vectorization-based methods could be easily adapted to centerline perception by alternating the supervision. Recently, InstaGram [40] constructs map elements as a graph by predicting vertices first and then utilizing a GNN module to detect edges. Its GNN produces all vertex features simultaneously, leading to the lack of instance-level interaction. Contrary to the aforementioned approaches, we leverage instance-wise feature transmission with a graph neural network, to extract prominent prediction hints from other elements in the topology graph.

1. <https://opendrivelab.com/AD23Challenge.html>

2.3 Driving Scene Understanding

Driving Scene Understanding mainly indicates summarizing positional relationships of elements in outdoor environments beyond perception [41, 42, 43, 44]. Previous works focus on utilizing the relationships of 2D bounding boxes for motion prediction [43, 45, 46, 47] and risk assessment [48, 49]. In the industrial context, Mobileye presents an optimization-based method to automatically construct lane topology and traffic light-to-lane relationships based on their internal data [50]. In the academy, Langenberg et al. [19] address the traffic light to lane assignment (TL2LA) problem with a convolutional network by taking heterogeneous metadata as additional inputs. In contrast, TopoNet takes RGB images only and additionally reasons about the topology for lane entities besides TL2LA. We instantiate TopoNet on the large-scale driving scene understanding benchmark, which covers complicated urban scenarios.

2.4 Graph Neural Network

Graph Neural Network and its variants, such as graph convolutional network (GCN) [51], GraphSAGE [52], and GAT [53], are widely adopted to aggregate features of vertices and extract information from graph data [54]. Witnessing the impressive achievements of GNN in various fields (e.g., recommendation system and video understanding) [55, 56, 57, 58], researchers in the autonomous driving community also attempt to utilize it to process unstructured data. Weng et al. [59, 60] introduce GNN to capture interactions among agent features for 3D multi-object tracking. LaneGCN [61] constructs a lane graph from HD map, while others [47, 62, 63] model the relationship of moving agents and lanelines as a graph to improve the trajectory forecasting performance. Inspired by prior works, we design a GNN for the driving scene understanding task to enhance feature interaction and introduce a class-specific knowledge graph to better incorporate semantic information.

3 TOPONET

3.1 Problem Formulation

Given images from multi-view cameras mounted on a vehicle, the goal of TopoNet lies in two perspectives - perceiving entities and reasoning their relationships. As an instance-level representation is preferable for topology reasoning, a directed centerline is described as an ordered list of points. We denote it as $v_l = [p_0, \dots, p_{n-1}]$, where $p = (x, y, z) \in \mathbb{R}^3$ describes a point's coordinate in 3D space, p_0 and p_{n-1} are the starting and ending point respectively. Traffic elements are represented as 2D bounding boxes in different classes on the front-view images. All existing lanes V_l and traffic elements V_t within a predefined range are required to be detected.

On the perceived entities, the topology relationships are built. The connectivity of directed lanes establishes a map-like network on which vehicles can drive and is denoted as the lane graph (V_l, E_{ll}) , where the edge set $E_{ll} \subseteq V_l \times V_l$ is asymmetric. An entry (i, j) in E_{ll} is positive if and only if the ending point of the lane v_i is connected to the starting point of v_j . The graph $(V_l \cup V_t, E_{lt})$ describes

the correspondence between lanes and traffic elements. It can be seen as a bipartite graph that positive edges only exist between V_l and V_t . Both edge sets are required to be predicted in the task of topology reasoning.

3.2 Overview

Fig. 2 illustrates the overall architecture of the proposed **TopoNet**. Given multi-view images as input, the feature extractor generates multi-scale image features, including the front-view feature F_{PV} , and then convert them into a BEV feature F_{BEV} through a view transform module. Two independent decoders with the same deformable attention architecture [21] consume F_{PV} and F_{BEV} to produce instance-level embeddings Q_t and Q_l separately. The proposed **Scene Graph Neural Network (SGNN)** then refines centerline queries Q_l in positional and topological aspects. Note that the decoder and SGNN layers are stacked iteratively to obtain local and global features in a sequential fashion. Finally, the task-specific heads take the refined queries to produce prediction results. Next, we elaborate on the proposed SGNN module.

3.3 Scene Graph Neural Network

A representative embedding (or query) could provide ideal instance-wise detection or segmentation results, as discussed in conventional perception works [20, 64]. Nevertheless, being discriminative is not enough to recognize correct topology relationships. The reason is that it takes a pair of instance queries as input to determine their relationship, in which feature embeddings are actually not independent. Meanwhile, adopting the local feature aggregation scheme of point-wise queries [14, 15] for centerline perception is inadequate. Specifically, a key difference between centerlines and physical map elements is that centerlines naturally encode lane topology and traffic rules, which cannot be inferred from local features alone. Therefore, we aim to simultaneously acquire perception and reasoning results by modeling not only discriminative instance-level representations but also inter-entity relationships.

To this end, we present SGNN, which has several designs and merits compared to previous works. (1) It adopts an embedding network to extract TE knowledge within a unified feature space. (2) It models all entities in a frame as vertices in a graph, and strengthens interconnection among perceived instances to learn their inherent relationships with a graph neural network. (3) Alongside the graph structure, SGNN incorporates prior topology knowledge with a scene knowledge graph. We detailedly introduce each part below.

3.3.1 Embedding Network

As traffic elements are labeled on the perspective view, it is hard to harness their positional features in the BEV space. However, their semantic meaning imposes a great effect that could not be neglected. For instance, a road sign indicating the prohibition of left turn usually corresponds to lanes that lay in the middle of the road. This predefined knowledge is beneficial for locating corresponding lanes. Towards this, we introduce an embedding network to extract semantic

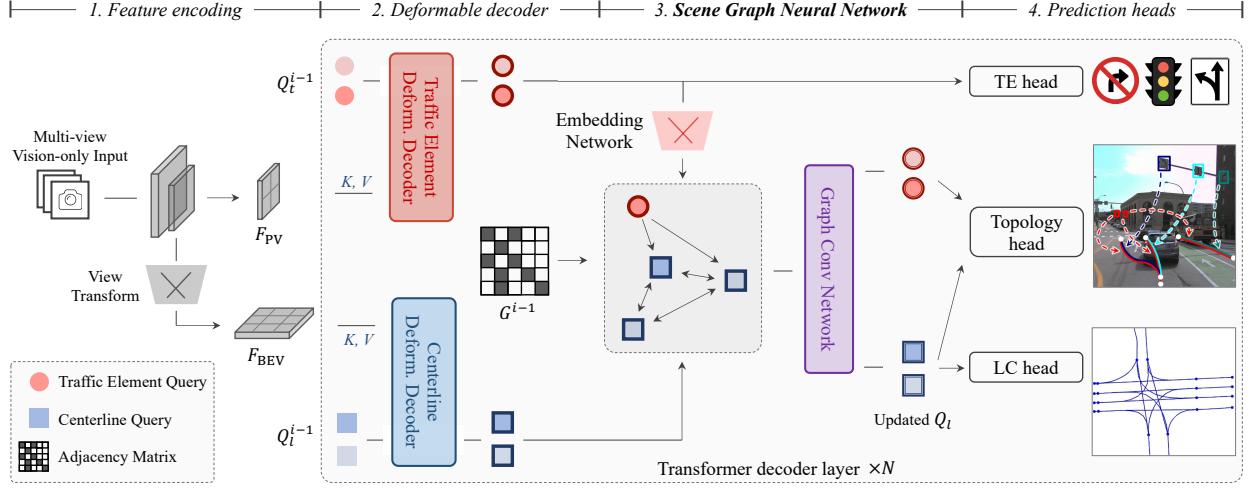


Fig. 2: **Systematic diagram of TopoNet.** TopoNet addresses the crucial problem of topology reasoning for driving scenes in an end-to-end fashion. It consists of four stages, with the last three being compacted in a Transformer decoder architecture. TopoNet handles traffic elements and centerlines as two parallel branches at the Deformable decoder stage. Various types of instance queries (red, blue) then interact, exchange messages, acquire and aggregate prominent knowledge in the proposed Scene Graph Neural Network stage. The explicit relationship modeling inside the network serves as a favorable scheme for feature learning and topology prediction. We abbreviate traffic elements and lane centerlines as “TE” and “LC” in this paper, respectively.

information and transform it into a unified feature space to match with centerlines:

$$\tilde{Q}_t^i = \text{embedding}^i(Q_t^i), \quad (1)$$

where i denotes the i -th decoder layer. Note that the queries \tilde{Q}_t^i remain intact in the SGNN. This is intended since imagining traffic elements from centerlines is relatively challenging, and we empirically find that updating Q_t with too many feature interactions places negative effects on predicting their attributes.

3.3.2 Feature Propagation in GNN

In this part, we introduce how topological relationships are modeled and how knowledge from different queries is exchanged. Regarding the concept of relationship, GNN is a natural choice. Relations can be conveniently formulated as edges in a graph where entities are seen as vertices, while it is nontrivial in an open world without any explicit constraint. As there is no prior knowledge of topology structure, a trivial way is to construct a fully connected graph (V, E) , where $V = V_l \cup V_t$ and $E \subseteq V \times V$. This inevitably increases computational cost and introduces unnecessary information transmission, such as between two traffic elements that are placed subjectively by humans. Instead, we form two directed graphs to propagate features, namely $G_{ll} = (V_l, V_l \times V_l)$ for lane graph estimation and $G_{lt} = (V_l \cup V_t, V_l \times V_t)$ representing the TE to LC assignments.

In graph G_{ll} and G_{lt} , lane queries Q_l are refined by the connected neighbors and corresponding traffic elements. Due to the fact that Q_l and Q_t represent different objects, the semantic gap still exists. We introduce an adapter layer to combine this heterogeneous information into the informa-

tion gain denoted as R . The overall process in an SGNN layer can be formulated as follows:

$$\begin{aligned} Q_l^{i'} &= \text{SGNN}_{ll}^i(Q_l^i, G_{ll}^{i-1}), \\ Q_l^{i''} &= \text{SGNN}_{lt}^i(Q_l^i, \tilde{Q}_t^i, G_{lt}^{i-1}), \\ R^i &= \text{downsample}^i(\text{ReLU}(\text{concat}(Q_l^{i'}, Q_l^{i''}))), \\ \tilde{Q}_l^i &= Q_l^i + R^i. \end{aligned} \quad (2)$$

Next, we detail how the topological weights in SGNN are constructed and learned, based on a naive scene graph structure and an elaborately designed knowledge graph.

3.3.3 Vanilla Scene Graph

Given the adjacency matrix A_{ll}^{i-1} , which is a representation of neighboring relationships in graph G_{ll}^{i-1} from the previous layer, we construct a weight matrix T_{ll}^i to control the flow of messages in the graph. In the directed graph, messages are passed in a single direction, e.g., from a centerline to its successor. However, as the structure of lanes depends on each other, the position of a lane is a good indication of the locations of its neighbors. Thus, we supplement A_{ll}^{i-1} with a backward adjacency matrix to allow message exchange for two connected centerlines. The matrix T_{ll}^i of the i -th layer is calculated by:

$$\begin{aligned} T_{ll}^0 &= I, \\ T_{ll}^i &= \beta_{ll} \cdot (A_{ll}^{i-1} + \text{transpose}(A_{ll}^{i-1})) + I, \end{aligned} \quad (3)$$

where I denotes the identical mapping for self-loop, β_{ll} is a hyperparameter to control the ratio of features propagated between nodes.

In the bipartite graph G_{lt} , where only the correspondence between lanes and traffic elements is presented, we

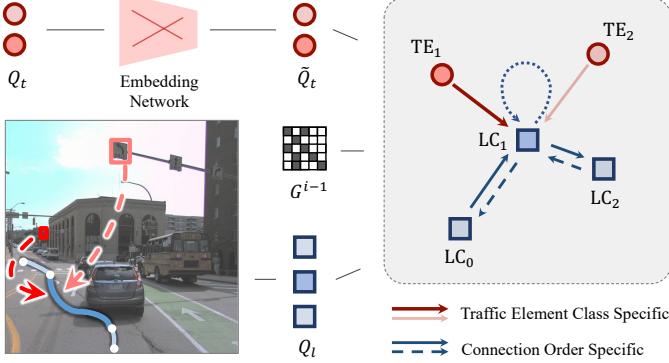


Fig. 3: **Scene knowledge graph** illustration. For the centerline colored blue in the left case, related weight matrices in the graph are categorically independent. Different traffic elements and lane-directed connections bring different information to the centerline, which is encoded as a scene knowledge graph on the right.

utilize features of traffic elements to refine centerline embeddings as follows:

$$\begin{aligned} T_{lt}^0 &= O, \\ T_{lt}^i &= \beta_{lt} \cdot A_{lt}^{i-1}, \end{aligned} \quad (4)$$

where O is a matrix in which all entries are zero.

After obtaining the weight matrices, SGNN utilizes the graph convolutional layer (GCN) [51] to perform feature propagation among queries:

$$\begin{aligned} Q_l^{i'} &= \text{GCN}_{ll}^i(Q_l^i, T_{ll}^i), \\ Q_l^{i''} &= \text{GCN}_{lt}^i(Q_l^i, \tilde{Q}_t^i, T_{lt}^i). \end{aligned} \quad (5)$$

3.3.4 Scene Knowledge Graph

Though GCN enables feature propagation in the built graphs and treats nodes differently based on their connectivity, the semantic meaning of vertices remains unexplored. For example, the information from a traffic element indicating to go straight is not equally important to a red light. To address the issue and incorporate categorical prior, we design the scene knowledge graph to treat vertices in different classes differently. Fig. 3 illustrates an example process of updating a centerline query LC_1 on the given knowledge graph.

On the graph G_{lt} , we use $\mathbf{W}_{lt}^i \in \mathbb{R}^{|C_t| \times F_l \times F_t}$ to denote the learnable weights, where C_t describes the attribute set of traffic elements, F_l and F_t are the number of feature channel of LC and TE queries respectively. A centerline query with index x aggregates information from its corresponding traffic elements based on their classification scores:

$$\begin{aligned} K_{lt}^i &= A_{lt}^{i-1}, \\ Q_{l(x)}^{i''} &= \sum_{\forall y \in N(x)} \sum_{\forall c_t \in C_t} \beta_{lt} \cdot S_t^i K_{lt(x,y)}^i \mathbf{W}_{lt(c_t)}^i \tilde{Q}_{t(y)}^i, \end{aligned} \quad (6)$$

where $N(x)$ outputs the indices of all neighbors of the vertex with index x , and $S_t^i \in \mathbb{R}^{|C_t| \times |Q_t^i|}$ represents the classification scores of traffic element queries.

Although all centerlines fall into the same category, the directed connection nature, namely predecessor and

successor, still poses an impact on the process of feature propagation. To this end, we formulate the learnable weight matrix for the lane graph as $\mathbf{W}_{ll}^i \in \mathbb{R}^{|C_l| \times F_l \times F_l}$, where $C_l = \{\text{successor, predecessor, self-loop}\}$. The centerline queries are further updated by:

$$\begin{aligned} K_{ll}^i &= \text{stack}(A_{ll}^{i-1}, \text{transpose}(A_{ll}^{i-1}), I), \\ Q_{l(x)}^{i'} &= \sum_{\forall y \in N(x)} \sum_{\forall c_l \in C_l} \beta_{ll} \cdot K_{ll(c_l,x,y)}^i \mathbf{W}_{ll(c_l)}^i Q_{l(y)}^i. \end{aligned} \quad (7)$$

3.4 Learning

We employ multiple losses to train TopoNet in an end-to-end manner. As depicted in Fig. 2, all heads consume queries to provide perception and reasoning results. Nevertheless, they are not entirely independent, as the topology head requires matching results from perception heads. Similar to Transformer-based networks [20, 21], the supervision is applied on each decoder layer to optimize the query feature iteratively. The overall loss of the proposed model is:

$$\mathcal{L} = \mathcal{L}_{det_{TE}} + \mathcal{L}_{det_{LC}} + \mathcal{L}_{top}. \quad (8)$$

3.4.1 Perception

Following the head design in DETR [20], the TE head predicts 2D bounding boxes with classification scores. Note that for predicting traffic elements, we take Q_t instead of \tilde{Q}_t to preserve their positional information in the perspective view. The LC head produces 11 ordered 3D points and a confidence score from each centerline query $\tilde{Q}_l \in \tilde{Q}_L$. The ground truth of centerlines is normalized based on the predefined BEV range. For both heads, the Hungarian algorithm is utilized to generate matchings between ground truth and predictions, with the matching cost the same as the loss function. Then task-specific losses $\mathcal{L}_{det_{TE}}$ and $\mathcal{L}_{det_{LC}}$ are applied accordingly. Specifically, for the TE head, we employ the Focal loss [65] for classification, an L1 regression loss, and an IOU loss for localization. Meanwhile, for centerlines, we use Focal loss and L1 loss as the classification and regression loss, respectively.

3.4.2 Reasoning

The topology head reasons pairwise relationships on the given embeddings. Similar to STSU [17], for a pair of instances, we use two MLP layers to reduce the feature dimension for each instance. Then the concatenated feature is sent into another MLP with a sigmoid activation to predict their relationship. Based on the matching results from perception heads, the ground truth of each pair of embeddings is assigned. Different from the TE head, we adopt embeddings from the SGNN module, i.e., the refined queries \tilde{Q}_l for lanes and the semantic embeddings \tilde{Q}_t for traffic elements. Due to the sparsity of the graph, there exists a severe imbalance in sample distribution, towards which the Focal loss is deployed in \mathcal{L}_{top} .

4 EXPERIMENTS

In this section, we describe experimental settings in detail. We validate the effectiveness of our design compared to various SOTA methods. Visualizations are also provided.

4.1 Protocols

4.1.1 Implementation Details

Feature Encoding. We adopt a ResNet-50 [66], which is pre-trained on ImageNet [67], with an FPN [68] to obtain multi-scale image features. Following previous works [21, 31], the output features are from stage $S_{8\times}$, $S_{16\times}$ and $S_{32\times}$ of ResNet-50, where the subscripts $n\times$ indicates the down-sampling factor. In the FPN module, the features are transformed into a four-level output with an additional $S_{64\times}$ level. The number of output channels of each level is set to 256. Then we adopt a simplified view transformer with 3 encoder layers proposed in BEVFormer [31]. Note that we do not use temporal information, and thus the temporal self-attention layer in the BEVFormer encoder is replaced by a deformable attention [21] layer. The size of BEV grids is set to 200×100 , with four different height levels of $\{-1.5m, -0.5m, +0.5m, +1.5m\}$ relative to the ground.

Deformable Decoder. For the decoder, we utilize the decoder layer in Deformable DETR [21] that each decoder layer contains three layers: a self-attention layer with 8 attention heads, a deformable attention layer with 8 attention heads and 4 offset points, and a two-layer feed-forward network with 512 channels in the middle. After each operation, a dropout layer with a ratio of 0.1 and a layer normalization is applied. The dimension of initial queries $q = [q_p, q_o] \in Q$ is set to 256, where q_p is utilized to generate the initial reference point, and q_o is the initial object query. Notably, the reference points will remain unchanged across different layers. The query number for centerlines and traffic elements is set to 200 and 100 respectively.

Scene Graph Neural Network. We utilize a simplified version of Graph Convolutional Network (GCN) [51] as our GNN layer. Given an input matrix $Q^i \in \mathbb{R}^{N \times C}$, with N representing the number of nodes and C denoting the number of channels, the output of the operation is:

$$Q^{i'} = \sigma(T^i Q^i \mathbf{W}^i), \quad (9)$$

where $\mathbf{W}^i \in \mathbb{R}^{C \times C}$ is the learnable weight matrix, $T^i \in \mathbb{R}^{N' \times N}$ describes the adjacency matrix with N' output nodes, and $\sigma(\cdot)$ is the activation function. Note that the matrix T is inferred without gradients during training. For the traffic element branch, an embedding network is employed before each GNN layer. The embedding network is a two-layer MLP, in which the output channels are 512 and 256. In between the MLP, a ReLU activation function and a dropout layer are included. β_{il} and β_{lt} are set to 0.6.

Prediction Heads. The prediction head for perception comprises a classification head and a regression head. For the traffic element branch, the classification head is a single-layer MLP, which outputs the sigmoid probability of each class. The regression head is a three-layer MLP with ReLU as the activation function in between, which predicts the normalized coordinates of 2D bounding boxes in the form of $\{cx, cy, width, height\}$. For centerline, the classification head consists of a three-layer MLP with LayerNorm and ReLU in between, which predicts the confidence score. The regression head is a three-layer MLP with ReLU, which predicts the normalized point set of 11×3 for a centerline. To predict topology relationships, relationship heads are

applied. Given the instance queries \tilde{Q}_a and \tilde{Q}_b with 256 feature channels, the topology head first applies a three-layer MLP:

$$\tilde{Q}'_a = \text{MLP}_a(\tilde{Q}_a), \quad \tilde{Q}'_b = \text{MLP}_b(\tilde{Q}_b), \quad (10)$$

where the number of output channels is 128. For each pair of queries $\tilde{q}'_a \in \tilde{Q}'_a$ and $\tilde{q}'_b \in \tilde{Q}'_b$, the output is the confidence of the relationship:

$$\text{conf.} = \text{sigmoid}\left(\text{MLP}_{\text{top}}(\text{concat}(\tilde{q}'_a, \tilde{q}'_b))\right), \quad (11)$$

where the MLP is independent in different types of relationships.

Loss. The $\mathcal{L}_{\text{det}_{\text{TE}}}$ is decomposed into a classification, a regression, and an IoU loss:

$$\mathcal{L}_{\text{det}_{\text{TE}}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \lambda_{\text{iou}} \cdot \mathcal{L}_{\text{iou}}, \quad (12)$$

where λ_{cls} , λ_{reg} , and λ_{iou} is set to 1.0, 2.5, and 1.0 respectively. The classification loss \mathcal{L}_{cls} is a Focal loss. Note that the regression loss \mathcal{L}_{reg} is an L1 Loss calculated on a normalized format of $\{cx, cy, width, height\}$, while the IoU loss \mathcal{L}_{iou} is a GIoU loss computed on the denormalized coordinates. For centerline detection, the $\mathcal{L}_{\text{det}_{\text{LC}}}$ comprises a classification and a regression loss:

$$\mathcal{L}_{\text{det}_{\text{LC}}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}}, \quad (13)$$

where λ_{cls} and λ_{reg} is set to 1.5 and 0.025 respectively. Note that the regression loss is calculated on the denormalized 3D coordinates. For topology reasoning, we adopt the same Focal loss but different weights on different types of relationships:

$$\mathcal{L}_{\text{top}} = \lambda_{\text{top}_{ll}} \cdot \mathcal{L}_{\text{top}_{ll}} + \lambda_{\text{top}_{lt}} \cdot \mathcal{L}_{\text{top}_{lt}}, \quad (14)$$

where both $\lambda_{\text{top}_{ll}}$ and $\lambda_{\text{top}_{lt}}$ are set to 5.0.

Training. 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 . For data augmentation, resizing with a factor of 0.5 and color jitter is used by default. We adopt the AdamW optimizer [69] and a cosine annealing schedule with an initial learning rate of 1×10^{-4} . TopoNet is trained for 24 epochs with a batch size of 8 with 8 NVIDIA Tesla A100 GPUs.

4.1.2 Re-implementation of SOTA Methods

Since there are no prior methods for the task of driving scene understanding, we adapt three state-of-the-art algorithms which are initially designed for lane graph estimation or map learning: STSU [17], VectorMapNet [14], and MapTR [15]. For a fair comparison, we utilize the same image backbone and adopt the same TE head as in TopoNet. As for topology reasoning, we treat it differently based on their own modeling concepts. Specifically, STSU predicts centerlines as Bezier curves and their relationships. Thus most of its designs are retained, while other irrelevant modules are removed. Additionally, we append a topology head for predicting correspondence between centerlines and traffic elements. The goal of VectorMapNet and MapTR is to predict vectorized map elements such as lanelines and pedestrian crossings. VectorMapNet utilizes a DETR-like decoder to estimate key points and then introduces

Data	Method	DET _l ↑	TOP _{ll} ↑	DET _t ↑	TOP _{lt} ↑	OLS↑
<i>subset_A</i>	STSU [17]	12.7	0.5	43.0	<u>15.1</u>	25.4
	VectorMapNet [14]	11.1	0.4	41.7	<u>5.9</u>	20.8
	MapTR [15]	8.3	0.2	<u>43.5</u>	5.9	20.0
	MapTR* [15]	17.7	<u>1.1</u>	<u>43.5</u>	10.4	<u>26.0</u>
	TopoNet (Ours)	28.5	4.1	48.1	20.8	35.6
<i>subset_B</i>	STSU [17]	8.2	0.0	43.9	<u>9.4</u>	21.2
	VectorMapNet [14]	3.5	0.0	49.1	<u>1.4</u>	16.3
	MapTR [15]	8.3	0.1	<u>54.0</u>	3.7	21.1
	MapTR* [15]	<u>15.2</u>	0.5	<u>54.0</u>	6.1	<u>25.2</u>
	TopoNet (Ours)	24.3	2.5	55.0	14.2	33.2

TABLE 1: **Comparison with state-of-the-art methods** on the OpenLane-V2 benchmark. TopoNet outperforms all previous works by a wide margin, especially in directed centerline perception and topology reasoning. *: Topology reasoning evaluation is based on matching results on Chamfer distance. The highest score is bolded, while the second one is underlined.

an auto-regressive module to generate detailed graphical information for a laneline instance. MapTR directly predicts polylines with a fixed number of points using a Transformer decoder. To adapt them to our task, we supervise VectorMapNet and MapTR with centerline labels and use the element queries in their Transformer decoders as the instance queries to produce the driving scene topology.

STSU. The original model utilizes EfficientNet-B0 [70] as the backbone and detects centerlines and vehicles under the monocular setting. It employs a BEV positional embedding and a DETR head to predict three Bezier control points for each centerline, and uses object queries in the decoder to predict the connectivity of centerlines. To adapt to the multi-view inputs, we re-implement STSU by replacing the original backbone with ResNet-50 and aligning the image input resolution with our model. Subsequently, we compute and concatenate the BEV embedding of images from different views. The concatenated embedding is then fed into the DETR encoder. We retain the original DETR decoder to predict the Bezier control points, which are interpolated into 11 equidistant points as outputs. The lane-lane relationship prediction head of STSU is preserved as well. For the newly added traffic element detection branch and lane-traffic element topology head, we adopt the same design as in TopoNet to ensure a fair comparison.

VectorMapNet utilizes a DETR-like decoder to estimate key points and an auto-regressive module to generate detailed graphical information for a laneline instance. It uses a smaller input image size with a shorter edge length of 256. We increase the length to 775 while maintaining the aspect ratio. The backbone setting is aligned with TopoNet. The perception range is defined as $\pm 30m \times \pm 15m$ in the original setting, but we expand it to $\pm 50m \times \pm 25m$. The centerline outputs of VectorMapNet are interpolated to fit our setting during the prediction process. For topology prediction, we use the key point object queries in the VectorMapNet decoder as instance queries of centerlines. We implement the modification on the given codebase of VectorMapNet while retaining other settings. However, due to their lack of support for 3D centerlines, we only predict 2D centerlines in the BEV space and ignore the height dimension.

MapTR. We align the original backbone setting with TopoNet. The perception region is expanded to $\pm 50m \times \pm 25m$.

Method	Topology	DET _l ↑	TOP _{ll} ↑	DET _{l, chamfer} ↑	FPS
STSU [17]	✓	14.2	0.6	13.8	12.8
VectorMapNet [14]	✗	12.7	-	10.3	1.0
MapTR [15]	✗	10.0	-	21.7	11.5
TopoNet	✓	27.7	4.6	27.4	10.1

TABLE 2: **Comparison on centerline perception with a ResNet-50 backbone.** “Topology” denotes that the network is trained with topology supervision.

Method	mIoU↑
HDMapNet [12]	18.3
STSU [17]	31.1
VectorMapNet [14]	25.0
MapTR [15]	35.7
TopoNet (Ours)	39.0

TABLE 3: **Comparison on BEV segmentation.** When rendering centerlines on the BEV grids, TopoNet also outperforms the previous approach.

We adopt MapTR to predict 11 points for each centerline. For topology prediction, we use the average of point queries of an instance in the MapTR decoder as the object query of a centerline. The traffic element head and the topology head are with the same setting as in TopoNet. The implementation is also done on the open-source codebase of MapTR. Due to the lack of support for 3D centerlines, we only predict 2D centerlines in BEV and ignore the height dimension during inference.

4.1.3 Dataset and Metrics

We conduct experiments on the OpenLane-V2 benchmark [22], which covers complex urban scenarios. The dataset contains topological structures in the driving scenes, and raises huge challenges for algorithms to perceive and reason about the environment accurately. Ablation studies are conducted on the *subset_A* of OpenLane-V2.

Dataset. Built on top of the Argoverse 2 [4] and nuScenes [3] datasets, the OpenLane-V2 benchmark includes images from 2,000 scenes collected worldwide under different environments. The dataset is split into two subsets, namely *subset_A* and *subset_B*. Each subset contains 1,000 scenes with multi-view images and annotations at 2Hz. All lanes within $[-50m, +50m]$ along the x-axis and $[-25m, +25m]$ along the y-axis are annotated in the 3D space. Centerlines are described in the form of lists of points. Each list is ordered and comprises 201 points in 3D space. Statistically, about 90% of frames have more than 10 centerlines while about 10% have more than 40. Traffic elements follow the typical labeling style in 2D detection that objects are labeled as 2D bounding boxes on the front-view images. Each element is denoted as a 2D bounding box on the front view image, with its attribute. There are 13 types of attributes, including *unknown*, *red*, *green*, *yellow*, *go_straight*, *turn_left*, *turn_right*, *no_left_turn*, *no_right_turn*, *u_turn*, *no_u_turn*, *slight_left*, and *slight_right*. The topology relationships are provided in the form of adjacency matrices based on the ordering of centerlines and traffic elements. In the adjacency matrices, an entry (i, j) is positive (i.e., 1) if and only if the elements at i and j are connected. Statistically, most of the lanes have one predecessor and successor. The number of connections is up to 7, as a lane leads to many directions

Method	$\text{DET}_l \uparrow$	$\text{TOP}_{ll} \uparrow$	$\text{DET}_t \uparrow$	$\text{TOP}_{lt} \uparrow$	$\text{OLS} \uparrow$
Baseline	25.7	4.0	47.2	20.6	34.6
+ SG	27.7	3.7	48.0	20.1	35.0
+ SKG	28.5	4.1	48.1	20.8	35.6

TABLE 4: **Ablation on the design of scene graph neural network.** “SG” represents the vanilla scene graph, and “SKG” is the enhanced SGNN with the proposed scene knowledge graph.

Method	$\text{DET}_l \uparrow$	$\text{TOP}_{ll} \uparrow$	$\text{DET}_t \uparrow$	$\text{TOP}_{lt} \uparrow$	$\text{OLS} \uparrow$
LL only	27.9	3.8	47.8	20.3	35.1
LT only	27.8	3.9	47.5	20.5	35.1
TopoNet	28.5	4.1	48.1	20.8	35.6

TABLE 5: **Ablation on feature propagation** in the scene knowledge graph. “LL only” denotes that SGNN only aggregates the spatial information from lane-lane connectivity, and “LT only” indicates that only semantic information from the lane-traffic element relationship is included in SGNN.

in a complex intersection. For the correspondence between centerlines and traffic elements, the majority have less than 5 correspondence, while a tiny number of elements have up to 12.

Perception Metrics. The DET score is the typical mean average precision (mAP) for measuring instance-level perception performance. Based on the Fréchet distances [71], the DET_l score is averaged over match thresholds of $\mathbb{T} = \{1.0, 2.0, 3.0\}$:

$$\text{DET}_l = \frac{1}{|\mathbb{T}|} \sum_{t \in \mathbb{T}} AP_t. \quad (15)$$

Note that the defined BEV range is relatively large compared to other lane detection datasets, so accurate perception of lanes in the distance is hard. As a result, thresholds \mathbb{T} are relaxed based on the distance between the lane and the ego car. The DET_t uses IoU as the similarity measure and is averaged over different attributes \mathbb{A} of traffic elements:

$$\text{DET}_t = \frac{1}{|\mathbb{A}|} \sum_{a \in \mathbb{A}} AP_a. \quad (16)$$

Reasoning Metrics. The TOP score is an mAP metric adapted from the graph domain. Specifically, given a ground truth graph $G = (V, E)$ and a predicted one $\hat{G} = (\hat{V}, \hat{E})$, it builds a projection on the vertices such that $V = \hat{V}' \subseteq \hat{V}$, where the Fréchet and IoU distances are utilized for similarity measure among lane centerlines and traffic elements respectively. Inside the predicted \hat{V}' , two vertices are regarded as connected if the confidence of the edge is greater than 0.5. Then the TOP score is the averaged vertex mAP between (V, E) and (\hat{V}', \hat{E}') over all vertices:

$$\text{TOP} = \frac{1}{|V|} \sum_{v \in V} \frac{\sum_{\hat{n}' \in \hat{N}'(v)} P(\hat{n}') \mathbb{1}(\hat{n}' \in N(v))}{|N(v)|}, \quad (17)$$

where $N(v)$ denotes the ordered list of neighbors of vertex v ranked by confidence and $P(v)$ is the precision of the i -th vertex v in the ordered list. The TOP_{ll} is for topology among centerlines on graph (V_l, E_{ll}) , and the TOP_{lt} for topology between lane centerlines and traffic elements on graph $(V_l \cup V_t, E_{lt})$.

Method	$\text{DET}_l \uparrow$	$\text{TOP}_{ll} \uparrow$	$\text{DET}_t \uparrow$	$\text{TOP}_{lt} \uparrow$	$\text{OLS} \uparrow$
w/o embedding	28.4	4.1	46.9	20.5	35.2
TopoNet	28.5	4.1	48.1	20.8	35.6

TABLE 6: **Ablation on traffic element embedding.** TE embedding is necessary to deal with inconsistency in the feature space of different queries.

Overall Metrics. The primary task of the dataset is scene structure perception and reasoning, which requires the model to recognize the dynamic drivable states of lanes in the surrounding environment. The OpenLane-V2 Score (OLS) summarizes metrics covering different aspects of the primary task:

$$\text{OLS} = \frac{1}{4} \left[\text{DET}_l + \text{DET}_t + f(\text{TOP}_{ll}) + f(\text{TOP}_{lt}) \right], \quad (18)$$

where DET and TOP describe performance on perception and reasoning respectively, and f is the square root function.

4.2 Main Results

We compare the proposed TopoNet to several state-of-the-art methods in Table 1. TopoNet outperforms all previous algorithms by a large margin. As the SOTA map learning method MapTR ignores the direction of centerlines with the permutation-equivalent modeling, we additionally evaluate MapTR based on Chamfer distance matching. However, its performance on DET_l , as well as topology metrics, significantly degenerates. The performance of centerline queries without directional information indicates that understanding the complex scenario and perceiving presented instances are two totally different stories. All methods achieve similar DET_t , since we adopt the same traffic element detection branch. However, TopoNet still has a better performance on LC-TE topology reasoning, indicating the necessity of the proposed SGNN module, in which different entities are modeled differently.

Comparison on Centerline Perception. To have a fair comparison, we use a unified backbone architecture and PV-to-BEV transformation module for various SOTA methods on centerline perception task. We keep the topology supervision for STSU, as it is originally designed for detecting centerlines and their topology relationship. Since VectorMapNet and MapTR are for the task of laneline detection where there is no relationship between visible lanelines, we alter the supervision from laneline to centerline and ignore topology supervision to preserve their design choice. As shown in Table 2, TopoNet outperforms other methods, indicating the effectiveness of the model design. To better align with previous works [14, 15], we also provide $\text{DET}_{l,\text{chamfer}}$ with the Chamfer distance as the similarity measure. It does not take the lane direction into account and is thresholded on {0.5, 1.0, 1.5}. Under this metric, TopoNet also shows a prevailing performance compared to other methods. Besides, we measure the runtime of each model on an A100 bare machine. The FPS of TopoNet is 10.1. Compared to other methods on the same machine with aligned input size 512×676 , our method has comparable online efficiency but higher performance.

Comparison on BEV Segmentation. The prediction results of our method are rendered to BEV with a fixed line width of

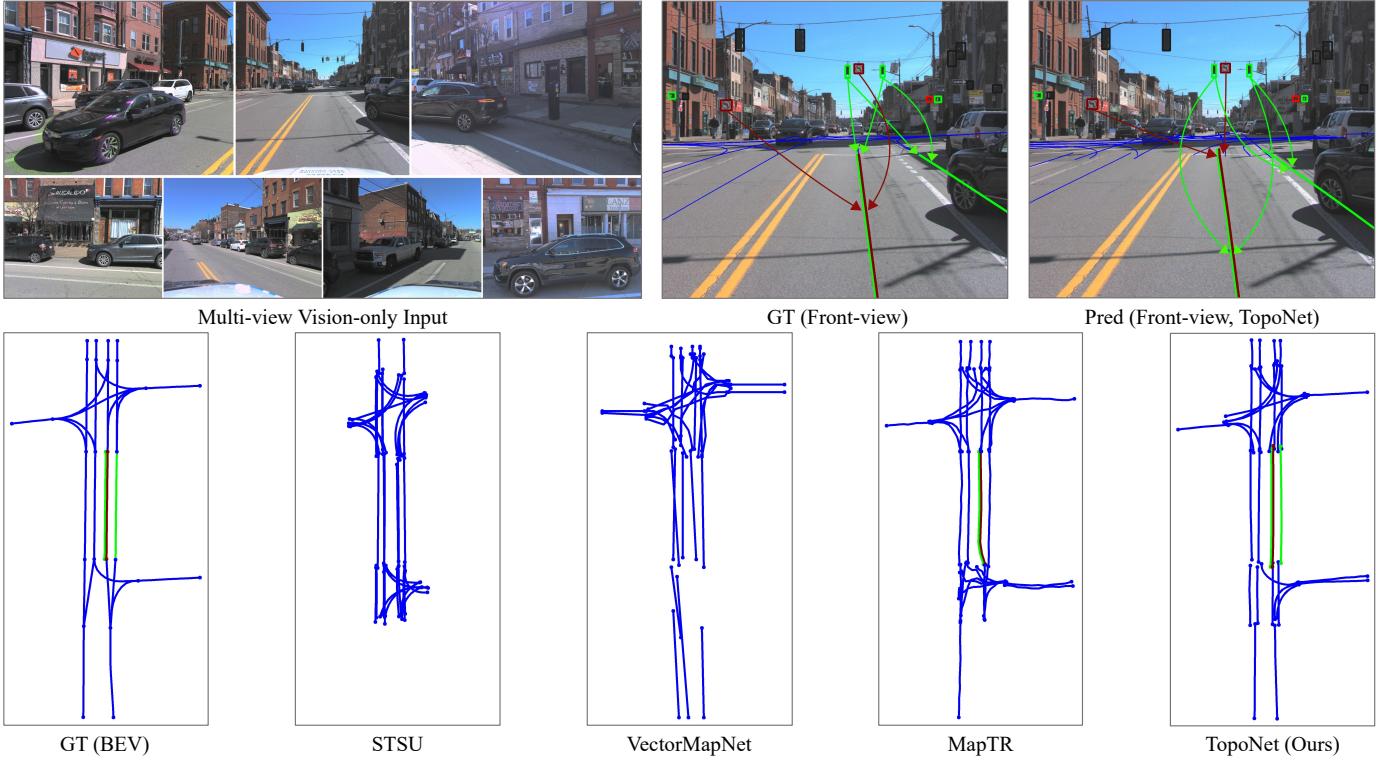
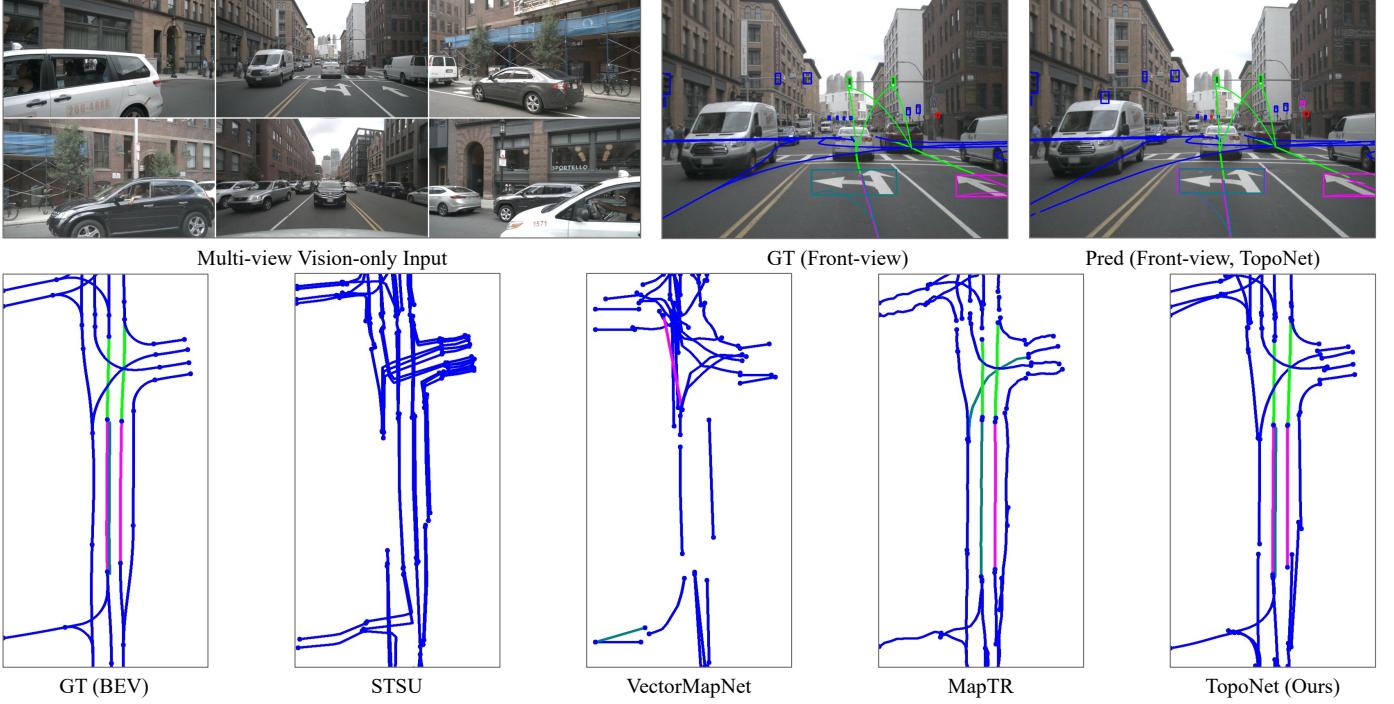
(a) Qualitative results on *subset_A* of the OpenLane-V2 dataset.(b) Qualitative results on *subset_B* of the OpenLane-V2 dataset.

Fig. 4: Qualitative results of TopoNet and other algorithms. While driving in complex scenarios, TopoNet achieves superior lane graph prediction performance compared to other SOTA methods. It also successfully builds all connections between traffic elements and lanes (top right, and correspondingly colored lines in BEV). Colors denote categories of traffic elements.

0.75m aligned with HDMapNet [12], and the mIoU metric is adopted following common segmentation tasks. We reproduce HDMapNet with an enlarged BEV perception range of $\pm 50m \times \pm 25m$ to get aligned with TopoNet. As shown in Table 3, TopoNet significantly surpasses other methods in

terms of mIoU, verifying the advance of our framework.

4.3 Ablation Study

Effect of Design in Scene Graph Neural Network. For ablations on the proposed SGNN, we alternate the proposed

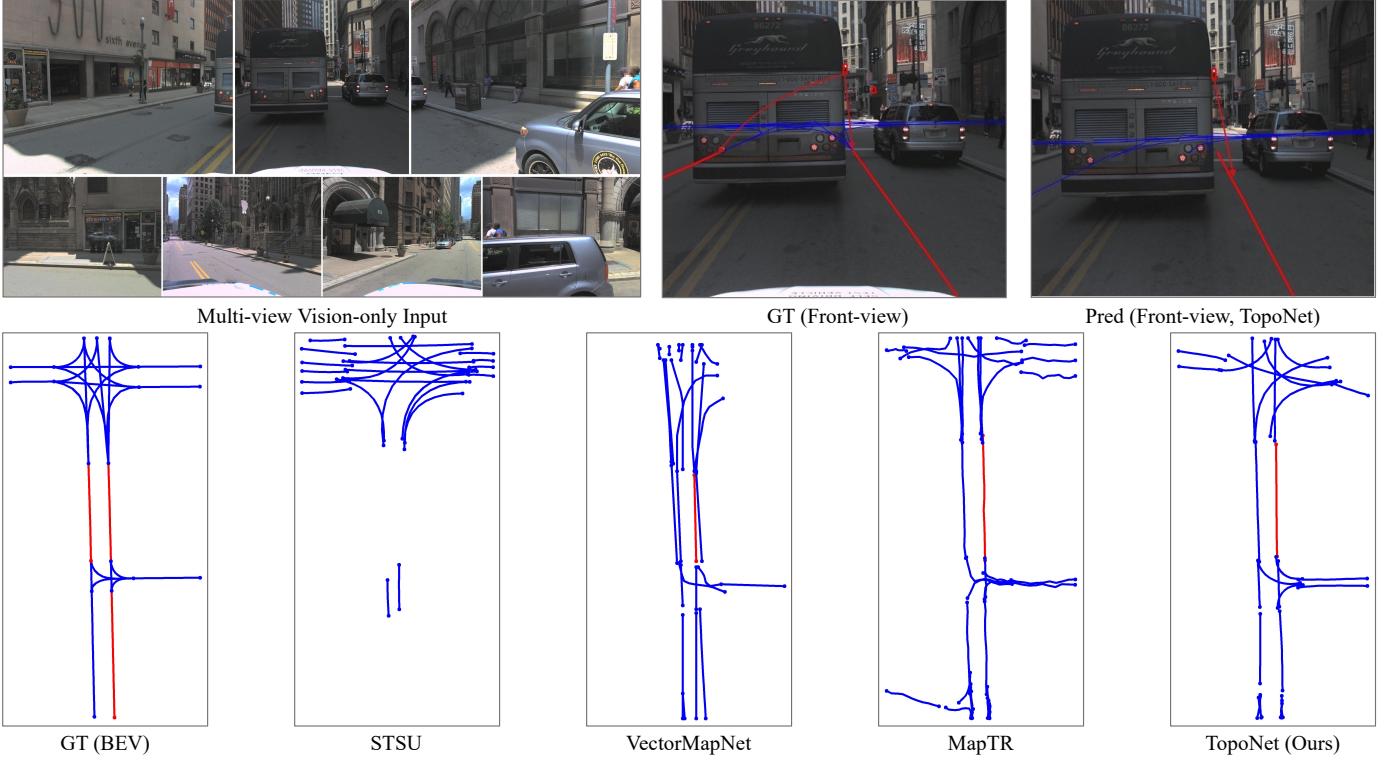


Fig. 5: Failure case under large-area occlusion. TopoNet fails to predict centerlines and the lane graph in the intersection with a large bus colluding in front. Note that the relationship between the left lane and the red light is an incorrect annotation where our algorithm reasons about the direction of the left lane and avoids the false positive prediction.

# GNN	DET _t ↑	TOP _u ↑	DET _t ↑	TOP _{lt} ↑	OLS↑
1	28.5	4.1	48.1	20.8	35.6
2	27.9	4.0	47.5	20.9	35.3
3	20.4	0.5	46.1	15.7	28.3
4	12.2	0.0	48.5	8.2	22.6

TABLE 7: **Ablation on the number of GNN layers** in the scene knowledge graph. Model performance drops as the number of SGNN layers increases.

network into a baseline without feature propagation by downgrading the SGNN module to an MLP and the supervision of topology reasoning only occurs at the final decoder layer. The concatenation and down-sampling operations, as well as the traffic element embedding, are also removed. As illustrated in Table 4, the proposed SKG outperforms models in other settings, demonstrating its effectiveness for topology understanding. Compared to the SG version, the scene knowledge graph provides an additional improvement of 0.8% for centerline perception, owing to the pre-defined semantic prior encoded in the categories of traffic elements. The improvement of traffic element detection and topology reasoning is also consistent.

Effect on Feature Propagation. In the “LL only” setting, we set the β_{lt} parameter to 0. Similar to the baseline, we remove the concatenation and down-sampling operations, as well as the traffic element embedding. For “LT only”, we set the β_{ll} parameter to 0, while other modules remain intact. Results are reported in Table 5. In the “LL only” setting, the drop on TOP_{lt} demonstrates the importance of the graph G_{lt} . With the “LT only” design, DET_t degenerates when removing the

Edge Weight	DET _t ↑	TOP _u ↑	DET _t ↑	TOP _{lt} ↑	OLS↑
0.5	28.4	4.0	47.7	20.8	35.4
0.6	28.5	4.1	48.1	20.8	35.6
0.7	27.3	4.1	47.7	20.7	35.1

TABLE 8: **Ablation on edge weight** in the scene knowledge graph.

graph G_{ll} , showing the importance of feature propagation between centerline queries. These experiments show that both branches are necessary for achieving satisfactory model performance on the primary task.

Effect on Traffic Element Embedding. In the “w/o embedding” setting, we remove the traffic element embedding network and use Q_t as the input of SGNN directly. As shown in Table 6, removing the embedding results in a 1.2% performance drop in traffic element recognition. The reason is that TE queries contain a large amount of spatial information in the PV space due to the 2D detection supervision signals, resulting in significant inconsistencies in the feature spaces. In all, the experiments demonstrate that TE embedding effectively filters out irrelevant spatial information and extracts high-level semantic knowledge to help centerline detection and lane topology reasoning.

Effect on the Number of GNN Layers. Though GNN is beneficial for propagating features in the knowledge graph, raising the number of GNN layers leads to degenerated performance. As shown in Table 7, SGNN with a single GNN layer achieves the best performance. The reason is that a GNN layer increases the similarity of adjacent vertices. When stacking multiple GNN layers, features of all vertices

become less discriminative.

Effect on Edge Weight. Edge weight in the scene knowledge graph represents how much information is propagated through the SGNN layers. In Table 8, 0.6 corresponds to the most appropriate ratio and is used as the default value.

4.4 Qualitative Analysis

We provide a qualitative comparison in Fig. 4. TopoNet predicts most centerlines correctly and constructs a lane graph in BEV. Yet, prior works fail to output all entities or get confused about their connectivity. Fig. 5 shows a case where a bus occludes the intersection in the front view image. TopoNet fails to predict lanes and the topology, especially those in the left half of the crossing. A large-scale dataset and learning techniques, such as active learning, would solve such failure cases in a real-world deployment.

5 CONCLUSION AND FUTURE WORK

In this paper, we discuss abstracting driving scenes as topology relationships and propose the first resolution, namely TopoNet, to address the problem. Importantly, our method models feature interactions via the graph neural network architecture and incorporate traffic knowledge in heterogeneous feature spaces with the knowledge graph-based design. Our experiments on the large-scale OpenLane-V2 benchmark demonstrate that TopoNet excels prior SOTA approaches on perceiving and reasoning about the driving scene topology under complex urban scenarios.

Limitations and Future Work. Due to the query-based design for feature interactions, TopoNet performs well in achieving most positive predictions, while post-processes such as merging or pruning are still needed to produce clean output as in lane topology works [17, 28]. How to incorporate the merging ability with auto-regressive or other association mechanisms deserves future exploration. Meanwhile, it will be interesting to see if more categories of traffic elements, and correspondingly more sophisticated knowledge graphs will make any advances.

LICENSE OF ASSETS

The training data are from the OpenLane-V2 dataset [22], which is under the CC-BY-NC-SA 4.0 license. The evaluation code in the OpenLane-V2 dataset is under the Apache-2.0 license. We integrate the pre-trained ResNet model and part of the BEVFormer [31] in our code, which are under the MIT license and Apache-2.0 license respectively.

REFERENCES

- [1] M. Bansal, A. Krizhevsky, and A. Ogale, “Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst,” *arXiv preprint arXiv:1812.03079*, 2018.
- [2] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, “Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction,” in *CoRL*, 2020.
- [3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020.
- [4] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khan-delwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, “Argoverse 2: Next generation datasets for self-driving perception and forecasting,” in *NIPS*, 2021.
- [5] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, “Spatial as deep: Spatial cnn for traffic scene understanding,” in *AAAI*, 2018.
- [6] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, “Keep your eyes on the lane: Real-time attention-guided lane detection,” in *CVPR*, 2021.
- [7] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, and J. Yan, “PersFormer: 3d lane detection via perspective transformer and the openlane benchmark,” in *ECCV*, 2022.
- [8] N. Garnett, R. Cohen, T. Pe’er, R. Lahav, and D. Levi, “3d-lanenet: End-to-end 3d multiple lane detection,” in *ICCV*, 2019.
- [9] Y. Guo, G. Chen, P. Zhao, W. Zhang, J. Miao, J. Wang, and T. E. Choe, “Gen-lanenet: A generalized and scalable approach for 3d lane detection,” in *ECCV*, 2020.
- [10] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, “Cross-view semantic segmentation for sensing surroundings,” *RA-L*, 2020.
- [11] T. Roddick and R. Cipolla, “Predicting semantic map representations from images using pyramid occupancy networks,” in *CVPR*, 2020.
- [12] Q. Li, Y. Wang, Y. Wang, and H. Zhao, “Hdmapnet: An online hd map construction and evaluation framework,” in *ICRA*, 2022.
- [13] Z. Xu, Y. Liu, Y. Sun, M. Liu, and L. Wang, “Centerlinedet: Road lane centerline graph detection with vehicle-mounted sensors by transformer for high-definition map creation,” in *ICRA*, 2023.
- [14] Y. Liu, Y. Wang, Y. Wang, and H. Zhao, “Vectormapnet: End-to-end vectorized hd map learning,” *arXiv preprint arXiv:2206.08920*, 2022.
- [15] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, “MapTR: Structured modeling and learning for online vectorized HD map construction,” in *ICLR*, 2023.
- [16] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, “Maptrv2: An end-to-end framework for online vectorized hd map construction,” *arXiv preprint arXiv:2308.05736*, 2023.
- [17] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, “Structured bird’s-eye-view traffic scene understanding from onboard images,” in *ICCV*, 2021.
- [18] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, “Topology preserving local road network estimation from single onboard camera image,” in *CVPR*, 2022.
- [19] T. Langenberg, T. Lüdecke, and F. Wörgötter, “Deep metadata fusion for traffic light to lane assignment,” *RA-L*, 2019.
- [20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *ECCV*, 2020.
- [21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” in *ICLR*, 2020.
- [22] OpenLane-V2 Dataset Contributors, “OpenLane-V2: The World’s First Perception and Reasoning Benchmark for Scene Structure in Autonomous Driving,” 2023.
- [23] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi, “Sat2graph: Road graph extraction through graph-tensor encoding,” in *ECCV*, 2020.
- [24] W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, “Spin road mapper: Extracting roads from aerial images via spatial and interaction space graph reasoning for au-

- tonomous driving," in *ICRA*, 2022.
- [25] N. Homayounfar, W.-C. Ma, J. Liang, X. Wu, J. Fan, and R. Urtasun, "Dagmapper: Learning to map by discovering lane topology," in *ICCV*, 2019.
- [26] J. Zürn, J. Vertens, and W. Burgard, "Lane graph estimation for scene understanding in urban driving," *RA-L*, 2021.
- [27] S. He and H. Balakrishnan, "Lane-level street map extraction from aerial imagery," in *WACV*, 2022.
- [28] M. Büchner, J. Zürn, I.-G. Todoran, A. Valada, and W. Burgard, "Learning and aggregating lane graphs for urban automated driving," *arXiv preprint arXiv:2302.06175*, 2023.
- [29] Tesla, "Tesla AI Day." https://www.youtube.com/watch?v=ODSjsvID_SU, 2022.
- [30] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *ECCV*, 2020.
- [31] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.
- [32] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *CVPR*, 2022.
- [33] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *CVPR*, 2023.
- [34] Y. Gao, C. Sima, S. Shi, S. Di, S. Liu, and H. Li, "Sparse dense fusion for 3d object detection," *arXiv preprint arXiv:2304.04179*, 2023.
- [35] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M²BEV: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," *arXiv preprint arXiv:2204.05088*, 2022.
- [36] Y. B. Can, A. Liniger, O. Unal, D. Paudel, and L. Van Gool, "Understanding bird's-eye view of road semantics using an onboard camera," *RA-L*, 2022.
- [37] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *ICRA*, 2023.
- [38] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022.
- [39] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformers," in *AAAI*, 2023.
- [40] J. Shin, F. Rameau, H. Jeong, and D. Kum, "Instagram: Instance-level graph modeling for vectorized hd map learning," *arXiv preprint arXiv:2301.04470*, 2023.
- [41] M. Zipfl and J. M. Zöllner, "Towards traffic scene description: The semantic scene graph," in *ITSC*, 2022.
- [42] Y. Tian, A. Carballo, R. Li, and K. Takeda, "Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles," *arXiv preprint arXiv:2011.13588*, 2020.
- [43] S. Mylavaram, M. Sandhu, P. Vijayan, K. M. Krishna, B. Ravindran, and A. Namboodiri, "Understanding dynamic scenes using graph convolution networks," in *IROS*, 2020.
- [44] A. V. Malawade, S.-Y. Yu, B. Hsu, H. Kaeley, A. Karra, and M. A. Al Faruque, "Roadscene2vec: A tool for extracting and embedding road scene-graphs," *Knowledge-Based Systems*, 2022.
- [45] C. Li, Y. Meng, S. H. Chan, and Y.-T. Chen, "Learning 3d-aware egocentric spatial-temporal interaction via graph convolutional networks," in *ICRA*, 2020.
- [46] S. Mylavaram, M. Sandhu, P. Vijayan, K. M. Krishna, B. Ravindran, and A. Namboodiri, "Towards accurate vehicle behaviour classification with multi-relational graph convolutional networks," in *IV*, 2020.
- [47] J. Fang, C. Zhu, P. Zhang, H. Yu, and J. Xue, "Heterogeneous trajectory forecasting via risk and scene graph learning," *arXiv preprint arXiv:2211.00848*, 2022.
- [48] S.-Y. Yu, A. V. Malawade, D. Muthirayan, P. P. Khargonekar, and M. A. Al Faruque, "Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions," *IEEE T-ITS*, 2021.
- [49] A. V. Malawade, S.-Y. Yu, B. Hsu, D. Muthirayan, P. P. Khargonekar, and M. A. Al Faruque, "Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction," *IoT-J*, 2022.
- [50] Mobileye, "Mobileye under the hood." <https://www.mobileye.com/ces-2022/>, 2022.
- [51] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.
- [52] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017.
- [53] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.
- [54] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE TNNLS*, 2008.
- [55] Z. Guo and H. Wang, "A deep graph neural network-based mechanism for social recommendations," *IEEE TII*, 2020.
- [56] P. Pradhyumna and G. Shreya, "Graph neural network (gnn) in image and video understanding using deep learning for computer vision applications," in *ICESC*, 2021.
- [57] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *PAMI*, 2021.
- [58] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *CVPR*, 2020.
- [59] X. Weng, Y. Wang, Y. Man, and K. M. Kitani, "Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning," in *CVPR*, 2020.
- [60] X. Weng, Y. Yuan, and K. Kitani, "Ptp: Parallelized tracking and prediction with graph neural networks and diversity sampling," *RA-L*, 2021.
- [61] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *ECCV*, 2020.
- [62] X. Jia, P. Wu, L. Chen, H. Li, Y. Liu, and J. Yan, "HDGT: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding," *arXiv preprint arXiv:2205.09753*, 2022.
- [63] X. Jia, L. Chen, P. Wu, J. Zeng, J. Yan, H. Li, and Y. Qiao, "Towards capturing the temporal dynamics for trajectory prediction: a coarse-to-fine approach," in *CoRL*, 2022.
- [64] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "Seqformer: Sequential transformer for video instance segmentation," in *ECCV*, 2022.
- [65] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
- [68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
- [69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.
- [70] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.
- [71] T. Eiter and H. Mannila, "Computing discrete fréchet distance," 1994.