TLSD: Breaking the Limit of Topological Lane Mapping with Graph Knowledge and Distance Awareness

Anh Trong Nguyen anh.nguyence0912@hcmut.edu.vn
Gia Bao Phan bao.phan0805@hcmut.edu.vn
Minh Tri Huynh tri.huynh_tk15nbk@hcmut.edu.vn
Duc Dung Nguyen* nddung@hcmut.edu.vn
AITech Lab., Ho Chi Minh City University of Technology, VNUHCM

Editors: Hung-yi Lee and Tongliang Liu

Abstract

High-Definition (HD) maps are essential for both Advanced Driver-Assistance Systems (ADAS) and autonomous driving. However, offline HD map construction remains costly and challenging to maintain due to the dynamic nature of real-world environments. Consequently, online HD map generation using onboard sensors has become a key area of research. Despite recent advancements, existing deep learning-based methods often provide inaccurate output even using computationally heavy architectures, limiting their practicality for real-world applications. We introduce TLSD, an efficient end-to-end neural network that generates HD maps, incorporating both topological and geometric road information. To enhance both accuracy and efficiency, we introduce four key innovations: (1) an iterative refinement scheme within the decoder to progressively improve map predictions, (2) a group-wise one-to-many assignment strategy that accelerates training convergence, (3) a graph neural network (GNN) module that integrates lane segment coordinates for improved spatial reasoning, and (4) a distance-aware topological post-processing method that enhances the quality of connectivity outputs.

We performed extensive experiments and showed that TLSD achieves a significant improvement in OLUS score compared to existing methods, setting a new state-of-the-art benchmark, producing accurate HDMaps, and a connectivity graph. In particular, TLSD outperforms previous methods on the lane segment perception task (+3.13 in OLUS) and the lane centerline perception task (+3.20 in OLS), demonstrating superior performance in lane-based HD map generation. In addition, we introduce an efficient version, eTLSD, which incorporates a lightweight ResNet-18 backbone and still achieves competitive results, outperforming previous ResNet-50-based methods.

Keywords: HDMaps; 3D lane detection; detection transformer; autonomous driving.

1. Introduction

High-definition (HD) maps offer rich semantic details of driving scenes but are difficult to build and maintain. A reliable HD map system must accurately represent map elements and quickly adapt to real-world changes, yet manual updates are labor-intensive and costly. Hence, constructing HD maps online from onboard sensors is crucial for long-term scalability. While companies like Waymo and Zoox use LiDAR, radar, and cameras for precise 3D mapping, these multi-sensor setups are expensive and pose data fusion challenges. In contrast, cameras provide a cost-effective alternative for scalable HD map construction.

© 2025 A.T. Nguyen, G.B. Phan, M.T. Huynh & D.D. Nguyen.

The rich semantic information captured by cameras makes them well-suited for visual perception tasks. Coupled with advancements in camera-based perception within the research community, online HD map construction using only camera data presents a compelling and sustainable approach. In the autonomous driving community, many perception works Li et al. (2022b),Liao et al. (2023),Li et al. (2023b) transform the perspective view of multi-camera images into a unified Bird-Eye-View (BEV) representation for downstream tasks such as occupancy prediction, 3D object detection, and HD map construction. This provides a versatile 3D representation of the driving scenes thanks to its broad field of view that clearly shows the position of the ego vehicle relative to surrounding driving objects. Perception plays an important role in the driving stacks since it is the first stage and directly influences motion planning. BEV representation is widely adopted because it can seamlessly integrate perception and planning into a unified end-to-end autonomous driving framework. Current approaches to HD-map construction typically decompose the task into two sub-problems: lane detection and topology reasoning Li et al. (2023a, 2024b); Wu et al. (2024). Many aim to address these jointly within a unified learning framework by transforming surrounding-view images into a Bird's-Eye View (BEV) representation. BEV offers a natural and versatile representation of autonomous driving due to its comprehensive field of view. The existing literature on online map learning can be broadly classified into two main streams: map element detection Li et al. (2022a); Liao et al. (2023); Liu et al. (2023) and centerline perception Li et al. (2024a, 2023a). While map element detection focuses primarily on perceiving road geometry, it often neglects the connectivity between map elements. Conversely, centerline perception emphasizes detecting road centerlines but largely disregards geometric details. Therefore, a framework capable of learning both road geometry and topology is needed. To address this, lane segment perception was proposed by Li et al. (2024b) as a means of combining geometric and topological road information using deformable attention Zhu et al. (2021). However, the stacked deformable attention operations in this approach lead to substantial random memory access, creating a bottleneck for edge computing devices. In this work, we contribute a lightweight lane segment perception model designed for efficient edge deployment, achieved through the following improvements:

- We propose an iterative refinement scheme to better guide the transformer decoder in learning lane relationships from the extracted features.
- We introduce a group-wise one-to-many assignment strategy to improve the convergence of the lane segment perception model during training.
- We design a novel module that leverages a graph neural network to incorporate lane segment coordinates into the learning of lane connectivity.
- We acclimate a distance-aware post-processing to enhance topological reasoning.

2. Related Work

2.1. Map Element Detection

With the growing adoption of BEV-based perception, several methods have been proposed for HD map construction. HDMapNet Li et al. (2022a) initiates this trend by segmenting

map elements on a BEV grid, followed by grouping and vectorization with a separate post-processing step. To streamline this process, VectorMapNet Liu et al. (2023) introduces a DETR-based Carion et al. (2020) architecture that enables end-to-end learning of vectorized map elements, eliminating the reliance on post-processing heuristics. Building upon this direction, MapTR Liao et al. (2023) proposes a unified permutation-equivariant formulation to improve the stability of DETR-style training. Its successor, MapTR-V2 Liao et al. (2024), further enhances training with auxiliary one-to-many set prediction and dense supervision applied in both perspective and bird's-eye views. While these approaches significantly advance vectorized map representation, they overlook a crucial aspect of HD maps: **the explicit modeling of connectivity between map elements**. Our work addresses this limitation by jointly learning both geometric structure and topological relationships.

2.2. Topological Reasoning

Structured topological reasoning has gained increasing attention in HD map construction. STSU Can et al. (2021) is one of the earliest attempts to adopt a DETR-like architecture for centerline prediction, employing a fully connected module to infer lane connectivity. However, its reliance on post-processing prevents it from fully exploiting the end-to-end learning capabilities and representational power of deep neural networks. To address this, subsequent approaches have aimed to integrate topological reasoning more directly into the learning process. LaneGAP Liao et al. (2025) formulates lane topology as overlapping paths and applies shortest-path algorithms to recover connectivity. TopoNet Li et al. (2023a) couples Deformable DETR Carion et al. (2020) with graph neural networks (GNNs) Kipf and Welling (2016) to aggregate features across connected lanes, while TopoMLP Wu et al. (2024) leverages PETR Liu et al. (2022) for lane detection and employs a multi-layer perceptron to predict connectivity relationships. Expanding beyond geometric cues, TopoLogic Fu et al. (2024) combines lane-lane geometric distances with semantic similarity to improve relational inference. TopoFormer Lv et al. (2025) introduces a transformer-based framework that constructs a unified traffic scene graph, enabling explicit modeling of inter-lane interactions. Complementary to these methods, SMERF Luo et al. (2024) incorporates SDMap as auxiliary input to improve lane detection, while LaneSegNet Li et al. (2024b) uses a lane attention mechanism to more effectively isolate relevant lane segments. Most recently, Topo2Seq Yang et al. (2025) treats lane topology as a sequence prediction problem, employing a token-based representation of ordered lane points and an auto-regressive decoder for structured connectivity reasoning. In this work, TLSD models the lane-lane interconnection by a graph network, together with location-based positional encoding and a distance-aware post-processing technique.

3. Proposed Method

3.1. Problem Formulation

Given multi-view images captured by the vehicle's surrounding cameras, the goal of TLSD is to predict a structured HD map comprising lane segments and their connectivity graph. Specifically, each lane segment S is defined as a triplet of polylines: $S = \{\mathcal{V}_{\text{left}}, \mathcal{V}_{\text{center}}, \mathcal{V}_{\text{right}}\}$, where each $\mathcal{V}_i$ represents an ordered sequence of 3D points: $\mathcal{V}_i = [l_0, l_1, \ldots, l_{N-1}]$, with

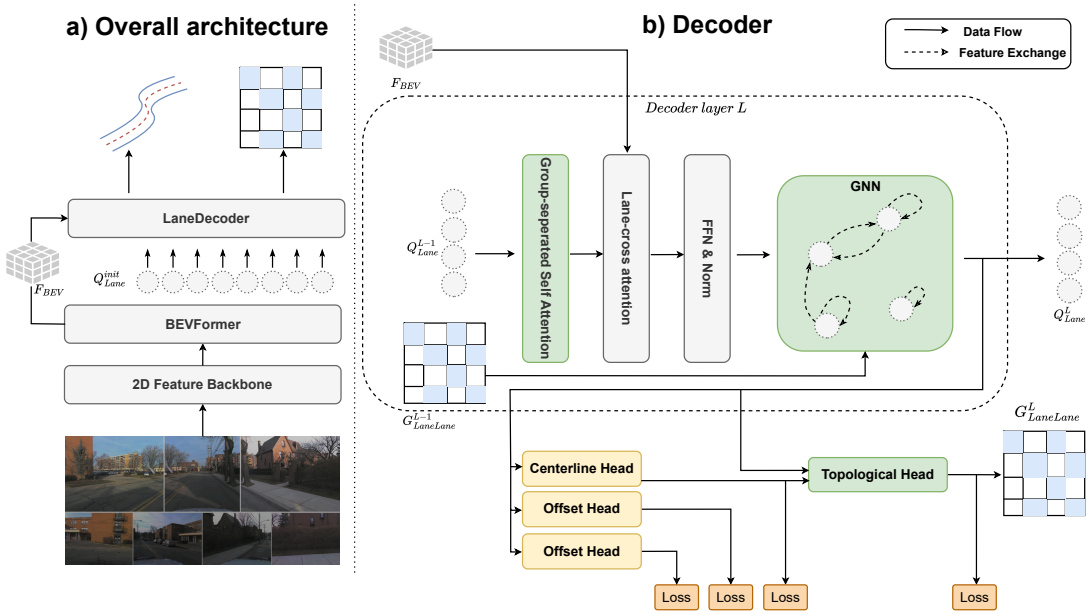

Figure 1: **Overview of the TLSD framework**. TLSD follows a widely adopted three-stage architecture. First, the surrounding multi-view images are processed by a 2D feature extractor to generate multi-scale features. These features are then transformed into bird's-eye view (BEV) representations using BEVFormer. Finally, a transformer-based decoder predicts 3D lane segment coordinates along with a connectivity matrix. To improve convergence and expressiveness, TLSD introduces multiple groups of learnable queries and leverages a graph neural network to capture high-level relational information among potentially connected segments.

$\mathcal{V}_i \in \mathbb{R}^{N\times 3}$. Here, N denotes the number of sampled points along the curve (typically set to 10), and each $l_i \in \mathbb{R}^3$ corresponds to a 3D point expressed in the vehicle-centric coordinate system.

3.2. Overall Architecture

As illustrated in Figure 1, TLSD takes multi-view images captured by vehicle-mounted sensors as input. Multi-scale 2D features are extracted using a standard backbone network (e.g., ResNet He et al. (2016)) combined with a Feature Pyramid Network (e.g., FPN Lin et al. (2017)). These high-dimensional features are then projected into the bird's-eye view (BEV) domain, producing BEV features $\mathbf{F}_{\text{BEV}}$ via BEVFormer Li et al. (2022b).

The BEV features serve as input to a transformer-based lane decoder, which refines a set of learnable lane queries Q_{Lane} through self-attention and cross-attention mechanisms. The refined queries are subsequently passed through prediction heads composed of Multi-Layer Perceptrons (MLPs) to generate 3D lane segment coordinates and their associated topological connectivity.

Our TLSD design introduces several key innovations to the decoder component, including the use of multiple groups of queries, a graph neural network for spatial reasoning, and a distance-aware post-processing module to enhance topological accuracy.

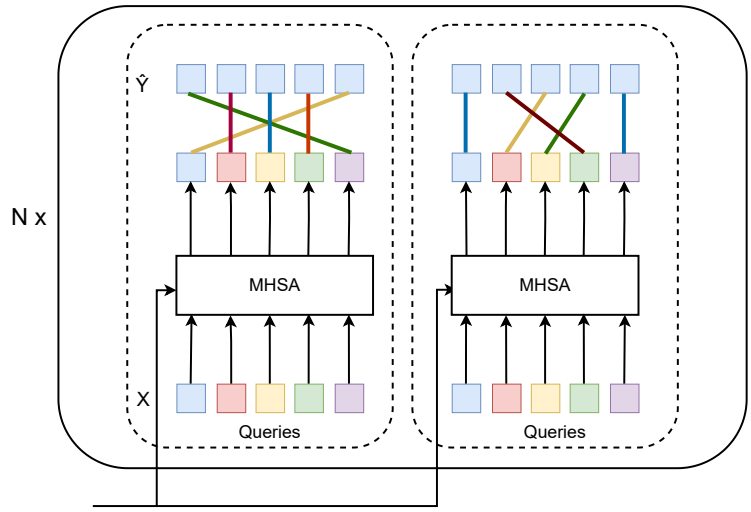

Figure 2: The group-wise one-to-many assignment training strategy.

3.3. Group Query

As previously mentioned, training DETR-based frameworks Carion et al. (2020) involves computing the optimal assignment between model predictions and ground truth prior to loss calculation. The original DETR employs a one-to-one assignment using the Hungarian algorithm, which can lead to slow convergence and prolonged training times. Several works have addressed this issue by incorporating many-to-one assignment strategies into DETR training. For example, MapTRv2 Liao et al. (2024) introduces hierarchical bipartite matching to enhance DETR training for map element detection, along with a new loss function designed to facilitate this matching scheme.

Chen et al. (2023) proposed a simple yet effective method for improving DETR training. They introduced a grouping scheme for object queries, termed group-wise one-to-many assignment. This approach performs one-to-one matching within each query group, allowing a single ground-truth object to be assigned to multiple predictions. Consequently, the prediction closest to the ground-truth object receives a high matching score, while redundant predictions within the same group receive lower scores.

We employ a group-wise one-to-many assignment strategy as follows: N lane segment queries form the primary group, denoted as $\mathbf{Q}$. Subsequently, $K-1$ additional groups, each containing N queries, are introduced, resulting in K groups in total, $\mathbf{Q}_1$ to $\mathbf{Q}_K$. Correspondingly, we have K groups of lane segment predictions, denoted as $\mathbf{Y}_1$ through to $\mathbf{Y_K}$. One-to-one assignment is then performed within each group to determine the optimal matching between each group of predicted lane segments and the ground truth $(\mathbf{Y_k}, \hat{\mathbf{Y}})$. Parallel Multi-Head Self-Attention is applied to each query group, and the resulting outputs are concatenated and fed to the lane attention module. With this group-wise assignment strategy, the training procedure proceeds as follows:

$$\begin{aligned}
&\text{Decoder}(\mathbf{F_{BEV}}, \mathbf{Q}_1) \rightarrow \tilde{\mathbf{Q}}_1, \quad \text{Predictor}(\tilde{\mathbf{Q}}_1) \rightarrow \mathbf{Y}_1, \\
&\text{Decoder}(\mathbf{F_{BEV}}, \mathbf{Q}_2) \rightarrow \tilde{\mathbf{Q}}_2, \quad \text{Predictor}(\tilde{\mathbf{Q}}_2) \rightarrow \mathbf{Y}_2, \\
&\quad \vdots \\
&\text{Decoder}(\mathbf{F_{BEV}}, \mathbf{Q}_K) \rightarrow \tilde{\mathbf{Q}}_K, \quad \text{Predictor}(\tilde{\mathbf{Q}}_K) \rightarrow \mathbf{Y}_K
\end{aligned} \tag{1}$$

During inference, the decoder operates similarly to the training phase, with the key difference being that only one group of queries is used. The total loss during training is the sum of the individual losses from each of the K decoders, expressed as follows:

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{n=1}^{N} \mathcal{L}(\mathbf{Y}_{\sigma_k(n)}, \tilde{\mathbf{Y}}_{kn}), \tag{2}$$

where $\mathcal{L}$ is the final cost function and $\sigma_k(.)$ is the optimal permutation of N indices for the k-th decoder as in Chen et al. (2023).

3.4. Topology Iterative Refinement

Previous map-learning methods skip iterative refinement, applying loss only to the final decoder output. Inspired by bounding-box refinement in Carion et al. (2020), TLSD iteratively refines lane segment queries after each decoder layer, mitigating noise sensitivity in intermediate predictions and their graph connectivity.

3.5. Graph Neural Network

Inspired by Li et al. (2023a) and Kipf and Welling (2016), we utilize graph neural network to model the relationship between potential connected lanes by learnable message weight $\mathbf{W}_{ll}^{i} \in \mathbb{R}^{|C_l| \times F_l \times F_l}$, where $C_l = \{\text{successor}, \text{predecessor}, \text{self-loop}\}$. The queries are further updated by

$$\begin{aligned}
A_{ll}^{i} &= \texttt{stack}\left(G_{ll}^{i-1}, \texttt{transpose}\left(G_{ll}^{i-1}\right), I\right), \\
Q_{l_{(x)}}^{i'} &= \sum_{\forall y \in N(x)} \sum_{\forall c_l \in C_l} \beta_{ll} \cdot A_{ll_{(c_l,x,y)}}^{i} \mathbf{W}_{l_{l(c_l)}}^{i} Q_{l_{(y)}}^{i}.
\end{aligned} \tag{3}$$

where $N(x)$ outputs the indices of all neighbors of the vertex with index x; G_{ll}^{i-1} is the adjacency matrix from the previous layer.

3.6. Distance-aware post-processing

The topology head reasons pairwise relationships on the given embeddings Q'_a and Q'_b to predict TOP_lsls. Other works such as Li et al. (2023a), Li et al. (2024b), Li et al. (2024a) and Yang et al. (2025) employ two MLPs to project the lane queries into other dimensions, and pass the concatenated refined queries to classifier MLPs to predict the probability of connectivity between lanes:

$$\begin{aligned}
Q'_a &= \text{MLP}_a(Q_a), \quad Q'_b = \text{MLP}_b(Q_b), \\
\texttt{conf} &= \texttt{sigmoid}(\text{MLP}_{\text{classifier}}(\texttt{concat}[Q'_a, Q'_b]))
\end{aligned} \tag{4}$$

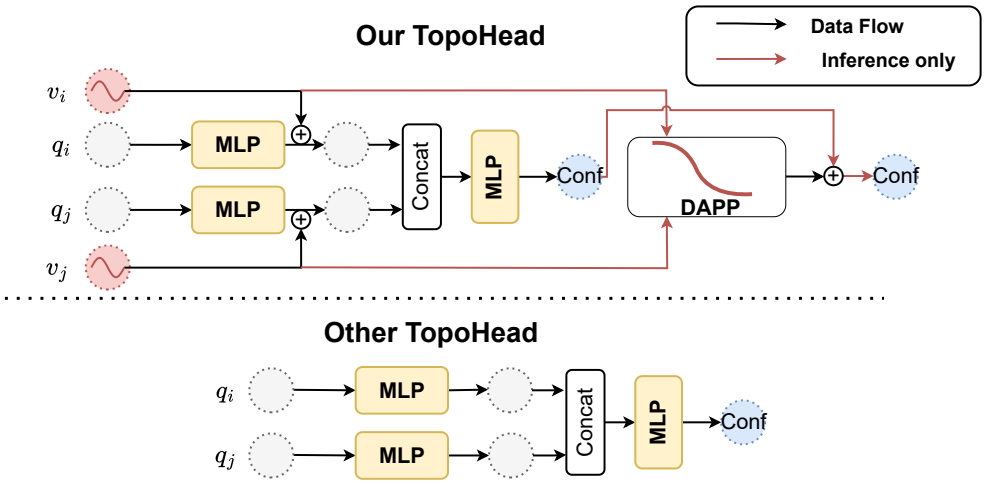

Figure 3: **Comparison between our and other topological head**. Our architecture takes into consideration the geometric information during both training and inference processes.

As illustrated in Figure 3, we project the 3D lane coordinate of center polylines V_{center} of each lane segment into high-dimensional space by a learnable linear projector. Therefore, the network now could leverage the 3D location of segments, further enhancing the connectivity probability among segments:

$$\begin{aligned} Q'_a &= \texttt{concat}(\text{MLP}(Q_a), \texttt{Linear}(l_a)), \\ Q'_b &= \texttt{concat}(\text{MLP}(Q_b), \texttt{Linear}(l_b)), \\ \text{confidence} &= \texttt{sigmoid}(\text{MLP}_{\text{top}}(\texttt{concat}[Q'_a, Q'_b])), \end{aligned} \tag{5}$$

where $\texttt{concat}, \texttt{sigmoid}$ are the concatenation and sigmoid activation function to scale the confidence score between 0 and 1.

The topological relationship between centerlines depends not only on semantic information but also on their geometric locations. If the endpoints of two centerlines are in close proximity, they are likely topologically related Fu et al. (2024). During the inference phase, at the final lane decoder layer, we predict N_l lane lines and then compute the geometric distances between the endpoint of one lane centerline and the starting points of the others.

$$\begin{aligned} l_0, \ldots, l_{N_l-1} &= \text{LaneHead}\left(Q_l^{L-1}\right) \\ d_{ij} &= \left| l_i^{\text{end}} - l_j^{\text{start}} \right| \\ D &= \{ d_{ij} \mid i, j = 0 \ldots N_l - 1 \} \end{aligned} \tag{6}$$

where $D \in \mathcal{R}^{N_l \times N_l}$ is the lane geometric distance matrix that contains L1 distance d_{ij} between l_i^{end} - the last point of lane line l_i, and l_j^{start} - the first point of lane line l_j,

To map distance to topology matrix, we utilize the following mapping function, inspired by Fu et al. (2024) $f : \mathcal{R} \to [0, 1]$ as follows:

$$f = \texttt{exp}\left\{ -\frac{x^{\alpha}}{\lambda \cdot} \right\} \tag{7}$$

where $x = d_{ij}$. In previous work, Fu et al. (2024), α, λ are the learnable parameters during the training phase while being chosen as the empirical value for the inference phase. We observed that these parameters are extremely sensitive, and directly employing the coordinate during training would lead to performance degradation.

We then construct the distance-aware connectivity matrix G_{dis} as follow:

$$G_{dis} = \{f(d_{ij}) \mid i, j = 0 \dots N_l - 1\} \tag{8}$$

The final connectivity matrix outputs combine both distance awareness and the high-space features:

$$G = \lambda_{dis} \cdot G_{dis} + \lambda_{hf} \cdot G_{hf}, \tag{9}$$

where G_{hf} is the matrix of potential connected segments derived from equation 5 and $\lambda_{dis}, \lambda_{hf}$ are set to 1 to balance the contribution of both information.

4. Experiments

4.1. Dataset and metrics

4.1.1. Dataset

We evaluate our method on the OpenLane-V2 dataset from Wang et al. (2023), which builds upon ArgoverseV2Wilson et al. (2021) and NuScenesCaesar et al. (2020). This dataset spans diverse real-world conditions across urban and suburban regions, with variations in lighting and weather. Our experiments use Subset A, featuring seven surrounding camera views per frame, with about 22K training and 4.8K validation frames.

4.1.2. Metrics

Our method focuses on predicting HDMaps with lane segment representation and topological reasoning. To quantify the accuracy of the lane-segment-based HDMaps, we follow the metrics proposed by Wang et al. (2023), Li et al. (2024b) and Yang et al. (2025):

- **AP**$_{ls}$: shows the average precision of lane segment computed under `Chamfer` and `Frechet` distance threshold of $\{0.5, 1.0, 1.5\}$ meters. Concretely, to measure the accuracy between the predicted segment $\tilde{S} = \{\tilde{\mathcal{V}}_{left}, \tilde{\mathcal{V}}_{center}, \tilde{\mathcal{V}}_{right}\}$ and ground truth $S = \{\mathcal{V}_{left}, \mathcal{V}_{center}, \mathcal{V}_{right}\}$, we employ the following equation:

$$D(\tilde{S}, S) = \frac{1}{2}\Big[D_{Frechet}(\tilde{\mathcal{V}}_{center}, \mathcal{V}_{center}) + \sum_{i \in \{left, right\}} D_{Chamfer}(\tilde{\mathcal{V}}_i, \mathcal{V}_i)\Big] \tag{10}$$

- **AP**$_{ped}$: illustrates the average precision of pedestrian crossing to evaluate map element construction quality.

- **mAP**: is the mean AP computed as the average of AP$_{ls}$ and AP$_{ped}$.

- **TOP**$_{lsls}$: measures the performance of topology reasoning. Specifically, the ground truth connectivity graph $G = (V, E)$ and the corresponding prediction $\tilde{G} = (\tilde{V}, \tilde{E})$, two vertices are considered connected if the predicted confidence score of the edge

between them exceeds 0.5. The $\mathbf{TOP}_{lsls}$ for a given vertex is determined by ranking all predicted edges associated with it and computing the cumulative mean of the precision values. Mathematically, the $\mathbf{TOP}_{lsls}$ is expressed as:

$$\mathbf{TOP}_{lsls} = \frac{1}{|V|} \sum_{v \in V} \frac{\sum_{\hat{n} \in N(v)} P(\hat{n}) \cdot 1(\hat{n} \in N(v))}{|N(v)|} \tag{11}$$

- **OLUS**: the overall performance for HDMap construction, equasl $\frac{1}{2}\left(\mathbf{mAP} + \sqrt{\mathbf{TOP}_{lsls}}\right)$

Moreover, to emphasize the performance of TLSD compared to the previous works that consider centerline only, we conduct a fair comparison by extracting the centerlines from the segments. To specify, we employ the following metrics, proposed by Wang et al. (2023), Li et al. (2023a) and Yang et al. (2025):

- $\mathbf{DET}_l$: similar to $\mathbf{AP}_{ls}$, but only measures the centerline.
- $\mathbf{TOP}_{ll}$: similar to $\mathbf{TOP}_{lsls}$, but only applying for the matched centerlines.
- **OLS**: average score, is computed as $\frac{1}{2}\left(\mathbf{DET}_l + \sqrt{\mathbf{TOP}_{ll}}\right)$

4.2. Results

We conduct a comprehensive evaluation of our proposed method, TLSD, against several state-of-the-art approaches, including TopoNet, MapTR, MapTRv2, LaneSegNet, TopoLogic, and Topo2Seq, using the training set of OpenLane-V2 Wang et al. (2023). Table 1 presents the results in terms of both map element detection accuracy and lane-to-lane topological prediction.

Under fair evaluation settings, TLSD achieves state-of-the-art performance. Specifically, in the 24-epoch training configuration, TLSD surpasses all baselines with a notable improvement of +2.6 **mAP** in lane segment detection and +1.2 **OLUS** in overall performance. Even with extended training to 48 epochs, TLSD maintains superior accuracy, achieving +1.1 **mAP**, +3.2 $\mathbf{TOP}_{lsls}$, and +1.7 **OLUS** over the strongest competing method.

We also introduce an efficient variant, eTLSD, which adopts a lightweight ResNet-18 backbone. Despite its compact architecture, eTLSD outperforms the ResNet-50-based Topo2Seq by +3.3% **mAP** and +6.6% $\mathbf{AP}_{ped}$ under identical training conditions. Furthermore, when compared to LaneSegNet using the same backbone, eTLSD consistently delivers better performance in both road geometry detection and topology prediction, confirming its efficiency and generalizability.

To further emphasize the performance of TLSD, we conduct comparison experiments on center polylines perception task in Table 2. We report performance under two training schedules: 24 epochs and 48 epochs, using three key metrics: $\mathbf{DET}_l$, $\mathbf{TOP}_{ll}$, and the derived composite score **OLS**. In the 24-epoch setting, our method TLSD achieves the highest performance across all metrics, significantly surpassing prior works such as Topo2Seq(+1.0 $\mathbf{DET}_l$ and +1.8 $\mathbf{TOP}_{ll}$). Notably, even the efficient version, eTLSD, which uses a lightweight ResNet-18 backbone, delivers strong results (**OLS** = 43.7), outperforming several methods that rely on heavier backbones, such as ResNet-50. Under the 48-epoch configuration, TLSD further improves performance, reaching $\mathbf{DET}_l = 37.8$, $\mathbf{TOP}_{ll} = 33.2$,

Method	Venue	mAP	AP_{ls}	AP_{ped}	TOP_{lsls}	OLUS
Backbone: ResNet-50, Epoch: 24						
TopoNet Li et al. (2023a)	Arxiv	23.0	23.9	22.0	-	-
MapTR Liao et al. (2023)	ICLR2023	27.0	25.9	28.1	-	-
MapTRv2 Liao et al. (2024)	IJCV2024	28.5	26.6	30.4	-	-
Topologic Fu et al. (2024)	NRIPS2024	33.2	33.0	33.4	**30.8**	44.3
LaneSegNet Li et al. (2024b)	ICLR2024	32.6	32.3	32.9	25.4	41.5
Topo2Seq Yang et al. (2025)	AAAI2025	33.6	33.7	33.5	26.9	42.7
TLSD (our)	ACML2025	**36.2**	**34.6**	**37.8**	30.1	**45.5**
Backbone: ResNet-50, Epoch: 48						
LaneSegNet Li et al. (2024b)	ICLR2024	36.4	34.9	37.9	27.3	44.3
Topo2Seq Yang et al. (2025)	AAAI2025	37.7	36.9	38.5	29.9	46.2
TLSD(our)	ACML2025	**38.3**	**38.2**	**38.4**	**33.1**	**47.9**
Backbone: ResNet-18, Epoch: 24						
LaneSegNet Li et al. (2024b)	ICLR2024	30.4	30.4	30.4	24.9	40.1
eTLSD (ours)	ACML2025	**34.7**	**33.7**	**35.6**	**29.1**	**44.3**

Table 1: **Comparison of TLSD with state-of-the-art methods on lane segment perception**. The best performances are highlighted in **bold**. All metrics are reported in percentage (%). We achieve state-of-the-art performance on all settings.

and **OLS** = 47.7, establishing a new state of the art. It outperforms the previous best, Topo2Seq, by margins of +1.1 $\mathbf{DET}_l$, +3.2 $\mathbf{TOP}_{ll}$, and +1.9 **OLS**. These results demonstrate not only the effectiveness of our full model but also the efficiency and scalability of the TLSD framework. The consistent improvements across training budgets and backbone configurations highlight TLSD's robustness in modeling both geometric and topological lane structure.

For qualitative comparison, we visualize bird's-eye-view HD maps generated by TLSD and LaneSegNet Li et al. (2024b), focusing on particularly challenging scenes characterized by dense lane structures and complex topologies. As shown in Figure 4, the first column (left to right) presents the multi-view input images, followed by the ground-truth annotations, the HD maps predicted by TLSD, and finally the outputs from LaneSegNet. TLSD consistently produces more refined and complete HD maps, even in scenes with highly intricate lane arrangements. Its outputs capture both rich geometric details and accurate connectivity graphs. Notably, in complex scenarios such as intersections, LaneSegNet usually fails to generate usable HDMaps, whereas TLSD successfully preserves the topological structure with only minor discrepancies.

Method	Venue	Backbone	$\mathbf{DET}_l$	$\mathbf{TOP}_{ll}$	$\mathbf{OLS}^*$
Epoch: 24					
STSU Can et al. (2021)	ICCV2021	R50	12.7	2.9	14.9
VectorMapNet Liu et al. (2023)	ICML2023	R50	11.1	2.7	13.8
MapTR Liao et al. (2023)	ICLR2023	R50	17.7	2.3	16.4
TopoNet Li et al. (2023a)	Arxiv	R50	28.6	10.9	30.8
Topo2D Li et al. (2024a)	Arxiv	R50	29.1	26.2	38.2
TopoMLP Wu et al. (2024)	ICLR2024	R50	28.3	21.7	37.4
TopoMLP* Wu et al. (2024)	ICLR2024	Swin-B	32.5	11.9	33.5
RoadPainter* Ma et al. (2025)	ECCV2024	R50	30.7	7.9	29.4
LaneSegNet Li et al. (2024b)	ICLR2024	R50	31.1	25.3	40.7
TopoLogic Fu et al. (2024)	NRIPS2024	R50	29.9	18.6	36.5
Topo2Seq Yang et al. (2025)	AAAI2025	R50	<u>33.5</u>	27.0	42.7
eTLSD(Our)	ACML2025	R18	33.3	<u>29.2</u>	<u>43.7</u>
TLSD(our)	ACML2025	R50	**34.5**	**31.0**	**45.1**
Epoch: 48					
LaneSegNet Li et al. (2024b)	ICLR2024	R50	34.3	27.3	43.3
Topo2Seq Yang et al. (2025)	AAAI2025	R50	36.7	30.0	45.8
TLSD(our)	ACML2025	R50	**37.8**	**33.2**	**47.7**

Table 2: **Comparison of TLSD with SOTA models on lane centerline perception on OpenLane-V2 validation set.** The best performances are highlighted in **bold**, while the second one is <u>underlined</u>. Metrics marked with * are extracted from the official $\mathbf{TOP}_{ll}$v1.0.0 results reported in prior work. The derived metric $\mathbf{OLS}^*$ reflects a composite of $\mathbf{DET}_l$ and $\mathbf{TOP}_{ll}$. All metrics are reported in percentage (%)

Index	TIR	GQ	LCE	GNN	DAPP	mAP	$\mathbf{AP}_{ls}$	$\mathbf{AP}_{ped}$	$\mathbf{TOP}_{lsls}$
1						30.4	30.4	30.4	24.9
2	✓					30.6	31.1	30.1	25.4
3	✓	✓				33.8	31.8	**35.8**	27.3
4	✓	✓	✓			33.4	33.3	33.4	27.3
5	✓	✓	✓	✓		**34.7**	**33.7**	**35.6**	28.1
6	✓	✓	✓	✓	✓	**34.7**	**33.7**	**35.6**	**29.1**

Table 3: **Ablation study on OpenLane-V2 benchmark**. The efficient version, eTLSD, is employed as a baseline. TIR, GQ, LCE, GNN, and DAPP refer to Topology Iterative Refinement, Group Query, Lane Coordinate Embedding, Graph Neural Network, and Distance-aware post-processing, respectively.

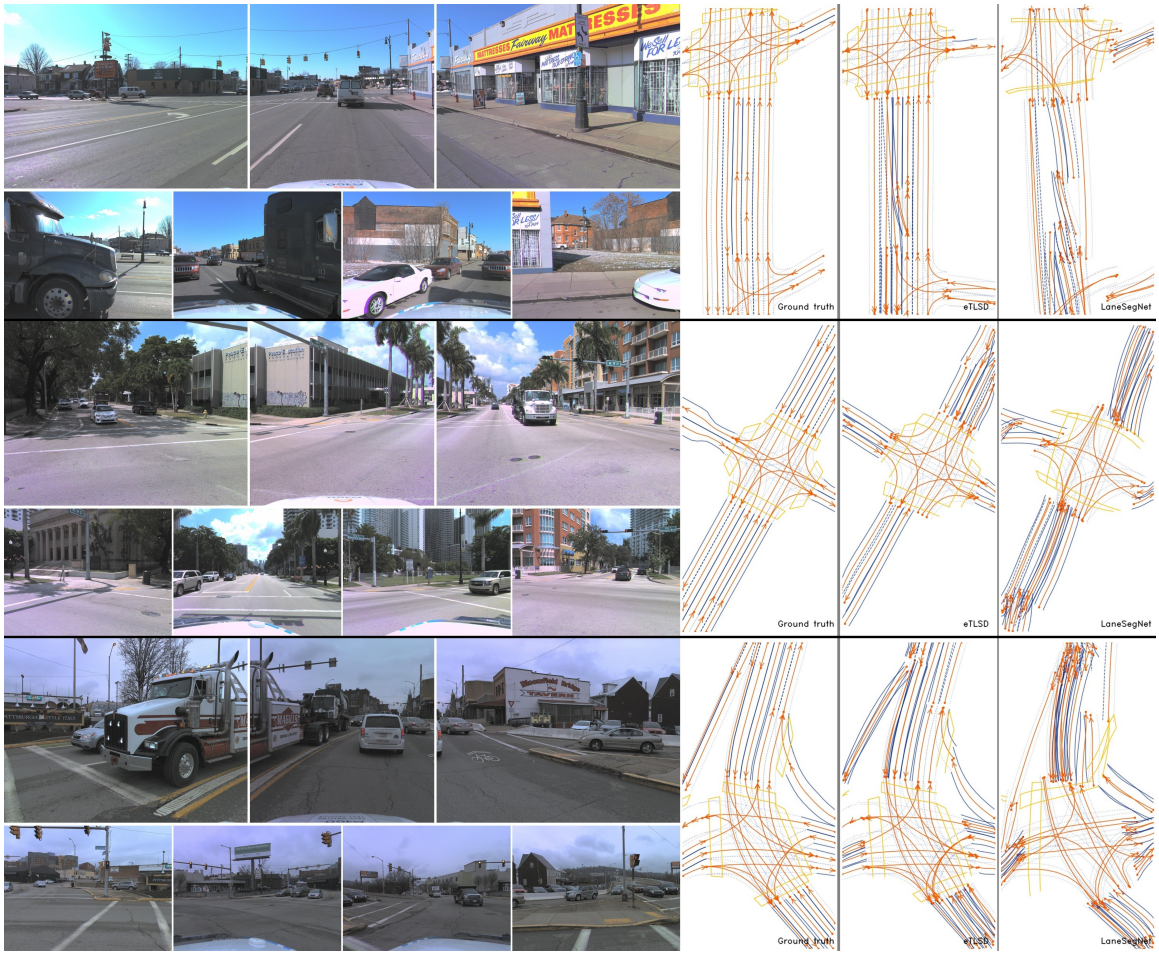

Figure 4: **Qualitative results between TLSD and baseline on OpenLane-V2 validation set**. Leftmost is the multi-view input image. TLSD exhibits more accurate HDMaps with minor discrepancies in the lane graph, compared to the baseline.

4.3. Ablation Study

4.4. Modules' effectiveness

In Table 3, we present an ablation study to assess the individual contributions of each proposed module. First, applying iterative refinement of topology queries across decoder layers yields a 2.0% improvement in lane prediction accuracy, indicating that progressive refinement enhances geometric reasoning. Next, replacing the standard one-to-one assignment in DETR with a group-based one-to-many assignment significantly boosts overall performance, most notably increasing $\mathbf{AP}_{ped}$ by 17.7%. This suggests that relaxing the strict matching constraint enables the model to retain diverse predictions, particularly benefiting fine-grained structures like pedestrian crossings. We also observe that incorporating Lane Coordinate Embedding improves lane detection accuracy. However, this gain comes with a slight decrease in $\mathbf{AP}_{ped}$, likely because the topology loss prioritizes lane-related features, which may overshadow supervision signals for pedestrian prediction. Replacing the base-

Setting	mAP	$\mathbf{AP}_{ls}$	$\mathbf{AP}_{ped}$	$\mathbf{TOP}_{lsls}$
PP-EL	0.324	0.298	0.349	0.263
PP-FL	**0.347**	**0.337**	**0.356**	**0.291**
T-EL	0.325	0.310	0.339	0.280
T-FL	0.328	0.306	0.349	0.280

Table 4: Ablation study on distance-aware mapping strategies.

Layer	mAP	$\mathbf{AP}_{ls}$	$\mathbf{AP}_{ped}$	$\mathbf{TOP}_{lsls}$
1	0.236	0.252	0.219	0.220
2	0.296	0.296	0.296	0.269
3	0.324	0.315	0.332	0.281
4	0.331	0.311	0.350	0.286
5	0.340	0.333	0.346	0.289
6	0.347	0.337	0.356	0.291

Table 5: Ablation study on decoder layer outputs.

line MLP reasoning module with a Graph Neural Network (GNN) results in a modest but consistent performance gain. This aligns with our hypothesis that GNN facilitates message passing among lane queries, improving feature aggregation and lane connectivity reasoning. Finally, introducing distance-aware post-processing provides a substantial 3.5% increase in $\mathbf{TOP}_{lsls}$, demonstrating the effectiveness of leveraging geometric priors to refine topological predictions between lane segments.

4.5. Training vs. post-processing

We also investigate the difference between post-processing and training, and applying to all layers or only the final. We begin by evaluating how the mapping function behaves when applied during training or post-processing, either across all decoder layers or exclusively at the final layer. As shown in Table 4, applying the mapping function only at the final decoder layer (PP-FL) yields the best performance across all metrics. In contrast, applying post-processing across all decoder layers (PP-EL) leads to a significant performance drop. We hypothesize that this degradation stems from the low precision of 3D coordinates predicted by early decoder layers, which are still under refinement and prone to noise. When such imprecise lane representations are passed into the graph reasoning module, they may introduce erroneous topological relationships due to inaccurate spatial anchoring.

To validate this hypothesis, we examine the prediction quality at each individual decoder layer. As shown in Table 5, performance metrics—including **mAP**, $\mathbf{AP}_{ls}$ and $\mathbf{TOP}_{lsls}$—consistently improve with deeper decoder layers. The first decoder output achieves only 0.236 **mAP** and 0.220 $\mathbf{TOP}_{lsls}$, indicating weak localization and topology understanding. As decoding progresses, the representations become increasingly refined, with the sixth (final) layer achieving the highest scores across all metrics, including 0.347 **mAP** and 0.291 $\mathbf{TOP}_{lsls}$. This clear upward trend confirms that decoder depth directly correlates with the precision of geometric and topological predictions. Therefore, applying the distance-aware mapping only at the final layer ensures that the model leverages the most reliable lane representations, avoiding noisy intermediate outputs that could corrupt downstream reasoning.

Additionally, we experiment with training the hyperparameters α and λ, initialized to 10 and 2, respectively. However, this adaptive training strategy also results in a degradation of overall performance. Specifically, training across all decoder layers (T-EL) leads to a performance drop of 6.34%, while restricting training to the final layer (T-FL) causes a

5.48% drop. These findings suggest that freezing the mapping function and applying it only at the final stage provides a better inductive bias for the model.

5. Conclusion

We present TLSD, a novel framework for high-definition map generation that achieves state-of-the-art performance in both map element detection and topological reasoning. We introduce key innovative modules, including iterative topology refinement, group-wise query assignment via DETR, and a graph neural network enhanced with lane coordinate embeddings. At inference, a distance-aware post-processing module further refines the predicted lane connectivity. Together, these components enable TLSD to deliver accurate, efficient, and topology-aware HD map predictions, advancing the current state of the art in vectorized mapping.

Acknowledgment

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15641–15650, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

Yanping Fu, Wenbin Liao, Xinyuan Liu, Hang xu, Yike Ma, Feng Dai, and Yucheng Zhang. Topologic: An interpretable pipeline for lane topology reasoning on driving scenes, 2024.

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

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

Han Li, Zehao Huang, Zitian Wang, Wenge Rong, Naiyan Wang, and Si Liu. Enhancing 3d lane detection and topology reasoning with 2d lane priors. *arXiv preprint arXiv:2406.03105*, 2024a.

Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022a.

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. Graph-based topology reasoning for driving scenes. *arXiv preprint arXiv:2304.05277*, 2023a.

Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. Lanesegnet: Map learning with lane segment perception for autonomous driving. In *ICLR*, 2024b.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485, 2023b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022b.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, Huang, and Chang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *International Conference on Learning Representations*, 2023.

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, pages 1–23, 2024.

Bencheng Liao, Shaoyu Chen, Bo Jiang, Tianheng Cheng, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 334–351, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72784-9.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

Katie Z Luo, Xinshuo Weng, Yan Wang, Shuang Wu, Jie Li, Kilian Q Weinberger, Yue Wang, and Marco Pavone. Augmenting lane perception and topology understanding with standard definition navigation maps. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4029–4035, 2024. doi: 10.1109/ICRA57147.2024.10610276.

Changsheng Lv, Mengshi Qi, Liang Liu, and Huadong Ma. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 17197–17206, June 2025.

Zhongxing Ma, Shuang Liang, Yongkun Wen, Weixin Lu, and Guowei Wan. Roadpainter: Points are ideal navigators for topology transformer. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 179–195, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73254-6.

Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. In *NeurIPS*, 2023.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: An simple yet strong pipeline for driving topology reasoning. *ICLR*, 2024.

Y. Yang, Y. Luo, B. He, E. Li, Z. Cao, C. Zheng, S. Mei, and Z. Li. Topo2Seq: Enhanced Topology Reasoning via Topology Sequence Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9318–9326, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.