RELTOPO: MULTI-LEVEL RELATIONAL MODELING FOR DRIVING SCENE TOPOLOGY REASONING

Anonymous authors
Paper under double-blind review

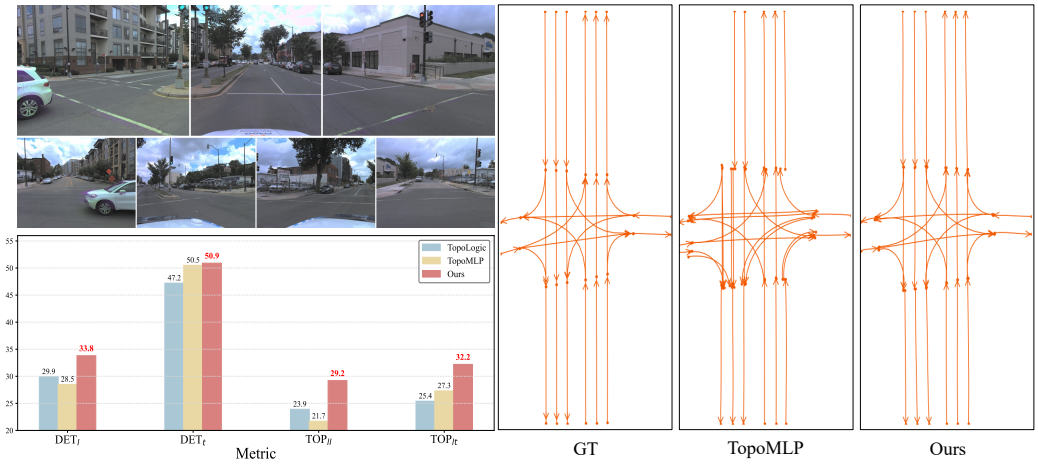

Figure 1: **Performance and visualization comparison between previous methods and ours.** The top-left shows multi-view input images. Our approach, RelTopo, integrates relational modeling across multiple levels, strengthening both perception and topology reasoning. As shown in the bottom-left quantitative results on OpenLane-V2, RelTopo significantly outperforms prior methods across all metrics. On the right, qualitative comparisons demonstrate that RelTopo produces more accurate lane geometries and well-aligned connectivity than prior approaches.

ABSTRACT

Accurate road topology reasoning is critical for autonomous driving, as it requires both perceiving road elements and understanding how lanes connect to each other (L2L) and to traffic elements (L2T). Existing methods often focus on either perception or L2L reasoning, leaving L2T underexplored and fall short of jointly optimizing perception and reasoning. Moreover, although topology prediction inherently involves relations, relational modeling itself is seldom incorporated into feature extraction or supervision. As humans naturally leverage contextual relationships to recognize road element and infer their connectivity, we posit that relational modeling can likewise benefit both perception and reasoning, and that these two tasks should be mutually enhancing. To this end, we propose RelTopo, a multi-level relational modeling approach that systematically integrates relational cues across three levels: **1) perception-level:** a relation-aware lane detector with geometry-biased self-attention and curve-guided cross-attention enriches lane representations; **2) reasoning-level:** relation-enhanced topology heads, including a geometry-enhanced L2L head and a cross-view L2T head, enhance topology inference via relational cues; and **3) supervision-level:** a contrastive InfoNCE strategy regularizes relational embeddings. This design enables perception and reasoning to be optimized jointly. Extensive experiments on OpenLane-V2 demonstrate that RelTopo significantly improves both detection and topology reasoning, with gains of +3.1 in DET$_l$, +5.3 in TOP$_{ll}$, +4.9 in TOP$_{lt}$, and +4.4 overall in OLS, setting a new state-of-the-art. Code will be released.

1 INTRODUCTION

Understanding road topology is fundamental for autonomous driving, as it provides structured knowledge of lanes and traffic elements for navigation and compliance with traffic rules. A complete topology model requires not only perceive road elements but also reason about how they connect: how lanes interconnect (Lane-to-Lane, L2L), and how traffic elements (*e.g.*, lights or signs) regulate specific lanes (Lane-to-Traffic-element, L2T). We follow the OpenLane-V2 benchmark Wang et al. (2024), where **perception** involves detecting lanes in 3D space from multi-view (MV) images and recognizing traffic elements in the front-view (FV), while **reasoning** infers L2L connectivity among lanes and L2T associations between lanes and traffic elements within only the FV.

Recent approaches (*e.g.*, Li et al. (2023a); Wu et al. (2023); Fu et al. (2024); Ma et al. (2024)) often adopt a two-stage pipeline: perception followed by reasoning. Despite recent advances, two major gaps remain: **First, fragmented task optimization:** methods often prioritize either perception or L2L reasoning, leaving L2T reasoning underexplored, and seldom optimizing them jointly Wu et al. (2023); Li et al. (2024); Fu et al. (2024); Lu et al. (2023). This fragmented treatment overlooks the opportunity for mutual enhancement between perception and reasoning, as the lack of end-to-end relational coupling prevents leveraging synergy across stages and constrains holistic optimization.

Concretely, in **perception**, many works Li et al. (2023a); Wu et al. (2023); Ma et al. (2024) adapt generic object detectors Liu et al. (2022b); Zhu et al. (2020) by turning bounding boxes into lane points or curves, but overlook the intrinsic geometric relationships Li et al. (2022a); Zhang et al. (2023b) (*e.g.*, parallelism, continuity) that humans naturally leverage. TopoLogic Fu et al. (2024) makes a partial attempt by encoding end-to-start distances as topology prior via a GNN, yet this focuses narrowly on connectivity and requires a separate graph module. A key question remains: *how can we embed structural cues directly into perception so that lane features become inherently relation-aware?* In **reasoning**, existing works rely heavily on coordinate encodings that are brittle to small endpoint shifts, making L2L prediction fragile. TopoLogic Fu et al. (2024) partly mitigates this using geometric distance in post-processing, but our study (*Tab. 4 in Supp.*) shows performance degrades sharply when removing that post-processing, implying the model fails to internalize relational modeling. L2T reasoning is even less explored: most methods naïvely fuse BEV lane features and FV traffic elements, ignoring cross-view misalignment. Some approaches Li et al. (2024) mitigates this issue by leveraging 2D lane features from an additional decoder, but adds computational overhead.

Second, weak relational modeling: Although topology reasoning inherently involves relation prediction, most works directly force the model relation prediction via supervision Li et al. (2023a); Wu et al. (2023); Li et al. (2024); Lu et al. (2023) or post-processing (*e.g.*, distance-based scoring in TopoLogic Fu et al. (2024)), rather than embedding relational cues into feature learning and connectivity prediction. As a result, structural priors (such as parallelism or proximity) between lanes and traffic elements are underutilized, leaving models brittle to geometric variation.

To address these gaps, we propose RelTopo, a **multi-level relational modeling** approach that that systematically integrates relational modeling across three levels. This enables perception and reasoning to be optimized jointly and coherently, grounded by structural relationships inherent in road scenes: **1) Relational Perception:** we develop a relation-aware lane detector based on a compact Bézier-curve representation. A *Geometry-Biased Self-Attention* module encodes inter-lane affinities (distance, orientation) as attention biases, while a *Curve-Guided Cross-Attention* module aggregates contextual cues along curves, ensuring relational awareness even with sparse control points. **2) Relational Reasoning:** we design relation-enhanced topology heads: a *Geometry-Enhanced L2L head* that embeds inter-lane distances to produce robust connectivity predictions, and a *Cross-View L2T head* that aligns BEV lanes with FV traffic elements to overcome spatial discrepancy. **3) Relational Supervision:** we introduce a InfoNCE Wu et al. (2018); He et al. (2020); Liu et al. (2021a)-based contrastive objective that pulls connected pairs closer and separates non-connected ones in the embedding space, regularizing relational learning.

Overall, our contributions can be summarized as follows:

- We identify two core limitations in existing topology reasoning, namely fragmented task optimization and weak relational modeling, and highlight the need for a unified relational perspective.
- We propose RelTopo, a multi-level relational modeling approach that integrates relational modeling through perception, reasoning, and supervision levels.

- Extensive experiments on the OpenLaneV2 dataset validate the effectiveness of our approach, demonstrating multi-level relation benefit mutually. With sufficient improvements, our method surpasses previous methods across both perception and reasoning, setting a new state-of-the-art.

2 Related Work

2.1 3D Lane Detection

3D lane detection provides the lane geometry for topology reasoning. Existing methods are typically designed for *monocular* front-view images and fall into BEV-based and front-view-based paradigms. BEV-based methods employ inverse perspective mapping (IPM) to transform FV images into BEV space for lane prediction Garnett et al. (2019); Guo et al. (2020); Efrat et al. (2020); Liu et al. (2022a); Chen et al. (2022); Li et al. (2022a). However, IPM-based methods suffer from distortions on non-flat roads due to their planar assumption, limiting its reliability in real-world settings. Front-view-based approaches instead predict 3D lanes directly from FV features, avoiding IPM distortions. Recent works Bai et al. (2023); Huang et al. (2023); Luo et al. (2023) adopt query-based detectors Carion et al. (2020); Zhu et al. (2020) to model 3D lanes information end-to-end, achieving stronger performance. Our work extends lane perception further by incorporating multi-view input and embedding relational cues directly into the perception process, enabling lane features to become inherently relation-aware.

2.2 Online HD Map Construction

Online HD map construction aims to dynamically generate structured map elements. Early methods Li et al. (2022b) use dense segmentation with heuristic vectorization. VectorMapNet Liu et al. (2023) advances this direction with an end-to-end detection-and-serialization pipeline. More recent works Liao et al. (2022; 2023); Qiao et al. (2023); Ding et al. (2023); Shin et al. (2023); Yu et al. (2023); Zhang et al. (2023b); Zhou et al. (2024); Hu et al. (2024) adopt Transformer-based architectures with curve- or point-based representations. MapTR Liao et al. (2022) introduces hierarchical query embeddings for polyline generation, BeMapNet Qiao et al. (2023) and PivotNet Ding et al. (2023) employ piecewise Bézier curves and dynamic points, and InstaGraM Shin et al. (2023) formulates map element generation as a graph problem with GNNs. GeMap Zhang et al. (2023b) models structural and relational properties of map elements, but focuses mainly on individual geometries without explicitly capturing topology relationships among lanes, and it relies on polyline representations composed of equidistant points, which lack flexibility and precision Ding et al. (2023); Zhang et al. (2023a) for nuanced lane description. To improve computational efficiency, various decoupled self-attention mechanisms Liao et al. (2023); Hu et al. (2024); Zhang et al. (2023b) have been proposed for integrating intra-/inter-instance information. However, topology reasoning among elements are limited in this area.

2.3 Driving Scene Topology Reasoning

Topology reasoning aims to capture the connectivity among road elements. Early works like STSU Can et al. (2021) and TPLR Can et al. (2022) construct lane graphs directly in BEV space, but focus only on L2L reasoning and ignore interactions with traffic elements. Recent methods begin to address both L2L and L2T reasoning. TopoNet Li et al. (2023a) employs a GNN-based framework that enhances topology prediction through message passing between element embeddings. TopoMLP Wu et al. (2023) adopts MLP heads for more efficiency, and LaneSegNet Li et al. (2023b) introduces lane segments augmented with left and right-side lane lines and proposes lane segment attention to capture intra-lane dependencies Yuan et al. (2024). However, these methods do not explicitly model relationships among lanes or between lanes and traffic elements, limiting their ability to capture dependencies among these elements. Topo2D Li et al. (2024) incorporates additional 2D lane detection to support topology reasoning, and TopoLogic Fu et al. (2024) combines geometry distance-based topology estimation with query similarity-based relational modeling via GNNs, but applies it primarily as post-processing for L2L. RoadPainter Ma et al. (2024) further refines point localization by utilizing BEV masks.

In contrast, our work takes a unified view: we embed relational modeling consistently across perception, reasoning, and supervision, enabling end-to-end training where relational cues guide both feature learning and topology inference.

3 METHOD

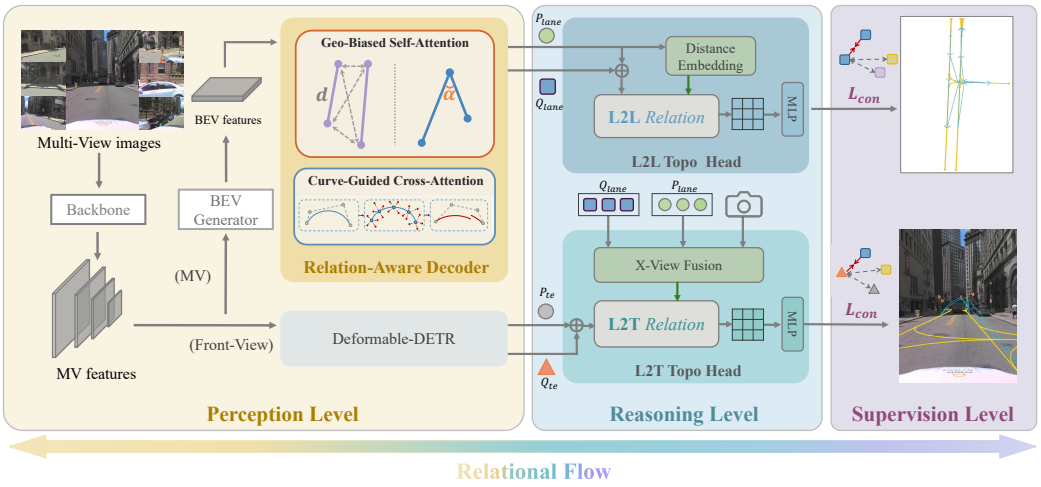

Figure 2: The overall framework of RelTopo, which employs multi-level relational modeling to jointly enhance lane perception and topology reasoning. Perception: a relation-aware lane decoder enrich lane features with structural cues. Reasoning: relation-enhanced topology heads infer topology between lanes (L2L) and between lanes and traffic elements (L2T) by encoding structural relations into relation embeddings, Supervision: relational embeddings are regularized via contrastive loss that pulls connected pairs closer and pushes non-connected pairs apart. This unified design ensures relational cues guide the entire pipeline. ■ (Q_{lane}), ▲ (Q_{te}) denote lane and traffic element queries, and ● (P_{lane}) and ● (P_{te}) for their predictions, respectively. L_{con} denotes our contrastive loss.

3.1 OVERVIEW

RelTopo weaves relational modeling through perception, reasoning, and supervision. Given multi-view images, the model comprises two branches: one decodes BEV lane features, and another detects front-view traffic elements. These outputs feed into relational topology heads to infer L2L and L2T connections. We inject relation cues early (in the decoder, Sec. 3.2), maintain them in reasoning (in the heads, Sec. 3.3), and sharpen them through contrastive loss in supervision (Sec. 3.4.2). The following subsections detail each module.

3.2 RELATIONAL PERCEPTION

Lanes in real driving scenes carry strong geometric relationships, such as parallelism and convergence, that can guide perception, which are neglected in prior works. We embed these geometric relational cues directly into the decoder so that lane features become structurally informed. Concretely, we introduce a **relation-aware decoder** that comprises two complementary mechanisms: 1) **geometry-biased self-attention** (Sec. 3.2.1) adds inter-lane geometric biases into the self-attention logits, steering each lane to attend to its peers in a relation-aware fashion; 2) **curve-guided cross-attention** (Sec. 3.2.2) samples points along the Bézier curve representation of each lane and performs deformable attention over those points. This enables context aggregation along the lane's shape rather than just at control points. These two design choices ensure that lane queries evolve under structural guidance, improving robustness in downstream topology reasoning.

3.2.1 GEOMETRY-BIASED SELF-ATTENTION

Training queries in DETR-like decoders can suffer from slow convergence due to lack of structural inductive bias Hou et al. (2025). To address this, we encode geometric relationships directly as bias terms in attention. For lane pairs (i, j), we compute features $\text{Dist}(\boldsymbol{l}_i, \boldsymbol{l}_j)$ (minimum end-to-start distance) and $\text{Angle}(\boldsymbol{l}_i, \boldsymbol{l}_j)$ (orientation difference). These are concatenated, embedded via a linear

projection GE (after sinusoidal encoding), and used as **Geometry**$(\cdot)$ biases:

$$\mathbf{Q} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{model}}}} + \textbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})\right)\mathbf{V} \tag{1}$$

$$\textbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})_{(i,j)} = \text{GE}(\text{Dist}(\boldsymbol{l}_i, \boldsymbol{l}_j) \| \text{Angle}(\boldsymbol{l}_i, \boldsymbol{l}_j)). \tag{2}$$

We illustrate this mechanism in ☐ of Fig. 3 (a), where the ***Geometry Bias*** represents the relation bias term $\textbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})_{(i,j)}$ between i-th and j-th lanes. This mechanism enhances query learning by introducing structural bias, as suggested in Hou et al. (2025), while also improving the capture of inherent lane geometric relationships. Unlike Topologic Fu et al. (2024), which relies on GNNs to encode connectivity, we take a simpler yet effective approach by directly encoding geometric relationships as attention biases within self-attention. Additionally, we incorporate angular information for comprehensive geometry relation modeling.

By enhancing geometrically proximal lanes, it allows the model to allocate greater attention to spatially relevant lanes through the interleaved attention process, leading to a more robust understanding of lane topology (detailed in Sec. 3.3.1). Our method differs from Hou et al. (2025), which encodes progressive cross-layer box positional relationships for individual objects.

3.2.2 Curve-Guided Cross-Attention

Polyline points with fixed spacing are often inflexible Ding et al. (2023); Zhang et al. (2023a). We instead adopt a cubic Bézier representation (four control points) for compactness. However, two challenges arise: 1) the sparsity of control points limits feature aggregation; 2) intermediate control points may be misaligned with the underlying curve Fig. 3(b).

To address these issues, instead of relying solely on sparse control points as reference points in deformable attention like TopoDBA Kalfaoglu et al. (2024), we sample K points along the Bézier curve, which serves as reference points for feature aggregation (Fig. 3 (b)). Furthermore, to capture long-range intra-lane dependencies, we employ a shared lane query to integrally generate offsets and weights for all sampled K points. This design enables contextual information flow holistically within each lane.

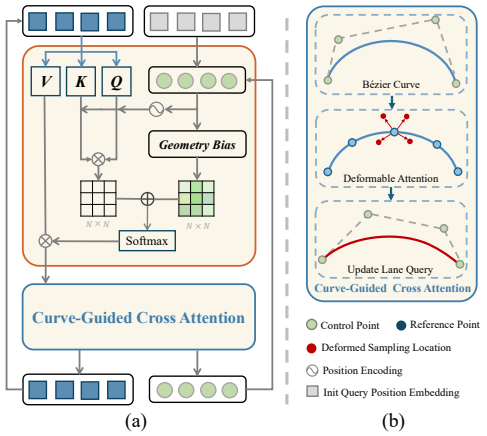

Figure 3: Illustration of our Bézier lane decoder layer, featuring our geometry-biased SA ☐ and curve-guided CA ☐.

Given the feature map $\boldsymbol{x}$, the l^{th} lane query $\boldsymbol{q}_l$ and its reference point $\boldsymbol{p}_l$, we adopt deformable attention mechanism Zhu et al. (2020) to update query features formulated as:

$$\text{DeformAttn}(\boldsymbol{q}_l, \boldsymbol{p}_l, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m [\sum_{i=1}^{N}\sum_{j=1}^{K} A_{mlij} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_l + \Delta \boldsymbol{p}_{mlij})], \tag{3}$$

where M is the number of attention heads, N denotes the number of offset locations per sampled point, and K is the number of sampled points along each curve. A_{mlij} and $\Delta \boldsymbol{p}_{mlij}$ denote the attentions weights and sampling offset, respectively, for the i^{th} offset of the j^{th} sampled points along the curve in the m^{th} attention head.

Unlike BézierFormer Dong et al. (2024), which projects sampled points onto image feature maps and performs `grid_sample`[1] to generate N_{ref} point queries, each undergoing separate deformable cross-attention, our approach is simpler and more effective, avoids the inefficiency of per-point

[1]This denotes the grid sampling operation defined in PyTorch.

queries and leverages global lane structure to guide local sampling, enhancing the modeling of long-range intra-lane dependencies. Besides, unlike BeMapNet Qiao et al. (2023), which represents each map element with multiple Bézier curves, our approach maintains efficiency and accuracy by using a single Bézier curve per lane. We also conduct comparative experiments against these two methods (see *Tab. 5, Supp.*), showing that our method achieves better prediction for lane structures.

3.3 Relational Reasoning

With relation-aware lane features in place, we infer connectivity in a unified relational space for both L2L and L2T. Our design departs from pipelines that either rely on brittle coordinate encodings Li et al. (2023a); Wu et al. (2023) or defer geometry to post-processing heuristics Fu et al. (2024). Instead, we embed relational priors *inside* the pairwise embeddings and align cross-view features end-to-end, thereby reducing sensitivity to endpoint shifts and eliminating reliance on external fixes.

3.3.1 Geometry-Enhanced L2L Reasoning

Existing methods Wu et al. (2023); Ma et al. (2024); Li et al. (2024) enhance lane features using coordinate-based encoding to predict L2L connectivity, which are, however, highly sensitive to endpoint shifts. Although Fu et al. (2024) attempt to remedy this by incorporating endpoint geometric distance as extra topology features (GDT), but the post-hoc design of their GDT fails to internalize relational modeling in model learning (as we discussed in Sec. 1 and analysis in Sec. A.3.1). This leaves relational structure underutilized. Motivated by these limitations, we propose a geometry-enhanced L2L reasoning head that embeds geometric priors directly into pairwise embeddings. Concretely, given refined lane features $\mathbf{Q}_{\text{lane}} \in \mathbb{R}^{N\times C}$ and endpoints $\mathbf{P}_{\text{lane}} \in \mathbb{R}^{N\times 2\times 2}$, we first project $\mathbf{Q}_{\text{lane}}$ into predecessor and successor embeddings with two MLPs (as L2L is directional, this choice is for capturing the asymmetric nature of L2L topology). Besides, we employ a sinusoidal positional encoding to produce lane positional embeddings based on $\mathbf{P}_{\text{lane}}$, yielding $\mathbf{PE}_{\text{lane}}$.

$$\mathbf{G}_{\text{L2L}} = (\text{MLP}_1(\mathbf{Q}_{\text{lane}}) + \mathbf{PE}_{\text{pre}}) \copyright (\text{MLP}_2(\mathbf{Q}_{\text{lane}}) + \mathbf{PE}_{\text{suc}}),\ \mathbf{G}_{\text{L2L}} \in \mathbb{R}^{N\times N\times C} \tag{4}$$

where © denotes broadcast concatenation. To reinforce connectivity relationships and stabilize against endpoint noise, we introduce a distance embedding:

$$\text{DistEmbed}^{i,j}_{\text{L2L}} = \text{MLP}(\texttt{distance}(\mathbf{P}^e_i - \mathbf{P}^s_j)),\ \ \text{DistEmbed} \in \mathbb{R}^{N\times N\times C} \tag{5}$$

where $\mathbf{P}^e_i$ and $\mathbf{P}^s_j$ denote the endpoint of lane i and the starting point of lane j, respectively. and then predict L2L topology via:

$$\mathbf{T}_{\text{L2L}} = \text{MLP}(\mathbf{G}_{\text{L2L}} + \text{DistEmbed}_{\text{L2L}}),\ \ \mathbf{T}_{\text{L2L}} \in \mathbb{R}^{N\times N\times 1} \tag{6}$$

By knitting geometry into the trainable relational embedding, our design internalizes robustness to small endpoint shifts and supports directional reasoning, unlike external post-processing hacks or pure coordinate encodings. This strengthens connected-pair discrimination and reduces false negatives in challenging scenarios. More details are illustrated in Appendix algorithm 1.

3.3.2 X-View L2T Reasoning

L2T reasoning requires bridging two inherently different representations: BEV features for lanes and FV features for traffic elements. The disparity between these representations poses a challenge for learning spatially consistent relationships. Many prior efforts treat these features as if they lie in the same space and naïvely fuse or concatenate them Wu et al. (2023); Ma et al. (2024). Topo2D Li et al. (2024) mitigates this by injecting 2D priors via auxiliary decoders.

In contrast, we introduce a X-view fusion module that aligns BEV and FV features more directly and cohesively. First, we project predicted 3D lane coordinates $\mathbf{P}^{3D}_{\text{lane}} \in \mathbb{R}^{N\times K\times 3}$ into the front-view plane to obtain $\mathbf{P}^{2D}_{\text{lane}} \in \mathbb{R}^{N\times K\times 2}$. We then sample spatial features $\mathbf{F}^{2D}_{\text{lane}}$ at those locations via `grid_sample`. These sampled features serve as a bridge between BEV and FV spaces, aligned by

3D geometry. Next, we fuse $\mathbf{F}_{\text{lane}}^{2D}$ into lane queries along with their 2D positional embeddings (using similar sinusoidal encoding strategy as in L2L module), producing enhanced queries:

$$\tilde{\mathbf{Q}}_{\text{lane}} = \text{MLP}_1(\mathbf{Q}_{\text{lane}}) + \mathbf{F}_{\text{lane}}^{2D} + \mathbf{PE}_{\text{lane}}^{2D}, \quad \tilde{\mathbf{Q}}_{\text{te}} = \text{MLP}_2(\mathbf{Q}_{\text{te}}) + \mathbf{PE}_{\text{te}}^{2D}. \tag{7}$$

Here the positional embeddings encode 2D spatial layout in the front-view, ensuring the fused features remain spatially coherent. We then form the relational embedding $\mathbf{G}_{\text{L2T}} \in \mathbb{R}^{N\times M\times C}$ by broadcast concatenation of $\tilde{\mathbf{Q}}_{\text{lane}}$ and $\tilde{\mathbf{Q}}_{\text{te}}$, and feed it into an MLP to predict the L2T topology:

$$\mathbf{T}_{\text{L2T}} = \text{MLP}(\mathbf{G}_{\text{L2T}}), \quad \mathbf{T}_{\text{L2T}} \in \mathbb{R}^{N\times M\times 1}.$$

By aligning BEV lane features to FV contexts via projected geometry and position embeddings, we reduce misalignment between two space representation, enabling more robust lane–traffic associations. Empirically, this design lowers erroneous associations in setups where naïve fusion would fail, especially at larger distances or oblique angles (shown in the qualitative comparison in Sec. A.3.4).

3.4 Loss Functions

3.4.1 Perception Loss

For lane detection, we employ Focal Loss Lin (2017) for classification and a combination of point-wise L1 loss and Chamfer distance loss over K sampling points to guide accurate geometric regression. For traffic element detection, we again use Focal Loss Lin (2017) for classification, along with L1 loss and GIoU loss Rezatofighi et al. (2019) to supervise bounding box regression over (x, y, w, h).

3.4.2 Relational Supervision

Most existing methods rely solely on Focal Loss to determine connectivity over pairs. However, this approach primarily emphasizes binary classification without explicitly distinguishing the relative importance of connected and non-connected pairs. To better capture topological relationships, we augment this with a **contrastive** InfoNCE loss Oord et al. (2018) over pairwise relation embeddings.

Concretely, consider the L2T case (the same logic applies to L2L). Let N be the number of lane queries and M the number of traffic element queries. The relational head produces an embedding tensor $\mathbf{G}_{\text{L2T}} \in \mathbb{R}^{N\times M\times C}$ and outputs a logit tensor $\mathbf{T}_{\text{L2T}}$. We treat pairs labeled as "connected" as positives, and all others as negatives. To sharpen discrimination, we introduce a **hard negative mining** strategy: for each positive pair, we select the top-n hardest negative pairs according to predicted logits, focusing the contrastive signal where confusion is highest. We then apply a symmetric InfoNCE loss (in the spirit of CLIP Radford et al. (2021)-style contrastive formulations) over these logits. Let $\mathbf{v}^+$ and $\{\mathbf{v}^-\}$ denote the positive and selected negative logits, then:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\mathbf{v}^+)}{\exp(\mathbf{v}^+) + \sum_{\mathbf{v}^-} \exp(\mathbf{v}^-)} = \log\left[1 + \sum_{\mathbf{v}^-} \exp(\mathbf{v}^- - \mathbf{v}^+)\right], \tag{8}$$

To handle multiple positives for a given anchor, we extend Eq. (8), following previous works Wu et al. (2022); Wang et al. (2023), to:

$$\mathcal{L}_{\text{con}} = \log\left[1 + \sum_{\mathbf{v}^+}\sum_{\mathbf{v}^-} \exp(\mathbf{v}^- - \mathbf{v}^+)\right]. \tag{9}$$

4 Experimental Results

4.1 Dataset and Metrics

Dataset. We evaluate our method on OpenLane-V2 Wang et al. (2024), a large-scale dataset specifically designed for topology reasoning in autonomous driving comprising two subsets: `subsetA` (from Argoverse-V2 Wilson et al. (2023)) and `subsetB` (from nuScenes Caesar et al. (2020)).

Table 1: **Performance comparison with state-of-the-art methods** on the OpenLane-V2 `subsetA` and `subsetB` dataset under the latest V2.1.0 evaluation. Results for RoadPainter‡ derive from their paper which uses old metrics (due to the absence of open-source code or model). TopoMLP† results were obtained with their provided model, while other results were sourced from the TopoLogic paper. Metrics are reported for detection accuracy (DET_l and DET_t) and topology reasoning accuracy (TOP_{ll} and TOP_{lt}), with overall score (OLS) indicating aggregated performance. Higher values indicate better performance across all metrics. Our method achieves the SOTA performance.

Subset	Method	Venue	Backbone	Epoch	DET_l ↑	DET_t ↑	TOP_{ll} ↑	TOP_{lt} ↑	OLS ↑
`setA`	VectorMapNet	ICML2023	ResNet-50	24	11.1	41.7	2.7	9.2	24.9
	MapTR	ICLR2023	ResNet-50	24	17.7	43.5	5.9	15.1	31.0
	TopoNet	Arxiv2023	ResNet-50	24	28.6	48.6	10.9	23.8	39.8
	TopoMLP†	ICLR2024	ResNet-50	24	28.5	<u>50.5</u>	21.7	<u>27.3</u>	<u>44.5</u>
	RoadPainter‡	ECCV2024	ResNet-50	24	<u>30.7</u>	47.7	7.9	24.3	38.9
	TopoLogic	NeurIPS2024	ResNet-50	24	29.9	47.2	<u>23.9</u>	25.4	44.1
	Ours	-	ResNet-50	24	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**
	Improvement	-	-	-	3.1 ↑	0.4 ↑	5.3 ↑	4.9 ↑	4.4 ↑
`setB`	TopoNet	Arxiv2023	ResNet-50	24	24.3	55.0	6.7	16.7	36.8
	TopoMLP†	ICLR2024	ResNet-50	24	26.0	<u>58.2</u>	21.0	<u>19.8</u>	<u>43.6</u>
	RoadPainter‡	ECCV2024	ResNet-50	24	<u>28.7</u>	54.8	8.5	17.2	38.5
	TopoLogic	NeurIPS2024	ResNet-50	24	25.9	54.7	<u>21.6</u>	17.9	42.3
	Ours	-	ResNet-50	24	**32.6**	**58.8**	**31.8**	**25.8**	**49.7**
	Improvement	-	-	-	3.9 ↑	0.6 ↑	10.2 ↑	6.0 ↑	6.1 ↑

Evaluation Metrics. Following the official evaluation of OpenLane-V2 Wang et al. (2024), we utilize DET_l and DET_t to measure detection accuracy for lanes and traffic elements, respectively. For topology reasoning, we employ TOP_{ll} and TOP_{lt} to assess L2L and L2T relationship prediction. The overall performance is quantified using the OpenLane-V2 Score (OLS):

$$\text{OLS} = \frac{1}{4}\left[\text{DET}_l + \text{DET}_t + f(\text{TOP}_{ll}) + f(\text{TOP}_{lt})\right], \tag{10}$$

where f denotes the square root function. Our evaluations follow the latest version (`V2.1.0`) of the metrics, as updated in the official OpenLane-V2 GitHub repository[2].

4.2 Implementation Details

Model Details. We use a ResNet-50 backbone to extract features, coupled with a pyramid network, FPN, for multi-scale feature learning. Following prior work Fu et al. (2024); Li et al. (2023a), a BEVFormer encoder Li et al. (2022c) with 3 layers is employed to generate a BEV feature map of size 100×200. We employ six decoder layers, using 300 queries for the lane decoder and 100 queries for the traffic element decoder following Wu et al. (2023).

Training Details. We utilize the AdamW optimizer Loshchilov & Hutter (2017) for model training, with a weight decay of 0.01 and an initial learning rate of 2.0×10^{-4}, which decays following a cosine annealing schedule. Training is conducted for 24 epochs using a total batch size of 8 on 8 NVIDIA 4090 GPUs. Input images are resized to 1024×800, following Wu et al. (2023). Due to space limitations, the training loss is provided in the Appendix Sec. A.2.1.

4.3 Main Results

We compare our model against SOTA methods on OpenLane-V2, with results shown in Tab. 1. On `subsetA`, RelTopo achieves an OSL of 48.9, surpassing all previous methods by a significant margin (+4.4). The gains are consistent gains across metrics, with particularly strong improvements in lane detection (+3.1 DET_l) and topology (+5.3 TOP_{ll}, +4.9 TOP_{lt}). On `subsetB`, RelTopo further improves to 49.7 OLS (+6.1), driven by large improvements in topology reasoning (+10.2 TOP_{ll} and +6.0 TOP_{lt}), while maintaining consistent improvements in detection metrics. These results confirm that our relational modeling at perception, reasoning, and supervision levels systematically

[2] https://github.com/OpenDriveLab/OpenLane-V2/issues/76

strengthens both perception and topology reasoning. Although our model adopts the same traffic decoder as prior work Wu et al. (2023), it still achieves +0.4/+0.6 gains in DET_t on `setA`/`setB`, suggesting that relational cues learned for L2T reasoning also benefit traffic element perception.

Overall, these results highlight the superiority of RelTopo , which sets a new state-of-the-art on both OpenLane-V2 `SubsetA` and `SubsetB`. To further illustrate its effectiveness, we present **qualitative results** in Fig. 4, where our model produces more accurate lane predictions and better-aligned connection points in complex driving scenes. More detailed visualizations are provided in Appendix Sec. A.3.4, together with a discussion of limitations and future directions in Sec. A.3.3.

4.4 Ablation Studies

To thoroughly evaluate our design, we conduct ablation studies on OpenLane-V2 `subsetA`. Additional analyses are included in the *Appendix*, covering: **1)** a comparison of our `SA` and L2L relational modeling against Topologic Fu et al. (2024) (Sec. A.3.1); **2)** alternative Bézier representations (Sec. A.3.2); **3)** analysis of DET_t variation (Sec. A.3.5); and **4)** evaluation of scalability and encoder generalizability (Sec. A.3.6). Our baseline model (#1) adopts a deformable-DETR decoder following Wu et al. (2023) with lightweight MLP-based topology heads, results are shown in Tab. 2.

Table 2: **Ablation study on our key components**. Columns correspond to Relational Perception (geometry-guided SA and curve-guided CA), Relational Reasoning, and Relational Supervision. Colored rows mark the milestones upon equipping the respective modules.

	Rel. Perception		Rel. Reasoning		Rel. Sup.	Metrics				
#L	SA	CA	L2L	L2T	NCE	DET_l	DET_t	TOP_{ll}	TOP_{lt}	OLS
#1						27.7	50.1	21.8	28.4	44.5
#2	✓					28.1	48.8	24.7	28.3	44.9
#3	✓	✓				32.7	50.1	26.8	29.9	47.3
#4	✓	✓	✓			33.7	48.0	28.5	29.7	47.4
#5	✓	✓		✓		33.5	48.4	26.4	31.4	47.4
#6	✓	✓	✓	✓		33.9	50.3	28.6	30.6	48.2
#7	✓	✓	✓	✓	✓	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**

Effect of Relational Perception. Our relation-aware lane decoder includes Geometry-Biased Self-Attention (`SA`) and Curve-Guided Cross-Attention (`CA`). Replacing standard self-attention with `SA` (#2) improves TOP_{ll} by +2.9, showing the benefit of encoding inter-lane geometry. Besides, we compare our geometry encoding with the distance topology method from Topologic in Appendix Tab. 4, further validating the advantages of our method. Introducing `CA` (#3) further boosts DET_l by +4.6, as curve-guided sampling better aggregates long-range features. Together, `SA` and `CA` (#3) deliver substantial gains over baseline (#1): +5.0 DET_l, +5.0 TOP_{ll}, +1.5 TOP_{lt}, and +2.8 OLS. These results confirm the combined benefits within our relational perception.

Effect of Relational Reasoning. We next evaluate our relational topology heads. Adding our geometry-enhanced L2L head (#3→#4) improves TOP_{ll} by +1.7 and also raises DET_t by +1.0, validating its effectiveness in capturing L2L relationships and showing that stronger lane connectivity reasoning indirectly benefits perception. Replacing the baseline L2T head with our cross-view design (#3→#5) improves TOP_{lt} by +1.5, demonstrating better lane–traffic alignment. Integrating both heads (#6) yields complementary gains (+1.8 TOP_{ll}, +0.7 TOP_{lt}, +0.9 OLS), confirming that relational reasoning at both L2L and L2T levels jointly strengthens performance.

Effect of Relational Supervision: Finally, we incorporate our contrastive loss for additional supervision in topology learning. Compared to #6, adding our contrastive regularization (#7) provides further gains (+0.6 TOP_{ll}, +1.6 TOP_{lt}), validating that relational supervision helps structure the embedding space for more discriminative topology learning.

Efficiency and Performance. To quantify the overhead of our relational modules, we report model size and inference latency for all ablation variants in Tab. 2 on a single RTX 4090 GPU, as shown in Tab. 3. Equipping our relational perception and relational reasoning increases latency only slightly,

Table 3: **Ablation study on efficiency and performance trade-offs**. Columns correspond to Relational Perception (SA and CA), Relational Reasoning (L2L and L2T), and Relational Supervision. We report Model Size (MB), Latency (ms), percentage increase in Cost (Δ Cost), and OLS.

	Rel. Perception		Rel. Reasoning		Rel. Sup.	Efficiency & Performance Metrics				
#L	SA	CA	L2L	L2T	NCE	Size (MB)	Lat. (ms)	Δ Cost	OLS	Δ OLS
#1						218.7	159.7	–	44.5	–
#2	✓					218.7	160.6	+0.6%	44.9	+0.4
#3	✓	✓				223.2	161.4	+1.1%	47.3	+2.8
#4	✓	✓	✓			224.2	161.8	+1.3%	47.4	+2.9
#5	✓	✓		✓		224.2	161.7	+1.3%	47.4	+2.9
#6	✓	✓	✓	✓		224.8	162.2	+1.6%	48.2	+3.7
#7	✓	✓	✓	✓	✓	224.8	162.2	+1.6%	48.9	+4.4

from 159.7 ms to 162.2 ms (about ≈1.6% over the baseline), while yielding a substantial OLS gain of +4.4 points (#1 → #7). Note that our relational supervision is used only during training (*i.e.*, #6 → #7 introduces no extra inference overhead). Thanks to the lightweight and effective design of our multi-level relational modeling, these gains come at negligible computational cost.

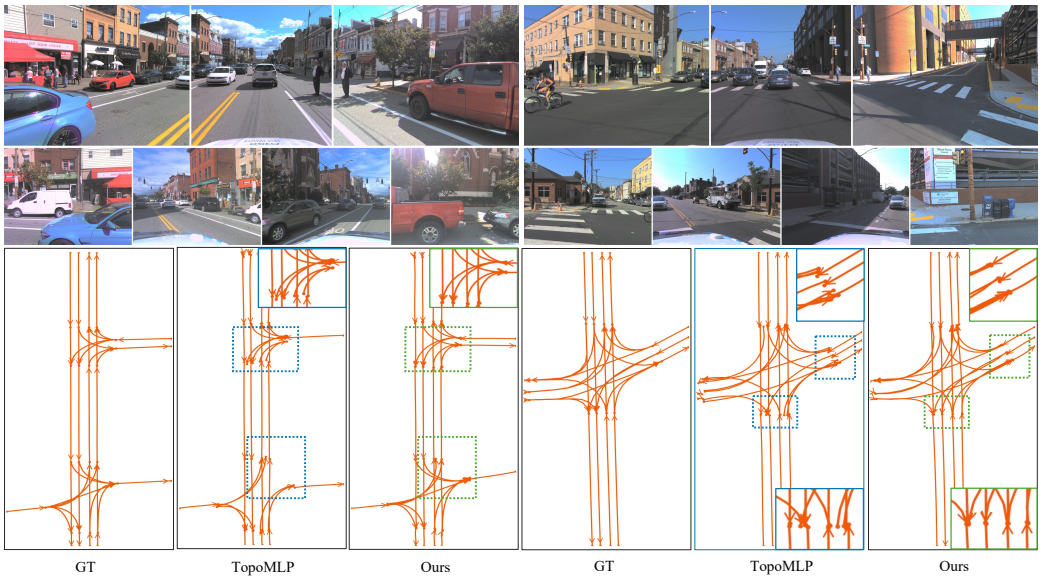

Figure 4: Comparative visual results on OpenLane-V2 `subsetA`. The top row shows multi-view input images, the bottom row shows lane predictions. We show comparison between ground truth, TopoMLP Wu et al. (2023) and ours. The blue box highlights misaligned connection point predictions from TopoMLP, and the **green** box shows the corresponding aligned predictions from our **RelTopo** . For clarity, zoomed-in views of selected regions are displayed at the top-right or bottom-right corners.

5 Conclusion

In this work, we present RelTopo , multi-level relational modeling method for driving scene topology reasoning. By systematically embedding relational cues at three levels: *relational perception*, *relational reasoning* and *relational supervision*, our design enables perception and reasoning to be optimized jointly and coherently. Extensive experiments on OpenLane-V2 demonstrate that this unified perspective yields consistent gains across detection and topology metrics, setting a new SOTA. We believe RelTopo highlights the importance of explicitly modeling structural relations in driving scenes, and can serve as a strong foundation for future research in this area.

REFERENCES

Yifeng Bai, Zhirong Chen, Zhangjie Fu, Lang Peng, Pengpeng Liang, and Erkang Cheng. Curveformer: 3d lane detection by curve propagation with curve queries and attention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7062–7068. IEEE, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

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

Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Topology preserving local road network estimation from single onboard camera image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17263–17272, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229. Springer, 2020.

Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pp. 550–567. Springer, 2022.

Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3672–3682, 2023.

Zhiwei Dong, Xi Zhu, Xiya Cao, Ran Ding, Caifa Zhou, Wei Li, Yongliang Wang, and Qiangbo Liu. Bézierformer: A unified architecture for 2d and 3d lane detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Netalee Efrat, Max Bluvstein, Shaul Oron, Dan Levi, Noa Garnett, and Bat El Shlomo. 3d-lanenet+: Anchor free lane detection using a semi-local representation. *arXiv preprint arXiv:2011.01535*, 2020.

Yanping Fu, Wenbin Liao, Xinyuan Liu, Yike Ma, Feng Dai, Yucheng Zhang, et al. Topologic: An interpretable pipeline for lane topology reasoning on driving scenes. *arXiv preprint arXiv:2405.14747*, 2024.

Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2921–2930, 2019.

Yuliang Guo, Guang Chen, Peitao Zhao, Weide Zhang, Jinghao Miao, Jingao Wang, and Tae Eun Choe. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 666–681. Springer, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, Badong Chen, and Xuguang Lan. Relation detr: Exploring explicit position relation prior for object detection. In *European Conference on Computer Vision*, pp. 89–105. Springer, 2025.

Haotian Hu, Fanyi Wang, Yaonong Wang, Laifeng Hu, Jingwei Xu, and Zhiwang Zhang. Admap: Anti-disturbance framework for reconstructing online vectorized hd map. *arXiv preprint arXiv:2401.13172*, 2024.

Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17451–17460, 2023.

Muhammet Esat Kalfaoglu, Halil Ibrahim Ozturk, Ozsel Kilinc, and Alptekin Temizel. Topobda: Towards bezier deformable attention for road topology understanding. *arXiv preprint arXiv:2412.18951*, 2024.

Chenguang Li, Jia Shi, Ya Wang, and Guangliang Cheng. Reconstruct from top view: A 3d lane detection approach based on geometry structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4370–4379, 2022a.

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*, 2024.

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)*, pp. 4628–4634. IEEE, 2022b.

Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, et al. 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. *arXiv preprint arXiv:2312.16108*, 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*, pp. 1–18. Springer, 2022c.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022.

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. *arXiv preprint arXiv:2308.05736*, 2023.

T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

Ruijin Liu, Dapeng Chen, Tie Liu, Zhiliang Xiong, and Zejian Yuan. Learning to predict 3d lane shape and camera pose from a single image via geometry constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1765–1772, 2022a.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021a.

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*, pp. 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*, pp. 531–548. Springer, 2022b.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, and Li Zhang. Translating images to road network: A non-autoregressive sequence-to-sequence approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23–33, 2023.

Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, and Zhen Li. Latr: 3d lane detection from monocular images with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7941–7952, 2023.

Zhongxing Ma, Shuang Liang, Yongkun Wen, Weixin Lu, and Guowei Wan. Roadpainter: Points are ideal navigators for topology transformer. *arXiv preprint arXiv:2407.15349*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13218–13228, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

Juyeb Shin, Francois Rameau, Hyeonjun Jeong, and Dongsuk Kum. Instagram: Instance-level graph modeling for vectorized hd map learning. *arXiv preprint arXiv:2301.04470*, 2023.

Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36, 2024.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*, 2023.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: An simple yet strong pipeline for driving topology reasoning. *arXiv preprint arXiv:2310.06753*, 2023.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pp. 588–605. Springer, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Jingyi Yu, Zizhao Zhang, Shengfu Xia, and Jizhang Sang. Scalablemap: Scalable map learning for online long-range vectorized hd map construction. In *Conference on Robot Learning*, pp. 2429–2443. PMLR, 2023.

Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7356–7365, 2024.

Gongjie Zhang, Jiahao Lin, Shuang Wu, Yilin Song, Zhipeng Luo, Yang Xue, Shijian Lu, and Zuoguan Wang. Online map vectorization for autonomous driving: A rasterization perspective. 2023a.

Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue. Online vectorized hd map construction using geometry. *arXiv preprint arXiv:2312.03341*, 2023b.

Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. Himap: Hybrid representation learning for end-to-end vectorized hd map construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15396–15406, 2024.

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