2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Confidence-Gated Topology Reasoning with Fiducial Marker Priors for Occlusion-Robust Lane Graph Prediction

Zirui Wu and Xianbiao Hu*

***Abstract*— Accurate lane topology perception is crucial for safe autonomous driving, yet vision-based models such as BEVFormer and TopoNet degrade under heavy occlusion and other visibility degradations (e.g., ambiguous road markings). Existing approaches augment vision with global priors like Standard Definition (SD) maps, but these rely on precise GNSS localization and global alignment, which can be unreliable in urban canyons, tunnels, or GNSS-denied areas. Fiducial markers provide a complementary alternative: compact infrastructure-embedded tags that encode structurally complete local lane graphs, mitigating blind spots in topology reasoning where visual pipelines fail. However, marker detections are not always reliable—pose estimates may degrade with distance, and detections may be intermittent under occlusion. To address these challenges, we propose a Confidence-Gated Marker Fusion framework that integrates marker-derived priors into BEV features through a dynamic gating mechanism, regulating the contribution of noisy long-range inputs. In addition, we introduce a temporal marker memory that caches and decays reliable priors across frames, propagating topology guidance during short-term detection gaps. Evaluated on a marker-augmented OpenLane-V2 benchmark, our method outperforms both vision-only and SD map-augmented baselines, achieving notable gains (27%) in lane graph completeness and occlusion robustness. These results demonstrate that fiducial marker priors, when fused with vision-based reasoning, provide a practical and reliable pathway toward resilient lane topology prediction in GNSS-denied urban scenarios.**

I. Introduction

Accurate lane topology perception is essential for safe and efficient autonomous driving, particularly in urban environments where vehicles must navigate multi-lane structures, intersections, and complex connectivity. This task requires identifying lane centerlines, determining their connectivity, and associating lanes with traffic elements. Recent advances in multi-view vision-based approaches, including BEV Transformer architectures and graph-based frameworks such as TopoNet [1], have demonstrated strong capabilities in predicting lane centerlines and their connectivity directly from camera inputs. However, the performance of these vision-centric models fundamentally relies on reliable visual cues, which are often compromised in real-world conditions. Occlusions from large vehicles or roadside structures can block critical regions, while other visibility degradations such as adverse weather or faded markings further reduce reliability. These blind-spot structures, once omitted, cannot be reconstructed by downstream topology reasoning, as GNN-based modules like TopoNet inherently depend on available centerline candidates to infer relationships.

(a) Fiducial Marker

(b) Topology Reasoning

Fig. 1: (a) An example of ArUco fiducial marker. (b) Lane topology reasoning in urban scenes. Markers can provide priors to complement vision-based reasoning.

To mitigate these challenges, prior work has explored augmenting lane topology models with global priors such as Standard Definition (SD) navigation maps. Frameworks like SMERF [2] show that SD maps can enhance lane detection and topology reasoning by injecting coarse road-level structure into BEV features. However, such methods require precise GNSS or LiDAR-based localization for map alignment, which can be unreliable in urban canyons, tunnels, or GNSS-denied areas, and may suffer from outdated or incomplete map data.

Fiducial markers—compact, infrastructure-embedded visual tags—offer a complementary alternative. Fiducial markers are machine-readable patterns designed for robust visual detection and pose estimation, with common examples including AprilTag [3] and ArUco [4]. When strategically deployed on roadways, they can encode local lane topology, enabling vehicles to directly access intersection layouts and lane connectivity without GNSS or global map registration. This localized, infrastructure-supported prior naturally fits cooperative perception paradigms and can fill critical structural gaps where visual features fail.

Despite their promise, fiducial markers face two main challenges when serving as priors for lane topology reasoning. First, detections are spatially uncertain: long-range observations, unfavorable viewing angles, or adverse conditions degrade pose estimation. In this scenario, fiducial markers are typically up to one meter in size and are mounted on roadside poles, traffic signals, or gantries. Their small physical footprint makes them challenging to detect reliably at long range, where small pixel-level errors can translate into large spatial uncertainties. Second, detections are temporally inconsistent: markers may be intermittently occluded by moving vehicles or roadside objects, or missed due to sparse deployment, resulting in unstable estimates

Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA, USA {zrwu, xbhu}@psu.edu
*Corresponding author.

and sudden disappearance of topology priors across frames. These limitations necessitate fusion models that can opportunistically exploit marker information when reliable, while remaining robust to uncertainty and intermittent availability.

To address these challenges, we present a Confidence-Gated Marker Fusion framework. Our method includes: (1) Confidence-Gated Marker Fusion which introduces a token-level early fusion mechanism that dynamically weights marker-derived features against vision-based BEV features to handle spatial uncertainty. (2) Temporal Marker Memory which caches reliable marker priors across frames, propagating them forward to smooth inconsistencies and bridge gaps when detections are missing. Together, these modules enable robust and opportunistic use of fiducial markers, significantly improving lane topology reasoning under occlusion. Unlike SMERF (SD map, global priors) and TopoNet (vision-only reasoning), our method introduces localized, infrastructure-based priors that opportunistically fill blind spots without requiring global maps.

Our contributions are summarized as follows:

- We highlight fiducial markers as a novel source of infrastructure-supported topology priors, extending their use beyond robotics and indoor localization to outdoor traffic environments.
- We propose a marker fusion framework that enhances BEV-based topology reasoning with mechanisms explicitly designed for the spatial uncertainty and temporal inconsistency of fiducial marker detections.
- We construct and evaluate on a marker-augmented OpenLane-V2 benchmark, demonstrating significant robustness improvements over state-of-the-art vision-only and map-assisted methods.

II. Related Work

A. Lane Topology Prediction

Accurate lane topology prediction is a core task for autonomous driving. Early methods used BEV segmentation and vectorization pipelines, such as HDMapNet [5] and VectorMapNet [6], which predict dense or polyline-based maps from multi-view cameras. These approaches improve structured map prediction but rely on post-processing and do not explicitly model lane-to-lane relationships. Transformer-based methods, such as STSU [7] and MapTR [8], moved to end-to-end prediction of centerlines and their relationships, but connectivity is often handled by separate modules. TopoNet [1] improves this by introducing a Scene Graph Neural Network (SGNN) that propagates features among lane and traffic element queries and uses a class-specific knowledge graph. This explicit modeling of spatial and semantic relationships yields state-of-the-art results on the OpenLane-V2 benchmark.

B. Map-Assisted Topology Reasoning

Map priors can help models overcome occlusions and perception degradation at long ranges. High-Definition (HD) maps offer centimeter-level accuracy for road boundaries and lane graphs, but their high cost and frequent updates limit scalability [9]. SD maps, available for most road networks, provide road-level topology at lower cost and are easier to maintain. Luo et al. [2] proposed SMERF, a Transformer encoder that represents SD maps as polylines and fuses them with BEV features through cross-attention. SMERF improves the performance of lane topology models, including BEVFormer [10] and TopoNet, especially for far-away lanes and intersections where camera inputs are weak. However, SMERF depends on GNSS or LiDAR alignment to register SD maps, which can fail in urban canyons, tunnels, and other GNSS-denied areas.

C. Cooperative Perception and Fiducial Markers

Fiducial markers, such as AprilTag [3] and ArUco [4], are machine-readable visual codes that enable robust detection and direct 6-DoF pose estimation. They are lightweight, reliable, and low-cost, but their data capacity is limited, serving mainly as unique identifiers linked to external map entries. Most prior work employs markers to support localization. They are widely used in visual-inertial odometry and SLAM [11] to correct drift by providing absolute pose anchors. Similar strategies are common in robotics, where sparse markers stabilize ego-pose in GNSS-denied environments [12]. However, their potential as priors for vision-based perception in driving scenes has not been explored. Existing methods rely on markers mainly for ego-localization, whereas using them to supply structural lane topology priors and directly guide scene reasoning remains an open problem.

D. Attention and Fusion Mechanisms

Transformers have become central to 3D perception, especially for integrating visual and geometric cues. DETR3D [13] connects 2D camera features with sparse 3D predictions, removing the need for non-maximum suppression (NMS) or rule-based label assignment. BEVFormer [10] extends this by projecting multi-view image features into dense BEV representations. PETR [14] incorporates 3D positional encodings to enhance geometric reasoning, while multi-modal fusion frameworks such as FUTR3D [15] integrate multiple sensor modalities via cross-attention. These designs show that attention-based fusion is effective for combining heterogeneous inputs. Beyond Transformers, graph-based propagation, as used in TopoNet [1], enables interaction between lane and traffic element queries. Dynamic gating mechanisms [16] provide an adaptive way to weight features based on confidence or spatial relevance. Building on these ideas, we propose a confidence-gated early fusion framework. By integrating marker-derived topology priors into BEV encoding before lane candidate generation, our approach improves lane graph prediction under severe occlusion or degraded visual conditions while remaining reliable when markers are absent.

III. Methodology

Problem setup. We consider urban driving with synchronized multi-view cameras. Along selected road segments, e.g., intersections, merges/splits, or complex corridors, fiducial markers are installed on pavement or roadside assets.

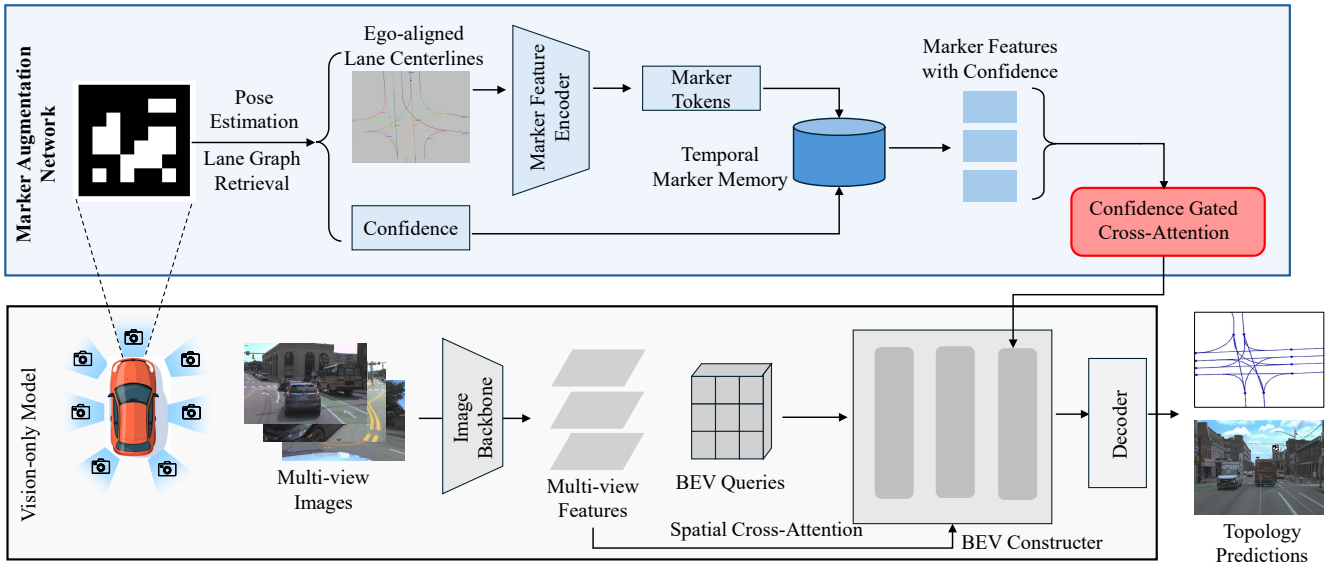

Fig. 2: Overview of our marker-augmented lane topology framework. Fiducial marker-derived lane centerlines are encoded into feature tokens with BEV anchors and a per-frame confidence, optionally aggregated by a temporal memory. These marker features are fused with image-derived features inside the BEVFormer encoder's deformable cross-attention layers, enhancing the BEV representation for robust lane geometry and topology prediction.

Each marker carries a unique ID that indexes a locally complete lane graph (lane centerlines and directed connectivity) stored in the vehicle's onboard database. When a marker is visible, the vehicle detects it in one or more cameras, estimates its 6-DoF pose, retrieves the corresponding local topology, and transforms it into the ego/BEV frame to form a localized topology prior. Because markers are sparse and may be intermittently visible, and because long-range or oblique observations yield noisier poses, we treat marker information as an opportunistic prior to be fused with image evidence and propagated over time.

At each timestamp t, given synchronized images $\{\mathcal{I}_t^{(k)}\}_{k=1}^{K}$ and an optional marker prior Φ_t (possibly empty), our goal is to predict a structured lane graph $\hat{\mathcal{G}}_t = (\hat{\mathbf{L}}_t, \hat{\mathcal{E}}_t)$ within a fixed BEV range. Here, $\hat{\mathbf{L}}_t$ denotes a set of 3D lane centerlines, each represented as a polyline with P resampled points for supervision, and $\hat{\mathcal{E}}_t$ denotes directed topological relations (e.g., continuation, merge/split) among these lanes. We denote by $\mathbf{H}_t$ the BEV latent features produced by a multi-view encoder; when available, marker-derived features augment $\mathbf{H}_t$, and a graph reasoning head decodes $\mathbf{H}_t$ into $\hat{\mathcal{G}}_t$.

For frames with a visible marker, standard PnP or planar homography yields the marker pose relative to the ego frame. The decoded ID queries an onboard store to obtain a local lane graph $\mathbf{L}_t$ aligned to the marker. We attach a per-frame confidence to the resulting prior based on detection score, viewing distance, and angular incidence. While the retrieved topology is structurally complete, its absolute placement can be biased by pose noise; fusion should therefore preserve structure while correcting geometry. The system must produce $\hat{\mathcal{G}}_t$ when Φ_t is absent (no marker), uncertain (noisy pose), or temporally inconsistent (occlusion/sparse visibility across frames).

Figure 2 shows the overall framework. Multi-view images are processed by a BEVFormer-based encoder to produce a latent BEV representation. When markers are detected, their lane graphs are converted into feature tokens with an associated confidence score. These tokens are integrated with image features via a confidence-gated fusion module (Sec. III-B), which down-weights unreliable priors. To handle missing or inconsistent detections, a temporal marker memory (Sec. III-C) caches recent marker features and decays their influence over time. The fused BEV features are then decoded into a lane graph with predicted centerlines and connectivity.

A. Encoding Representation from Markers

We use a structured encoding approach to utilize information from markers, which contain multiple polylines representing lane centerlines. The goal is to convert this set of geometric inputs into a fixed-size feature vector that can be fused with BEV features in the main perception backbone.

Each marker provides L polylines, where the ℓ-th polyline consists of a sequence of P_ℓ points $\{\mathbf{p}_i^{(\ell)}\}_{i=1}^{P_\ell}$ in 3D space. We first project each point onto the BEV plane by dropping the vertical coordinate. To reduce computation and unify input size, each polyline is uniformly subsampled to K points using a fixed stride. This keeps the number of points manageable while preserving the overall shape and direction of each polyline.

To convert each point into a high-dimensional representation that reflects its position, we use a fixed sinusoidal embedding [17]. This choice enables the model to capture both low-frequency global shape and high-frequency local

curvature of the polyline. Importantly, sinusoidal embeddings are sensitive to positional differences, which enhances the model's ability to distinguish between geometric variations. For each coordinate p and dimension index j, we define:

$$E(p)_{2j} = \sin\left(\frac{p}{T^{2j/d}}\right), \quad E(p)_{2j+1} = \cos\left(\frac{p}{T^{2j/d}}\right) \tag{1}$$

where d is the number of frequency bands per axis and T is a temperature constant. This encoding captures both low-frequency and high-frequency information and does not require any additional parameters or training. All point coordinates are globally normalized to a fixed BEV range and then scaled to $[0, 2\pi]$ before encoding, ensuring consistent spatial alignment across all polylines.

We use a Transformer encoder [17] to model the spatial relationships between the sampled points of lane centerlines. The input to the encoder is the sequence of point embeddings. The encoder consists of M layers of self-attention and feedforward networks. It allows the feature at each point to access information from other points in the sequence, which helps capture the overall shape and structure of lane centerlines.

After encoding, we flatten the sequence of outputs into a single vector. Finally, we use a linear layer to project this flattened vector to a fixed feature dimension C that aligns with the BEV feature.

The resulting vector $\mathbf{F}_{\text{marker}}$ is used as the feature representation for this marker in the cross-attention fusion module. This design keeps the input and output dimensions fixed and allows the model to learn from the full spatial context of each marker.

B. Marker Fusion under Spatial Uncertainty

Building on the marker representations, we now address how to fuse them into BEV-based lane topology reasoning in a uncertainty-aware manner. While markers provide structured lane priors that are locally accurate and semantically rich, their detections are not always equally reliable. In practice, fiducial markers may be missed due to occlusion, affected by distance-related degradation, or exhibit temporal jitter in pose estimation. Directly fusing such observations without accounting for their varying quality can introduce noise and even harm downstream reasoning. To address this, we design a marker fusion module that explicitly incorporates uncertainty, ensuring that BEV features are selectively enriched by trustworthy marker cues while down-weighting unreliable ones. This uncertainty-aware perspective is crucial for robust performance in realistic driving environments, where both perception noise and partial observability are inevitable.

Let $\mathbf{F}_{\text{bev}} \in \mathbb{R}^{B\times N\times C}$ represent the BEV feature tokens at a given Transformer encoder layer, where B denotes the batch size, N the number of BEV queries (spatial grid locations), and C the embedding dimension. Concurrently, we encode the local topology graph extracted from fiducial markers into a set of feature tokens $\mathbf{F}_{\text{marker}} \in \mathbb{R}^{B\times P\times C}$, where P is the number of positional embeddings of polylines represented per scene. These marker features are obtained via a Transformer encoder that preserves geometric and positional attributes, and are spatially aligned to the BEV frame.

To modulate the contribution of marker features, we introduce a confidence-aware gating mechanism. Confidence is derived from two observable reliability cues: (i) the detection distance between the ego-vehicle and the marker, and (ii) the short-term motion smoothness of the marker polyline across frames. The first cue is motivated by empirical observations that spatial localization noise typically grows with distance [18], making faraway markers less reliable. The second cue addresses temporally inconsistent detections: when a marker exhibits sudden jumps or discontinuities across frames, its pose estimation is unstable and should be down-weighted. Formally, we define the confidence of a marker as

$$c = \exp\left(-\frac{r}{\tau_r}\right) \cdot \exp\left(-\frac{d}{\tau_d}\right), \tag{2}$$

where τ_r and τ_d are scale parameters controlling sensitivity to residual instability and detection range. This formulation ensures that nearby reliable and temporally stable markers are assigned higher confidence, while distant or noisy markers are down-weighted. Note that most fiducial marker detection pipelines already provide a raw detection score (e.g., a classification probability or decoding Hamming distance), which can serve as a supplementary indicator of visual detection quality.

We incorporate marker information through a cross-attention mechanism that allows BEV queries to selectively attend to marker priors. Specifically, we define the fusion as:

$$\mathbf{F}'_{\text{bev}} = \text{LayerNorm}(\mathbf{F}_{\text{bev}} + c \cdot \text{Attn}(\mathbf{F}_{\text{bev}}, \mathbf{F}_{\text{marker}}, \mathbf{F}_{\text{marker}})), \tag{3}$$

where $\text{Attn}(\cdot)$ denotes multi-head self-attention with BEV tokens as queries, and marker tokens as both keys and values. The residual structure ensures that the original BEV representation is preserved and only enriched by marker-guided signals. This design enables the network to rely more heavily on stable, nearby marker priors, while suppressing contributions from unreliable detections.

By grounding marker confidence in distance and temporal smoothness, the model avoids over-reliance on noisy priors and adaptively balances vision-based reasoning with infrastructure cues. This confidence-gated fusion is applied in each encoder layer, allowing progressive refinement of BEV features with reliable marker guidance.

C. Marker Memory under Temporal Inconsistency

Although fiducial marker–derived lane priors provide structurally complete centerlines, they are not always available in every frame due to occlusions, sensor failures, or partial field of view. To ensure consistent topological guidance across frames, we introduce a Temporal Marker Memory module that stores and reuses recently observed lane tokens, enabling the network to propagate structural information forward in time, even in the absence of current marker detections.

At each time step t, the marker-derived lane tokens $\mathbf{F}_t \in \mathbb{R}^{N_t\times D}$, together with a scalar confidence score $c_t \in [0,1]$, are

stored in the form $\mathscr{M}_t = (\mathbf{F}_t, c_t, t)$, where N_t is the number of visible lanes, and D is the token dimension. A memory bank $\mathscr{B}_t = \{\mathscr{M}_{t-i}\}_{i=0}^{T-1}$ maintains the most recent T observations.

To control the temporal influence of past frames, each stored confidence is decayed exponentially with its age $\hat{c}_{t-i} = c_{t-i} \cdot \gamma^i, \gamma \in (0,1)$, where γ is a temporal decay factor.

Instead of computing a weighted average, we use confidence-scaled concatenation to aggregate lane tokens from multiple frames. This design retains fine-grained structure from each time step and supports variable token counts per frame:

$$\bar{\mathbf{F}}_t = \bigcup_{i=0}^{T-1} \left(\hat{c}_{t-i}^{\alpha} \cdot \mathbf{F}_{t-i}\right), \tag{4}$$

where $\alpha \in (0,1]$ is a scaling factor and the union denotes row-wise concatenation along the token dimension. If no valid prior frames exist, we set $\bar{\mathbf{F}}_t = \varnothing$.

When a new marker is observed, the encoded tokens $(\mathbf{F}_t, c_t)$ are pushed into the memory bank. If no marker is detected or the derived confidence is lower than a threshold ε in the current frame, the model queries the memory bank and retrieves $\bar{\mathbf{F}}_t$ as the guidance source. These temporally aggregated tokens are fused into the BEV encoder via the cross-attention mechanism described in Sec. III-B.

This mechanism enables the model to maintain coherent lane topology reasoning even across frames with missing marker observations. Confidence decay and token-level fusion ensure that outdated priors are attenuated while structurally informative markers continue to enhance perception under real-world conditions.

IV. Experiments

A. Dataset Augmentation

We conduct experiments on the OpenLane-V2 dataset [19], a large-scale benchmark designed for structured scene understanding in autonomous driving. OpenLane-V2 provides 3D lane centerlines, traffic elements (e.g., traffic lights and signs), and their topology relationships, making it well-suited for lane graph prediction. All experiments are performed on the primary evaluation subset (Subset A), which is built upon Argoverse 2 scenes and includes multi-camera images at 2 Hz. Following the standard evaluation protocol, we consider perception and topology reasoning within a BEV range of ± 50 m longitudinally and ± 25 m laterally from the ego vehicle.

To simulate realistic infrastructure-based priors, we inject synthetic fiducial markers and their associated local topology graphs. Each scene is assigned a virtual ArUco marker of fixed size (1×1 m) near complex road geometries (e.g., intersections), assumed to be visible only as the vehicle approaches frontally. Marker pose noise and detection rates are calibrated against empirical measurements reported by Jurado-Rodriguez et al. [18].

To emulate spatial uncertainty, we perturb ground-truth lane centerlines with distance-dependent translation and orientation noise. Specifically, a forward translation error along the ego-vehicle axis is sampled from a zero-mean Gaussian

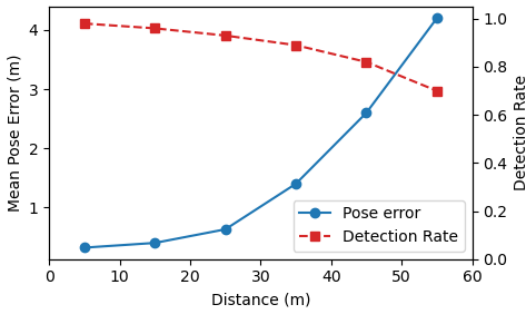

Fig. 3: Simulated marker pose error and detection probability as a function of distance. The trends align with empirical measurements [18].

$\Delta x \sim \mathscr{N}(0, \sigma_t^2)$, with variance σ_t^2 that increases with detection distance, reaching up to 3 m standard deviation at 60 m. Orientation errors are similarly sampled as yaw and pitch deviations $\theta \sim \mathscr{N}(0, \sigma_\theta^2)$, with σ_θ corresponding to 2° at 60 m. These settings are calibrated against empirical error–distance observations of ArUco markers in [18] and reflect how marker-based localization degrades as detection distance increases. We rescale and convert the reported results to the driving scenario by accounting for differences in scene scale, camera intrinsics, and image resolution in OpenLane-V2. In addition to such persistent errors, we also introduce occasional abrupt pose jumps, where with a small probability ($p = 0.05$) the estimated pose is shifted by a large offset, such as $\sim$25 m in translation and $\sim 90^\circ$ in yaw. These jumps mimic failure cases caused by reflections, glare, or unstable PnP solutions, forcing the model to remain robust under both continuous Gaussian noise and rare but extreme instabilities.

Finally, intermittent visibility is modeled with a distance-dependent detection probability $P_{\text{detect}}(d) = \exp\left(-\frac{d^2}{2\sigma_d^2}\right)$, with σ_d also fitted from empirical measurements in [18]. For each frame, a Bernoulli trial determines whether marker information is injected, producing temporally sparse and noisy priors. This setting allows our model to learn robustness against both gradual errors and sudden instabilities in marker perception.

B. Evaluation Metrics

We adopt the official metrics provided by the OpenLane-V2 benchmark to evaluate both perception quality and topology reasoning performance. The evaluation is conducted within a BEV region of $[-50\,\text{m}, +50\,\text{m}]$ along the longitudinal axis and $[-25\,\text{m}, +25\,\text{m}]$ laterally with respect to the ego-vehicle.

1) Perception Metrics: DET_ℓ denotes the mean average precision (mAP) for lane centerline detection, based on Fréchet distance [20] thresholds $\{1.0, 2.0, 3.0\}$ meters. DET_t is the mAP for traffic element detection across multiple attributes of traffic elements.

2) Topology Metrics: $\text{TOP}_{\ell\ell}$ evaluates the mAP of directed edge predictions between lane centerlines, while

$\text{TOP}_{\ell t}$ measures the accuracy of lane-to-traffic element associations. As our method focuses on enhancing the perception and structural inference of lane topology, we primarily focus on improvements on DET_ℓ and consequent boost on $\text{TOP}_{\ell\ell}$, rather than traffic element perception.

3) Overall Score: The consolidated OpenLane-V2 Score (OLS) averages all four metrics with square-root scaling on the two topology terms:

$$\text{OLS} = \frac{1}{4}\left(\text{DET}_\ell + \text{DET}_t + \sqrt{\text{TOP}_{\ell\ell}} + \sqrt{\text{TOP}_{\ell t}}\right). \quad (5)$$

C. Implementation Details

Our framework builds upon the TopoNet architecture [1] incorporating a shared BEV Transformer encoder followed by a scene graph reasoning module. We adopt a ResNet-50 [21] backbone pretrained on ImageNet, and extract multi-scale features from S8×, S16×, and S32× stages using an FPN [22]. These features are projected to a BEV grid (200 × 100) covering a 100m × 50m region centered on the ego-vehicle using cross-attention. The BEV features are then processed by a Deformable DETR-style decoder to produce centerline and traffic element queries. For topology reasoning, we employ a Scene Graph Neural Network (SGNN) [1] following the original TopoNet, which refines lane and traffic element queries via feature propagation on heterogeneous graphs.

For marker encoding, we uniformly subsample $K = 11$ points per polyline and apply sinusoidal positional embeddings with $d = 32$ frequency bands and a temperature constant $T = 1000$. The resulting point embeddings are processed by a 6-layer Transformer encoder with 4 attention heads, and the outputs are flattened and projected to obtain fixed-length marker tokens. BEV features are extracted with the same embedding dimension $C = 256$.

For confidence estimation, we set the distance scale to $\tau_d = 40\,\text{m}$ and the temporal residual scale to $\tau_r = 10\,\text{m}$, reflecting typical spatial degradation and motion jitter ranges. Cross-attention fusion is applied in every encoder layer after the original spatial cross-attention.

We train the model for 24 epochs using AdamW with a cosine learning rate schedule and initial learning rate of 1×10^{-4}. Input images are resized to 2048 × 1550, and standard augmentations such as random resizing and color jitter are applied. Batch size is set to 1 using 1 RTX 6000 Ada.

All experiments follow the official OpenLane-V2 evaluation toolkit for computing DET and TOP scores.

D. Baselines

To evaluate the performance of our proposed method, we compare against several baselines that represent the state of the art in vision-based and prior-assisted lane topology reasoning.

1) BEVFormer-DeTR: This baseline model is released with the OpenLane-V2 benchmark as the official "baseline large" and is built upon BEVFormer [10]. It encodes multi-view images into BEV features using spatiotemporal transformers and predicts lane centerlines and traffic elements using a deformable DETR-style decoder. While it achieves solid detection performance, it lacks explicit topology modeling and thus struggles with lane connectivity reasoning under occlusion.

2) TopoNet: TopoNet [1] is a vision-only state-of-the-art method that extends the BEVFormer backbone with a scene graph neural network (SGNN) to explicitly model pairwise relationships among lane centerlines and traffic elements. It introduces a deformable decoder for lane and object queries, followed by a reasoning module that refines features over heterogeneous graphs. This model serves as the primary baseline for evaluating the benefit of introducing topology priors, as our method builds directly upon its architecture.

3) TopoNet + SMERF: This baseline augments TopoNet with the SMERF [2] framework, which integrates Standard Definition (SD) maps as global topology priors using a Transformer encoder. The SD map features are fused into BEV features through multi-head cross-attention after each spatial cross-attention layer. While SMERF significantly improves reasoning in occlusion-heavy scenarios, it relies on precise GNSS pose for global alignment and assumes high-quality SD map availability, limiting its applicability in GNSS-denied environments.

E. Prediction Performance

We compare our marker-based topology reasoning framework against three baselines including BEVFormer-DeTR, TopoNet, and TopoNet + SMERF. Table I summarizes the quantitative performance across four metrics related to lane perception: DET_ℓ, $\text{TOP}_{\ell\ell}$, $\text{TOP}_{\ell t}$, and the consolidated OpenLane-V2 Score (OLS).

TABLE I: Comparison of Prediction Performance on OpenLane-V2.

Method	$\text{DET}_\ell \uparrow$	$\text{TOP}_{\ell\ell} \uparrow$	$\text{TOP}_{\ell t} \uparrow$	OLS ↑
BEVFormer-DeTR	17.0	2.3	16.2	30.2
TopoNet	28.5	4.1	20.6	34.7
TopoNet + SMERF	33.4	7.5	**23.4**	39.4
Ours	**36.2**	**8.8**	22.6	**40.5**

Compared to the visual-only baselines, our method yields substantial gains in both lane centerline detection (DET_ℓ) and lane-to-lane topology prediction ($\text{TOP}_{\ell\ell}$). This demonstrates that fiducial marker priors provide strong complementary information to BEV features, especially under occlusions or visually ambiguous conditions. The improvement over TopoNet shows that purely vision-based topology reasoning remains limited when visual features are degraded, while marker priors offer reliable localized structure.

Compared to TopoNet + SMERF, our model achieves slightly better performance in DET_ℓ and $\text{TOP}_{\ell\ell}$, while remaining comparable in lane-to-element association ($\text{TOP}_{\ell t}$). Since our approach does not enhance traffic element modeling, the small decrease in $\text{TOP}_{\ell t}$ is expected and not significant, as the fusion may introduce additional interference. Overall, our model achieves the highest OpenLane-V2 Score, indicating a favorable trade-off between perception accuracy and structural reasoning.

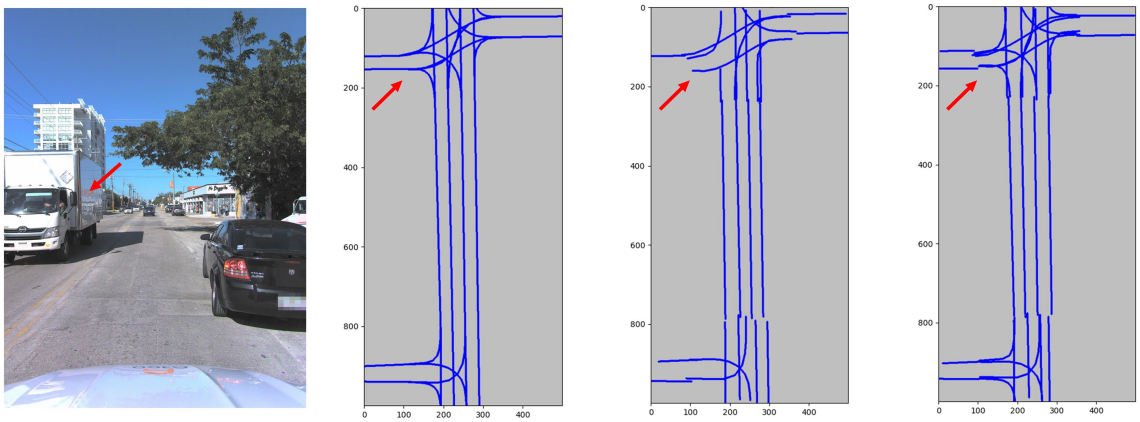

Fig. 4: Qualitative comparison under occlusion. In this example, large vehicles and roadside vegetation obscure multiple lanes, leading to incomplete topology predictions for the vision-only baseline, which misses several centerlines and yields fragmented connectivity. By contrast, our confidence-gated marker fusion recovers the occluded lanes and preserves correct lane-to-lane topology, demonstrating robustness in challenging occlusion scenarios.

F. Qualitative Analysis

Figure 4 illustrates qualitative comparisons under heavy occlusion scenarios. In the input frame, large vehicles and roadside vegetation obscure multiple lane markings, leading to incomplete topology predictions for vision-only baselines. As shown, TopoNet fails to recover several occluded centerlines and produces fragmented connectivity. In contrast, our confidence-gated marker fusion successfully completes the occluded lanes and preserves the correct topology structure by leveraging marker-derived priors. This qualitative evidence aligns with our quantitative results: marker priors are most beneficial in structurally missing regions where visual cues vanish, allowing our method to maintain lane continuity and connectivity reasoning even under severe occlusions.

G. Ablation Study

We perform ablations to quantify the contributions of confidence-gated fusion, temporal memory and marker representation encoding. Results are reported on the marker-augmented OpenLane-V2 benchmark.

TABLE II: Ablation on confidence-gated fusion.

Variant	DET_ℓ	$\text{TOP}_{\ell\ell}$	DET_t	$\text{TOP}_{\ell t}$	OLS
Uniform Fusion	23.3	3.9	47.8	18.8	33.6
Hard Masking	26.5	4.5	47.1	18.6	34.5
Confidence-Gated	36.2	8.8	48.5	22.6	40.5

Effect of Confidence-Gated Fusion. As reported in Table II, naïve uniform fusion performs poorly, since treating all marker priors equally introduces significant noise and harms lane centerline detection and topology reasoning. Hard masking improves robustness by discarding low-confidence priors, but remains brittle when confidence is underestimated. Our proposed confidence-gated fusion achieves the best performance across all metrics, demonstrating that adaptively balancing marker and visual features according to confidence is essential for exploiting marker priors without being misled by noisy detections.

TABLE III: Ablation on temporal marker memory.

Variant	DET_ℓ	$\text{TOP}_{\ell\ell}$	DET_t	$\text{TOP}_{\ell t}$	OLS
No Memory	26.5	6.4	46.7	21.0	36.1
Memory w/o Decay	34.3	8.8	48.9	22.3	40.0
Full Memory	36.2	8.8	48.5	22.6	40.5

Effect of Temporal Marker Memory. We then ablate the effect of temporal marker memory. As shown in Table III, removing the memory mechanism leads to a noticeable drop in performance, since the model cannot retain marker-derived priors once they are occluded. Introducing memory without decay substantially improves robustness, but the absence of confidence decay risks propagating stale or misaligned marker information. Our full design with decay achieves the best overall accuracy, indicating that temporal caching is crucial for handling intermittent marker availability while decay mechanisms prevent over-reliance on outdated priors.

TABLE IV: Ablation on marker representation encoding.

Variant	DET_ℓ	$\text{TOP}_{\ell\ell}$	DET_t	$\text{TOP}_{\ell t}$	OLS
MLP	20.5	3.2	45.1	19.2	31.8
Transformer	28.2	7.2	47.9	19.8	36.9
Transformer + PE	36.2	8.8	48.5	22.6	40.5

Effect of Representation Encoding. Table IV compares different encoding strategies for marker polylines. An MLP over raw point coordinates yields the lowest performance, suggesting that point-wise encoding is insufficient to capture polyline structure. Replacing the MLP with a Transformer improves all metrics, demonstrating the benefit of modeling point-wise interactions. Adding sinusoidal positional encoding further leads to the best performance, indicating that explicit spatial encoding helps the model better preserve layout and continuity of lane centerlines.

V. Discussion and Future Work

Our results show that incorporating fiducial marker priors via confidence-gated fusion significantly improves lane centerline detection and lane-to-lane topology prediction, particularly under occlusion conditions. The injected marker priors provide strong structural constraints that effectively complement visual features when local BEV evidence is incomplete. The proposed temporal marker memory further enhances robustness by propagating high-confidence priors across frames, enabling stable topology reasoning despite intermittent detections. Compared with global SD map-based methods such as SMERF, our approach avoids reliance on GNSS alignment or map registration, making it suitable for GNSS-denied or construction-heavy environments where global maps may be outdated or unavailable. In practice, markers can be sparsely deployed at structurally complex regions, offering a low-cost, passive alternative to dense HD map coverage or vehicle-to-everything (V2X) infrastructure.

While our design demonstrates clear benefits, several avenues remain for improvement. First, our current framework only exploits marker priors for lane centerline geometry, without incorporating their potential to provide explicit lane-to-lane connectivity. Future work may extend marker encoding to capture both geometry and connectivity, enabling more direct supervision of graph structure. Second, tighter integration between marker detection and topology prediction represents a promising direction. Rather than a two-stage design where marker pose estimation and graph retrieval are separate from the perception network, an end-to-end framework could jointly learn marker localization, pose refinement, and graph encoding together with BEV-based reasoning. Such integration would allow dynamic reasoning over both visual and symbolic inputs, further strengthening robustness in challenging conditions.

VI. Conclusion

We presented a Confidence-Gated Marker Fusion framework for lane topology perception that opportunistically integrates fiducial marker priors with BEV-based reasoning. By explicitly addressing spatial uncertainty through confidence-gated fusion and temporal inconsistency through a memory mechanism, our method achieves robust and structurally complete lane graphs under occlusion conditions. Experiments on a marker-augmented OpenLane-V2 benchmark demonstrate consistent improvements over vision-only and SD map-assisted baselines. These results highlight fiducial markers as a promising infrastructure-supported prior for resilient lane topology prediction, and open new directions for joint learning of marker-based and vision-based perception.

References

[1] T. Li, L. Chen, H. Wang, Y. Li, J. Yang, X. Geng, S. Jiang, Y. Wang, H. Xu, C. Xu, J. Yan, P. Luo, and H. Li, "Graph-based Topology Reasoning for Driving Scenes," Aug. 2023. arXiv:2304.05277 [cs].

[2] K. Z. Luo, X. Weng, Y. Wang, S. Wu, J. Li, K. Q. Weinberger, Y. Wang, and M. Pavone, "Augmenting Lane Perception and Topology Understanding with Standard Definition Navigation Maps," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4029–4035, May 2024.

[3] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *2011 IEEE International Conference on Robotics and Automation*, pp. 3400–3407, May 2011. ISSN: 1050-4729.

[4] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[5] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An Online HD Map Construction and Evaluation Framework," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 4628–4634, May 2022.

[6] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "VectorMapNet: End-to-end Vectorized HD Map Learning," in *Proceedings of the 40th International Conference on Machine Learning*, pp. 22352–22369, PMLR, July 2023. ISSN: 2640-3498.

[7] Y. B. Can, A. Liniger, D. P. Paudel, and L. 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.

[8] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction," Jan. 2023. arXiv:2208.14437 [cs].

[9] R. Liu, J. Wang, and B. Zhang, "High Definition Map for Automated Driving: Overview and Analysis," *The Journal of Navigation*, vol. 73, pp. 324–341, Mar. 2020.

[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, pp. 2020–2036, Mar. 2025.

[11] R. Muñoz-Salinas and R. Medina-Carnicer, "UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers," *Pattern Recognition*, vol. 101, p. 107193, May 2020.

[12] F. Wang, Y. Zou, C. Zhang, J. Buzzatto, M. Liarokapis, E. del Rey Castillo, and J. B. Lim, "Uav navigation in large-scale gps-denied bridge environments using fiducial marker-corrected stereo visual-inertial localisation," *Automation in Construction*, vol. 156, p. 105139, 2023.

[13] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries," in *Proceedings of the 5th Conference on Robot Learning*, pp. 180–191, PMLR, Jan. 2022. ISSN: 2640-3498.

[14] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position Embedding Transformation for Multi-view 3D Object Detection," in *Computer Vision – ECCV 2022* (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), (Cham), pp. 531–548, Springer Nature Switzerland, 2022.

[15] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 172–181, 2023.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] D. Jurado-Rodriguez, R. Muñoz-Salinas, S. Garrido-Jurado, and R. Medina-Carnicer, "Planar fiducial markers: A comparative study," *Virtual Reality*, vol. 27, no. 3, pp. 1733–1749, 2023.

[19] H. Wang, T. Li, Y. Li, L. Chen, C. Sima, Z. Liu, B. Wang, P. Jia, Y. Wang, S. Jiang, *et al.*, "Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18873–18884, 2023.

[20] T. Eiter, H. Mannila, *et al.*, "Computing discrete fréchet distance," 1994.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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