Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31472

UniBEV-Lane: A Unified End-to-End Framework for Multi-Modal 3D Lane Detection

Fuzhou Deng[1], Hongyi Chen[1*]

[1]College of Physics and Optical Engineering, Shenzhen University, Shenzhen, China
**Corresponding Author. Email: chenhy@szu.edu.cn*

Abstract. Accurate perception of three-dimensional lane is the key to realize automatic driving, but it faces two major challenges: monocular camera is difficult to accurately restore depth, and lidar point cloud is often sparse and discontinuous in long-distance or complex scenes. In addition, the current mainstream query driven methods usually rely on static anchors, which making difficult to flexibly model the complex topology common in real roads. To solve these problems, we propose UniBEV-lane, an end-to-end multimodal 3D lane sensing system. The method maps the camera and lidar data to the Bird's-Eye-View (BEV) space, and cooperatively fuses semantic and geometric information under this shared representation. Its core is a new type of lane detection head based on Instance Activation Maps (IAM): it can automatically generate instance aware dynamic queries from high probability areas, so as to get rid of the restriction of fixed anchor points and flexibly adapt to various lane forms. Experiments on the OpenLane benchmark show that UniBEV-lane has achieved 57.1% F-score, which is 6.6% higher than the current leading pure vision methods (such as performer). At the same time, the average error in the depth direction (Z-axis) is only 0.09m, showing excellent spatial accuracy and multimodal coordination ability, and providing a reliable solution for high-precision 3D lane perception.

Keywords: 3D lane detection, multi-modal fusion, bird's-eye-view, autonomous driving

1. Introduction

Reliable 3D lane perception is fundamental to autonomous navigation, underpinning trajectory synthesis and HD map generation. Unlike 2D tasks, this necessitates rigorous spatial reasoning within the ego-vehicle frame. While monocular methods suffer from depth ambiguity and limited range—especially in adverse weather—LiDAR lacks the semantic texture required for unambiguous feature extraction.

Existing methods usually use fixed query without scene information to detect lanes, which leads to slow training and is error-prone when encountering complex or unseen lane shapes (such as forks and sharp turns). Our proposed UniBEV-Lane unifies camera and lidar information to be processed under a Bird's-Eye-View (BEV), and uses Instance Activation Maps (IAM) [1] to dynamically generate queries matching real lanes instead of rigid fixed anchors. Then, a deep fusion module is used to further optimize the semantic and geometric information, so that the system is still stable and reliable in complex scenes.

DOI: 10.54254/2755-2721/2026.31472

Our primary contributions are summarized as follows:

• We have developed UniBEV-Lane, a unified system that can simultaneously process camera and lidar data. It converts the 2D image features of the camera and the 3D point cloud of the lidar to the same BEV perspective, so that the information of the two sensors can be better fused, so as to identify the 3D lane more reliably.

• We propose an IAM-based Grouped Sparse Lane Head. Unlike the traditional method which relies on fixed and preset anchor points, it uses IAM to dynamically generate queries that match the actual Lane shape. In this way, the model can be more flexible to deal with complex lane structures, such as bifurcation or merging.

• Comprehensive experiments carried out on the OpenLane benchmark show that UniBEV-Lane reaches comparable F-Score (57.1%), and our approach outperforms vision-centric ones, especially Persformer [2], by a wide margin of 6.6%, while we also improve the localization accuracy at long-range regions compared with existing fusion-based approaches.

2. Literature review

2.1. Monocular 3D lane detection

In essence, it is a very difficult problem to restore the three-dimensional spatial structure from a single ordinary image, because a planar photo itself lacks depth information. Some important early methods, such as 3DLaneNet [1] and Gen-LaneNet [3], used Inverse Perspective Mapping (IPM) to convert image features into a BEV projection. But this method has a big premise: the ground must be completely flat. Once encountering non planar terrain such as uphill and downhill roads and undulating roads, the converted lane position will be seriously distorted.

In recent years, the 3D lane detection method has gradually turned to learning to realize the perspective conversion. A class of mainstream methods (such as Persformer [2] and BEV-LaneDet [4]) intelligently transfer image features to BEV space, and then fuse 2D and 3D information. Another method (such as Anchor3Dlane [5]) skips the BEV and predicts the lane directly with the preset 3D anchor on the original image. Some transformer-based models (such as CurveFormer++ [6] use curve propagation scheme to iterate 3D lane query.

2.2. Lidar and multi-modal fusion

Lidar can provide accurate 3D structure, but its point cloud is very sparse and lacks semantic information such as color. Although PointLaneNet [7] attempts to directly use lidar for lane detection for the first time, the current research trend is more inclined to integrate multimodal data to retain rich semantic details of images and accurate spatial geometric information of lidar.

In 3D object detection, methods such as BEVFusion [8] and TransFusion [9] have proved that the fusion effect of unifying multi-sensor data under BEV is very good. But the lane is different from ordinary objects, it is long and curved, so it is not appropriate to directly apply these general methods. Although M2-3DLaneNet [10] designed a multimodal fusion framework for lanes for the first time, it still relies on dense fixed anchors and is not flexible enough. Our method is different: the sparse and instance aware dynamic query mechanism can more accurately understand the complex lane structure from the integrated aerial view.

DOI: 10.54254/2755-2721/2026.31472

3. Methodology

3.1. Overview

As shown in Figure 1, the overall design of UniBEV-Lane is based on an improved process of first promotion, then fusion, and finally query. The system consists of three core parts: (1) Dual stream encoder: extract features from camera images and lidar data respectively, and convert them to a unified Bev perspective; (2) Bev fusion module: through a powerful network, the information from these two sources is deeply integrated; (3) Sparse lane detection head: use IAM to dynamically generate queries suitable for the current Lane shape, so as to flexibly and accurately predict the complete 3D lane structure.

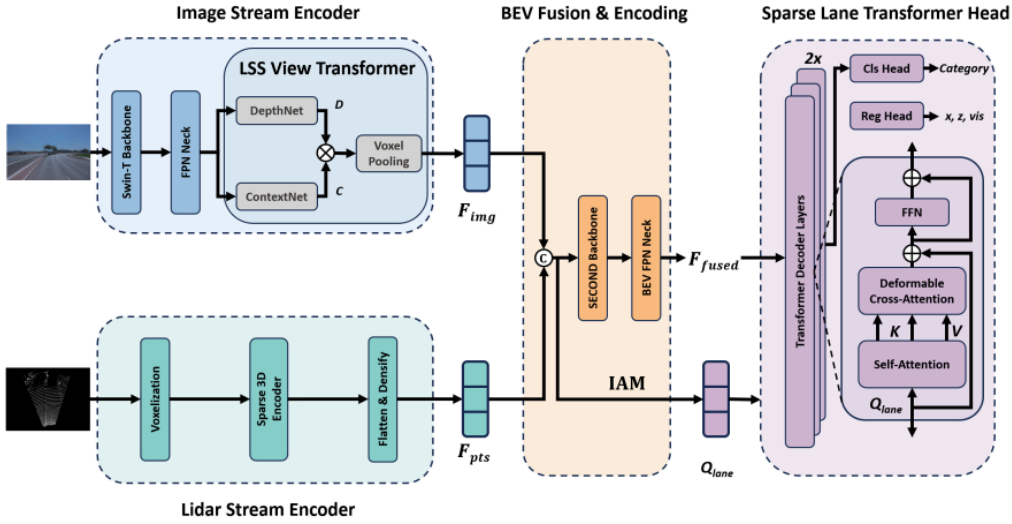

Figure 1. Overall architecture of the UniBEV-Lane framework

3.2. Dual-stream feature encoding

First, the input image is extracted with multi-scale features through the Swin-T [11] backbone and FPN network. Next, the LSS View Transformer [12] uses DepthNet and ContextNet to predict the depth D and semantic information C. Then, the system combines the two 2D information to lifts the image content to 3D space, and summarizes it into a 3D grid, and finally generates a camera-centric BEV feature F_{img}.

Input point clouds first go through voxelization and are used by a Sparse 3D Encoder (VoxelNet) [13], which will extract geometric features, which forms a 3D sparse tensor after processing, and then it is flattened and densified along the Z-axis and becomes the LiDAR BEV feature F_{pts}. F_{img} and F_{pts} have the same BEV resolution and coordinate system.

3.3. BEV fusion & encoding

In order to better integrate the semantic information of the image and the geometric information of the lidar, we first concatenated the aligned two BEV features in the channel dimension, as shown in Figure 1, and then optimized them with a SECOND Backbone [14] , and combined with the BEV FPN Neck to integrate the context information of different scales, and finally obtained the unified feature representation after fusion. The final synthetic expression of F_{fused} is:

DOI: 10.54254/2755-2721/2026.31472

$$F_{fused} = \mathscr{F}_{FPN}\left(\mathscr{F}_{SECOND}\left(Concat\left(F_{img}, F_{pts}\right)\right)\right) \tag{1}$$

In this formula, F_{FPN} and F_{SECOND} represent the operations performed by the neck module and the backbone module respectively.

3.4. Sparse lane transformer head

We abandoned the rigid fixed anchor point and designed a detection head that can dynamically focus on Lane instances. We use the IAM module for query initialization: this branch decodes F_{fused} to generate an activation map that accurately locates the foreground area with high confidence. Then, these maps aggregate the features from F_{fused} into instance aware embedding as the main lane queries Q_{lane}. The Q_{lane} queries advance into a Transformer Decoder structured with N = 2 layers. In each stage, the self-attention module coordinates the communication between queries, while the deformable cross attention unit obtains the context from the F_{fused} , regards it as Key K and Value V, and then processes the feedforward network (FFN).

Prediction heads receive the refined queries and feed them into 2 separate parallel MLPs:

• Cls Head: Predicts the lane's confidence score as well as its class.

• Reg Head: Regresses the 3D geometry (polynomial coefficients or offsets x, z) and point-wise visibility (vis).

3.5. Optimization objective

To facilitate end-to-end optimization, the framework employs a bipartite matching loss. The overarching objective, L_{total}, integrates the classification, regression, and auxiliary IAM components through the following synthesis:

$$\mathscr{L}_{total} = \lambda_{cls} + \lambda_{reg}\left(\mathscr{L}_x + \mathscr{L}_z + \mathscr{L}_{vis}\right) + \lambda_{aux}\mathscr{L}_{iam} \tag{2}$$

Here, L_{cls} use Focal Loss, while $L_{x/z}$ and L_{vis} use L_1 and Binary Cross Entropy for geometric and visibility monitoring, respectively. To enhance query initialization, L_{iam} provides an auxiliary training signal.

4. Experiments

In this section, we provide a detailed evaluation of the performance of our proposed method, namely UniBEV-Lane, including the dataset used in the experiment, the adopted evaluation metrics, and implementation details, and we further compare the performance of our approach with state-of-the-art approaches on the OpenLane benchmark.

4.1. Experimental setup

We evaluated UniBEV-Lane on OpenLane-V1 dataset. The dataset contains more than 200k images, covering a variety of complex scenes such as rainy days and nights. Following standard practice, we report F-Score (IoU ≥ 0.5) as well as lateral (X/Z-Error) in both near (0–40m) and far (40–100m) distances. It is implemented by using PyTorch combined with Swin-T [11] and VoxelNet [13]

DOI: 10.54254/2755-2721/2026.31472

backbones, the architecture runs on a BEV grid covering [-30m, 30m] × [3m, 103m] with a grid resolution of 0.5m. We use the AdamW optimizer [15] to perform 24 epochs of training in 798 scenes, and start at a learning rate of 2×10^{-4} learning rate with cosine decay—on a dual NVIDIA RTX 3080 Ti setup.

4.2. Results on OpenLane

We compare the methods for evaluating our UniBEV -Lane model against a broad suite of state-of-the-art models by putting them into the category of vision-centric baselines (e. g., Persformer [2], BEV-LaneDet [4]) and multimodal methods (e. g., M2-3DLaneNet [10]).

As shown in Table 1, our method achieved 57.1% F-score on OpenLane, which is significantly better than the early visual method Gen-LaneNet [3] (+24.6%) and formidable models like Persformer [2] (+6.6%). Although it is slightly lower than the BEV-LaneDet [4] (58.4%) with the best performance at present, UniBEV-Lane performs better in geometric positioning accuracy: the Z-Error (Far) is only 0.090m, which is far lower than the 0.631 m reported by BEV-LaneDet [4]. This shows that although our model may miss a small number of lanes in some difficult scenes, the lanes it successfully detects have higher 3D accuracy. Considering the sensitivity of downstream tasks such as trajectory planning to height estimation, we believe that this is a compromise worth making.

Table 1. Quantitative results on OpenLane dataset regarding F-score and localization errors

Method	F-Score (%)	X-Error (m) ↓		Z-Error (m) ↓	
		Near	Far	Near	Far
3D-LaneNet [1]	44.1	0.479	0.572	0.367	0.443
Gen-LaneNet [3]	32.5	0.591	0.684	0.411	0.521
Persformer [2]	50.5	0.485	0.553	0.364	0.431
CurveFormer [6]	32.5	0.452	0.556	0.497	0.491
Anchor3DLane [5]	50.5	0.276	0.311	0.107	0.138
DB3D-L [16]	53.7	0.254	0.682	0.202	0.621
M2-3DLaneNet [10]	55.2	0.283	0.256	0.078	0.106
BEV-LaneDet [4]	58.4	0.309	0.659	0.244	0.631
UniBEV-Lane (Ours)	57.1	0.323	0.334	0.072	0.090

For safe-critical tasks such as trajectory planning, we think accurate and reliable height estimation can be a proper trade-off, which would help draw wide attention to HED development. As shown in Figure 2 using qualitative experimental results derived from the OpenLane validation set. Our method could provide wide range of values through different cases such as different times (daytime/night), road surface condition(fog). The representation of model outputs meet the ground truth value on three views, such as image plane 2D lane projection view(top), BEV middle view, and ground 3D lane prediction(bottom). In all figures, the red lines represent ground truth and green lines represent our predictions. The close agreement of red and green lines indicates that our UniBEV-Lane has a strong ability to predict lane annotations with geometric consistency in complex topology or bad weather scenes despite illuminating and environmental condition variations.

DOI: 10.54254/2755-2721/2026.31472

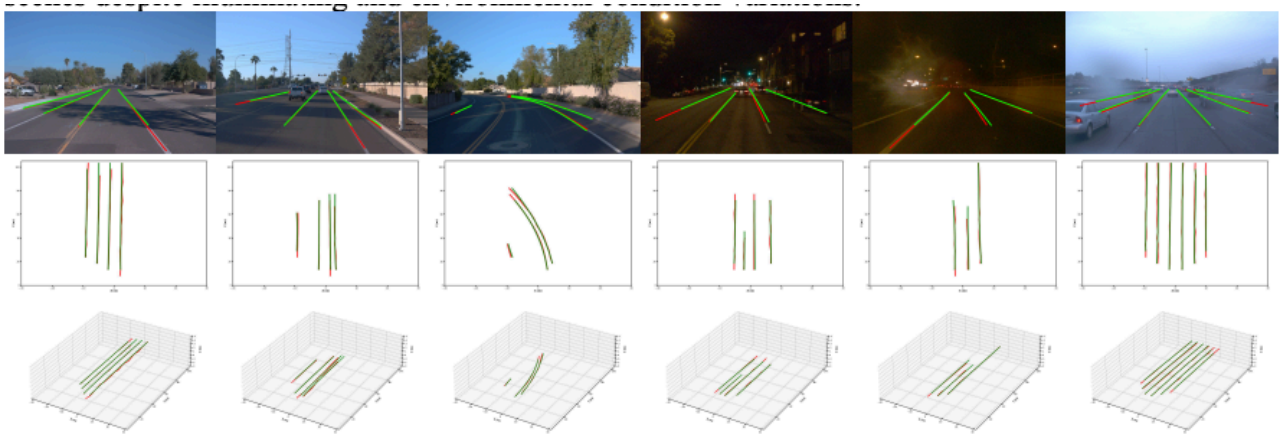

Figure 2. Qualitative results of UniBEV-Lane on the OpenLane validation set

5. Conclusion

This paper proposes UniBEV-lane, a unified architecture for multimodal 3D lane perception, which aims to bridge the difference between image semantics and lidar geometry. We integrate the camera and lidar features into a unified Bev space, and use the IAM based grouped sparse Lane detector to get rid of the limitations of the traditional static anchor design. Experiments on OpenLane benchmark show that this method not only achieves competitive performance in lane detection, but also performs well in spatial accuracy, especially in height (Z-axis) estimation. In future work, we plan to introduce time series information to improve prediction stability and explore the application potential of the system in downstream trajectory planning.

References

[1] Garnett, N., Cohen, R., Pe'er, T., Lahav, R. and Levi, D. (2019) 3D-LaneNet: End-to-End 3D Multiple Lane Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2921–2930.

[2] Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., Hongyang, L., Conghui, H., Jianping, S., Yu, Q. and Yan, J. (2022) Persformer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark. In European Conference on Computer Vision (ECCV).

[3] Guo, Y., Chen, G., Zhao, P., Zhang, W., Miao, J., Wang, J. and Choe, T.E. (2020) Gen-LaneNet: A Generalized and Scalable Approach for 3D Lane Detection. In European Conference on Computer Vision, pp. 666–681. Springer.

[4] Wang, R., Qin, J., Li, K., Li, Y., Cao, D. and Xu, J. (2023) BEV-LaneDet: An Efficient 3D Lane Detection Based on Virtual Camera via Key-Points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1002–1011.

[5] Huang, S., Shen, Z., Huang, Z., Ding, Z., Dai, J., Han, J., Wang, N. and Liu, S. (2023) 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.

[6] Bai, Y., Chen, Z., Fu, Z., Peng, L., Liang, P. and Cheng, E. (2022) CurveFormer++: 3D Lane Detection by Curve Propagation with Temporal Curve Queries and Attention. arXiv preprint arXiv: 2209.07989.

[7] Chen, Z., Liu, Q. and Lian, C. (2019) PointLaneNet: Efficient end-to-end CNNs for Accurate Real-Time Lane Detection. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 2563–2568.

[8] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L. and Han, S. (2023) BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 2774–2781. IEEE.

[9] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H. and Tai, C.L. (2022) TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1090–1099.

DOI: 10.54254/2755-2721/2026.31472

[10] Luo, Y., Yan, X., Zheng, C., Zheng, C., Mei, S., Kun, T., Shuguang, C. and Li, Z. (2022) Mˆ2 -3DLaneNet: Exploring Multi-Modal 3D Lane Detection. arXiv preprint arXiv: 2209.05996.
[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Stephen, L. and Guo, B. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[12] Philion, J. and Fidler, S. (2020) Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pp. 194–210. Springer.
[13] Zhou, Y. and Tuzel, O. (2018) VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4490–4499.
[14] Yan, Y., Mao, Y. and Li, B. (2018) SECOND: Sparsely Embedded Convolutional Detection. Sensors, 18(10), 3337.
[15] Loshchilov, I. and Hutter, F. (2018) Decoupled Weight Decay Regularization. In International Conference on Learning Representations.
[16] Liu, Y., Xu, X., Wang, Z. and Yao, Y. (2025) DB3D-L: Depth-aware BEV Feature Transformation for Accurate 3D Lane Detection. arXiv preprint arXiv: 2505.13266.