Freq-3DLane: 3D Lane Detection From Monocular Images via Frequency-Aware Feature Fusion

Yongchao Song, Jiping Bi, Lijun Sun, Zhaowei Liu, *Member, IEEE*, Yahong Jiang, and Xuan Wang, *Senior Member, IEEE*

***Abstract*— 3D lane detection provides richer spatial information than 2D lane detection planar position results. It improves vehicle perception in complex scenes, which is becoming increasingly important in intelligent driving. However, existing frameworks mainly focus on mapping front-view (FV) and bird's-eye view (BEV) features and ignore the intrinsic correlations between different perspectives and scales. It can lead to incomplete feature extraction, affecting the perception accuracy of lane detection and adaptation ability to complex scenes. To alleviate these problems, we present a novel Freq-3DLane framework, an efficient end-to-end 3D lane detector. Instead of directly superimposing deeper and lower-level features, we propose a strategy for multi-scale information integration that exploits the frequency characteristics of features for image feature extraction. To enhance perception, we fuse image features at each scale through frequency processing to ensure that detailed information and global structure are fully utilized. Next, spatial transformation fusion captures the association between the FV and the BEV feature at any two-pixel position of both, thus enabling view feature transformation. In addition, attentional guidance enhances the lane semantic information to ensure recovery of the lane geometry for accurate 3D lane detection. Extensive results on two challenging benchmarks (Apollo 3D Lane Synthetic, and OpenLane) show that our model performs favorably against the state of the arts.**

***Index Terms*— 3D lane detection, frequency aware, feature fusion, attention mechanism, intelligent driving.**

I. Introduction

AUTONOMOUS driving technology has become a hot topic in academia and industry [1]. Lane detection is one of the core tasks of autonomous driving systems (ADS). It uses advanced sensor technology to accurately identify the location and shape of lanes [2]. This process provides critical data for path planning, vehicle steering control, and lane keeping. However, lane detection faces many complex environmental factors in practical applications, such as light variations, road occlusion, and bad weather [3]. These factors often pose serious challenges to lane detection technology. To solve these problems effectively, researchers have made a lot of explorations and attempts in different dimensions [4].

In the past, lane detection was mainly regarded as a 2D image processing task. The core objective is to identify the lanes on the road through the image information acquired by the camera, thus helping ADS to understand the current road structure. With the advancement of deep learning technology, lane detection using methods like convolutional neural networks (CNN) has become widely adopted [5]. By training the deep neural network, the model can automatically extract and learn lane features from the image, thereby enhancing the accuracy and robustness of the detection [6], [7]. It is because lanes appear as lines with specific directions and widths in images. By analyzing the image, the lane location can be accurately recognized [8]. However, with the advancement of autonomous driving technology, simple 2D lane detection can no longer meet the needs of practical applications. To further ensure vehicle stability in complex environments, lane detection is gradually expanding from traditional 2D detection to 3D detection tasks [9]. This shift allows lane detection beyond planar lane markings and considers 3D spatial information about the lane, including position, width, curvature, and height [10]. As a result, the ADS obtains key information such as the distance and angle between the current and adjacent lanes, and the road structure. By combining lane boundaries, road gradients, and other data, ADS can accurately understand road spatial patterns to make safer driving decisions and avoid potential risks such as oversteering in corners or changing lanes too early [11].

While some approaches rely on LiDAR sensors or multi-sensor fusion, monocular cameras have become an increasingly popular option due to their low cost and ability to provide high-resolution visual information that allows for effective lane recognition [12]. The BEV-based approach is considered a breakthrough in 3D lane detection. It adapts the lane representation to the 3D domain by modeling vertical anchors or local segments within 3D space, oriented to match the aerial view [9]. Two challenges are involved: effective extraction of features in BEV space and accurate matching of the extracted features in 3D space. The BEV view provides overhead road information, which allows lanes to be represented more clearly as geometric shapes. However, lane geometry may vary significantly depending on road type, section curvature, and vehicle movement states [13].

Received 6 January 2025; revised 13 March 2025; accepted 24 April 2025. Date of publication 9 May 2025; date of current version 16 September 2025. This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF037, in part by the Graduate Innovation Foundation of Yantai University (GIFYTU), and in part by the School and Locality Integration Development Project of Yantai City in 2022. The Associate Editor for this article was V. Chamola. *(Corresponding authors: Jiping Bi; Yahong Jiang.)*

Yongchao Song, Jiping Bi, Lijun Sun, Zhaowei Liu, and Xuan Wang are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China (e-mail: ycsong@ytu.edu.cn; bijiping@s.ytu.edu.cn; sunlijun@s.ytu.edu.cn; lzw@ytu.edu.cn; xuanwang91@ytu.edu.cn).

Yahong Jiang is with the School of Transportation Engineering, Chang'an University, Xi'an 710064, China (e-mail: 2018023002@chd.edu.cn).

Digital Object Identifier 10.1109/TITS.2025.3565272

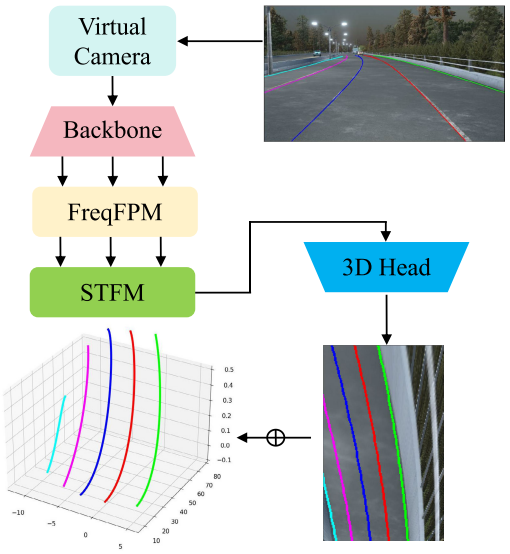

Fig. 1. End-to-end framework diagram of the Freq-3DLane process. First, the input images are fed into the backbone network to extract multi-scale features. Next, the frequency feature pyramid module and spatial transformation fusion module perform deep processing of these features to generate BEV features. Finally, the BEV features are passed to the 3D Head for prediction, and the results are further processed to obtain the 3D lane detection output.

Effectively dealing with the scale variations of lanes across different road sections, the lane detection model focuses on both the detailed parts (e.g., endpoints, bifurcations of lanes) and the global aspects (e.g., the overall shape of the lanes, curvature variations, etc.). Smaller scale lanes require fine-grained feature extraction, while larger scale lanes require the network to be able to capture the macroscopic road structure. Moreover, the raw data captured by sensors such as cameras usually have differences in viewpoints and coordinate systems, and converting them into a unified BEV view and effectively extracting spatial features is a difficult task [14]. Especially in complex urban environments, issues of sensor resolution, viewing angle differences, and noise may affect the accuracy of feature extraction. Bev-LaneDet [15] effectively solves this problem by introducing virtual camera technology, which ensures the consistency of the spatial relationship of the cameras and reduces data distribution discrepancies. In 3D lane detection, feature matching is no longer limited to a simple planar coordinate system due to the need to incorporate depth information [16]. Accurately matching these features in 3D space is crucial for autonomous driving systems [17]. In addition, in the practical application of autonomous driving, the algorithms must not only ensure accuracy but also meet real-time requirements. Especially at high speeds or in complex road environments, the system must be able to quickly sense and make decisions about lane information to ensure safety [4].

To address these challenges, we design Freq-3DLane, an efficient and real-time 3D lane detection method. It exploits feature frequency characteristics to achieve accurate and robust perceptual behavior, as shown in Fig.1. The first step is to unify the visual space through a virtual camera. Then, Backbone performs multi-level feature extraction to fully explore the rich information contained in different network depths and feature sizes. The Frequency-aware Feature Pyramid Module (FreqFPM) effectively integrates information across various scales extracted by the backbone network. Incorporating frequency domain features enhances spatial consistency and improves detail processing in lane detection. Specifically, the frequency feature fusion strategy not only enhances the ability to capture local details, but also improves the global understanding of a wide range of scenes. In addition, the Spatial Transformation Fusion Module (STFM) enables network multi-scale features to move beyond purely parallel processing and dynamically adapt across different scales. STFM effectively addresses the issue of information loss caused by perspective variations or complex scenes through precise spatial alignment. By combining with the attention module, more robust and high-level features can be extracted, further enhancing the overall performance of the network. Finally, key points are used as the 3D lane representation method, enabling precise identification of key lane locations in the BEV space and achieving high-accuracy lane modeling. Through in-depth research and extensive experiments, we have thoroughly validated the exceptional performance and effectiveness of the Freq-3DLane method.

The contributions of this paper are as follows:

- We propose the Freq-3DLane end-to-end 3D lane detection framework. By analyzing the frequency characteristics of features, spatial awareness is enhanced to further performance improvement.
- We develop a multi-scale information integration mechanism, which constructs a feature fusion network through a frequency fusion mechanism to ensure that detailed information and global structure are fully utilized. After that, we perform perspectives variation and feature enhancement of the features to obtain better BEV fusion features.
- We conduct thorough experiments on the benchmark datasets of OpenLane, and Apollo 3D Lane Synthetic. Our proposed Freq-3DLane delivers outstanding performance, achieves a 59.7 F1 score on Openlane, and outperforms state-of-the-art methods with a 98.4 F1 score on Apollo 3D Lane Synthesis.

The rest of the article is organized as follows. Chapter II reviews the related work on lane detection and feature fusion. Chapter III presents detailed information references about the proposed Freq-3DLane. Chapter IV provides extensive experimental results to validate the performance of Freq-3DLane. Chapter V summarizes the full text.

II. Related Work

A. 2D Lane Detection

Recently, thanks to advances in deep learning and computer technology, 2D lane detection has made significant progress. Existing approaches explore the problem from four main ways.

Segmentation-based methods [6], [18] [19] mainly use pixel-by-pixel classification to distinguish between lane and non-lane areas, focusing on extracting more effective and semantically informative features. Anchor-based methods [20], [21] use predefined anchors to model lanes and

subsequently regress the offsets between the sampled points and the anchors, enabling 2D lane prediction. Additionally, 2D lanes are modeled by introducing row anchors in the row direction and grid cells in the column direction, which significantly enhances the accuracy and efficiency of lane detection [8], [22] [23]. Keypoint-based methods [24], [25] [26], [27] can model lanes flexibly, specifically by estimating the position of points to get the position and shape of the lane. Parametric curve-based methods [28], [29]. [30] make use of various mathematical models to fit lane curves on a 2D image space using curve-specific formulas. Despite a certain degree of progress in 2D lane detection techniques, there is still a non-negligible gap between the 2D results and the requirements of real-world applications. Especially in complex road environments, 2D detection results provide insufficient spatial information. To bridge this gap, the 3D lane detection technology approach has received attention. It can better understand and model road geometry in 3D space, thus providing more accurate 3D lane location.

B. 3D Lane Detection

Due to the short time frame of the study, fewer studies have explored model design compared to the 2D lane task. 3DLaneNet [9] is the first solution to the problem of 3D lane detection with monocular vision sensors. Specifically, 3DLaneNet uses Inverse Perspective Mapping (IPM) to convert features FV to BEV representation and then regresses the anchor offsets of the lanes in BEV space. Gen-LaneNet [10] enables image segmentation and 3D lane prediction through a two-stage framework. It presents a new geometrically guided lane anchor representation and applies specific geometric transformations to compute 3D lane points. Similarly, CLGo [31] converts the raw image into a BEV image with estimated camera spacing and height and later fits the lanes by predicting polynomial parameters. Persformer [13] uses deformable attention to adaptively and robustly generate BEV features. It uses a transformer-based framework to realize spatial feature transformations to enable accurate detection of 3D lanes. However, the method requires high computational resources. The above methods are valid only under the assumption of flat roads. However, this assumption does not apply to uneven terrain, such as sloping or downhill sections, which may lead to distortion and reduced reliability.

CurveFormer [32] introduces dynamic 3D anchors to model queries as parametric curves. Using transformers sparse query representation and cross-attention mechanism to efficiently regress polynomial coefficients of 3D lanes. SALAD [33] decomposes 3D lane detection into two tasks: 2D lane segmentation and dense depth estimation. The segmented 2D lanes are projected into 3D space, and 3D lane predictions are made by integrating camera parameters with depth information. Anchor3DLane [34] employs pitch and yaw ground modeling for 2D feature extraction using anchor points. These anchor points are then projected into FV elements, from which their features are extracted. These features, rich in structural and contextual information, enable accurate predictions.

However, these learned features may not accurately represent 3D characteristics, as they often overlook the correlations between different perspectives and scales. Therefore, BEV-LaneDet [15] introduces a virtual coordinate and learns the mapping between image features and BEV features using the Feature Pyramid Structure and View Relationship Module. LATR [35] introduces a Lane-aware Query Generator that uses dynamically extracted lane-aware features to initialize the query embedding. In addition, the proposed Dynamic 3D Ground Positional Embedding method builds a bridge between 3D space and 2D images. DV-3DLane [36] designs a bidirectional feature fusion strategy that comprehensively learns the features of each view in PV and BEV space. A unified query generator is designed to extract lane-aware queries from dual views, and a 3D dual-view deformation attention mechanism is introduced to efficiently aggregate dual-view features. PVALane [37] proposes the priori anchors with strong lane localization priori. The priori anchors are projected into FV and BEV spaces via the Prior-Guided View-agnostic Feature Alignment Module, aligning and fusing geometric and semantic information from different views. To better combine global semantic information and local texture information, our method performs frequency feature perception and fusion at multiple scales. It enhances spatial coherence and detailing capabilities, not only for the capture of detailed lane features but also for the global understanding of large-scale scenes.

C. Feature Fusion

In many studies, a fusion of features from different scales has been considered a key means to improve performance [38], [39]. Low-level features have higher resolution and contain more positional information and details. However, they have weaker and noisier semantic information due to less convolution undergone. High-level features have stronger semantic information but lower resolution, resulting in a weaker ability to perceive details. Feature fusion is classified as early fusion and late fusion according to the order of fusion and prediction.

Early fusion involves combining different features and inputting them into a training model. Specifically, features from multiple layers are fused first, and then the predictor is trained on the fused features. Detection is performed only after complete fusion. This type of approach is also known as skip connection, i.e., the use of concat and add operations. Representatives of this approach are Inside-Outside Net [40] and HyperNet [41].

Features are processed independently at different levels to generate individual predictions, which are later fused. This is known as late fusion. Examples include Single Shot MultiBox Detector (SSD) [42], Feature Pyramid Network (FPN) [43], and so on. In which FPNs can make predictions on multiple fused features at different scales to maximize detection accuracy. FPN works by resizing the original image to different scales, using the reduced image for global features and the enlarged image for detailed features. However, after feature fusion, there is often a problem of inconsistency within categories due to interference from high-frequency features. At the same time, the fuzzy boundaries lack precise high-frequency information, thus affecting the accuracy of

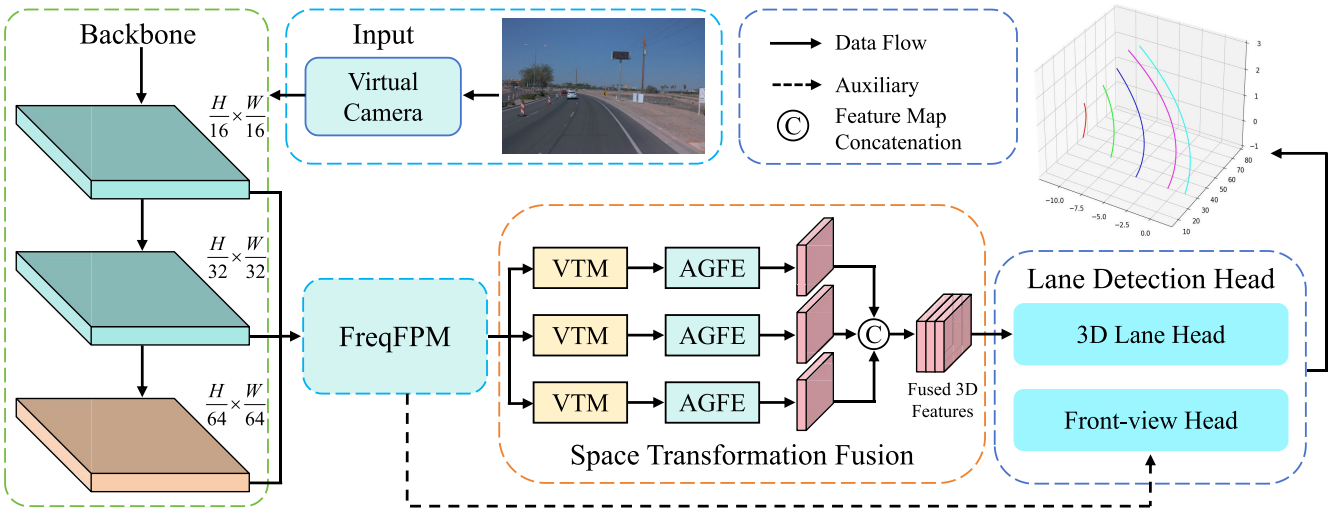

Fig. 2. The overall architecture of Freq-3DLane. Freq-3DLane takes a 2D image after processing as input and extracts multi-scale FV features through the CNN backbone. Freq-3DLane takes a 2D image after processing as input and extracts multi-scale FV features through the CNN backbone. By fusing multi-scale information through our FreqFPM, high-level semantic features can be efficiently fed back into the shallower layers of the network. The front view features were converted to BEV features using STFM, which were further enhanced before fusing the multi-scale information. Finally, lane positions are predicted using 3D lane heads. Additionally, 2D FV features are used as auxiliary information.

the results. To address this issue, FreqFusion [44] introduces a frequency-aware feature fusion method that leverages both low-level and high-level features, effectively enhancing feature consistency and sharpening object boundaries.

III. Methodology

3D lane detection is aimed to predict the 3D position of lanes in an input image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$. Lanes are represented by a collection of 3D points denoted as $P = \{L_i \mid i \in 1, 2, \ldots, N\}$, where N is the number of lanes in the image, and L_i denotes the i-th lane. Each lane consists of a collection of multiple points, denoted as $L_i = \{(x_i^j, y_i^j, z_i^j)\}_{j=1}^{M}$, where M is the total number of lane point collections [9], [13].

A. Overall Architecture

The overall architecture of Freq-3DLane is illustrated in Fig. 2. Firstly, the internal/external parameters of the image are entered uniformly via a virtual camera. Immediately after that, we use the 2D backbone to extract the multi-scale feature mapping $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ from the input image. We then process the multi-scale features using the Frequency-aware Feature Pyramid Module. It fuses low-level and high-level feature information to enhance the perception of lanes at different scales. Next, the STFM comes into play, which consists of the View Transformation Module (VTM) and the Attention-Guided Feature Enhancer (AGFE). The conversion from FV features to BEV features is performed, and the feature learning is then guided to enhance the semantic information of lane markings, ensuring that the lane geometry is accurately recovered from different perspectives. Subsequently, these features are fused. Finally, we apply the prediction header to the enhanced BEV features to obtain the final lane prediction. In addition, to enhance the network's ability to extract FV features, we introduce a FV lane detection header as an auxiliary supervision to improve the model performance.

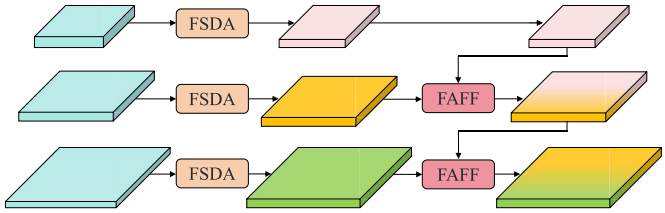

Fig. 3. The structure of FreqFPM, a three-layer feature pyramid, consists of two key components: feature sensitive dimensional aligner and frequency aware feature fusion.

B. Backbone

We adopt ResNet [45] as the backbone network for the Freq-3DLane feature extraction process. ResNet is a landmark model in the history of deep learning. It solves model degradation problems in deep networks by using shortcut connections. Using ResNet, we can extract multi-scale features and generate feature maps with $\frac{1}{16}$ and $\frac{1}{32}$ resolution, thus capturing rich spatial information at different scales and enhancing the perceptual capability of the model. To further extract deeper semantic information, we add a convolutional module after ResNet to improve the expressiveness of the backbone network. This module can effectively extract the feature maps with $\frac{1}{64}$ resolution, thus further improving the detection accuracy of the model.

C. Frequency-Aware Feature Pyramid Module

Feature Pyramid fuses high-level features (low-resolution, rich semantic information) with low-level features (high-resolution, detail edge information) through top-down, side-by-side connectivity, ensuring that semantic information is fully captured across all scales. To tackle the challenge of fine-grained extraction of small-scale lanes in lane detection and meet the need for capturing large-scale road structures, we propose FreqFPM, a frequency-aware feature pyramid module. The structure of FreqFPM is shown in Fig. 3.

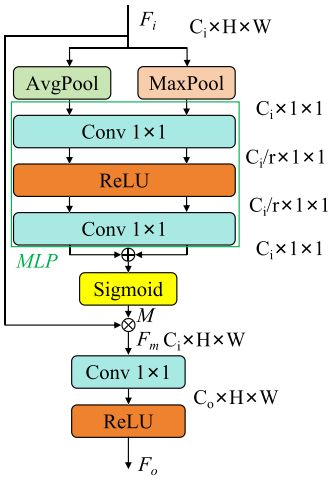

Fig. 4. A detailed framework for feature sensitive dimensional Aligner.

It consists of two key components: Feature Sensitive Dim Aligner (FSDA) and Frequency Aware Feature Fusion (FAFF). We collectively refer to the clear perception of local structure and global layout in an image as spatial awareness. Image information is represented as both original image features and frequency features. In this case, frequency features are image features divided into components of different frequencies. On the one hand, high-frequency features represent details and local structures (e.g., driveway edges). On the other hand, low-frequency features represent global structure and large-scale information (e.g., the general layout of the lane).

The FSDA enhances the representation of features in both horizontal and vertical directions, which later achieves accurate dimensional alignment of different spatial scale features. As shown in Fig. 4. The module receives an input feature map $\boldsymbol{F_i} \in \mathbb{R}^{C\times H\times W}$. The global average pooling and global maximum pooling performed on F_i results in two $C \times 1 \times 1$ feature maps. The pooling results are then fed into the multilayer perceptron (MLP). The MLP results are subjected to the Add operation, which is followed by Sigmoid activation to get the weight matrix M. As shown in Eq. (1):

$$M = \sigma(MLP(Ap(F_i)) + MLP(Mp(F_i))), \tag{1}$$

where σ stands for Sigmoid, Ap for Avg Pooling and Mp for Maximum Pooling. After multiplying M with F_i, we get $\boldsymbol{F_m} \in \mathbb{R}^{C\times H\times W}$. At last, F_m undergoes successive convolution and activation to obtain the final output F_o. FSDA can be formalised as follows Eq. (2):

$$F_o = \text{ReLU}(\text{Conv}(F_i \times M)) = \text{ReLU}(\text{Conv}(F_m)). \tag{2}$$

FAFF is the core component of FreqFPM that fuses features from different scales. As shown in Fig. 5. FreqFPM consists of three key components: the filter feature processor (FIPR), the low-pass filter (LPF), and the high-pass filter (HPF). The method extracts high-frequency information from high-resolution features, fused with low-resolution features to enhance the expressive power of low-resolution features. At the same time, high-resolution features, guided by low-resolution, can highlight key features at high resolution and further. Eventually, the enhanced two are fused to effectively combine the semantic information of the low-resolution features with the detailed information of the high-resolution features. In this way, lane features can be extracted more accurately.

The core function of the FIPR is to normalize the kernel on the input feature map. Its primary objective is to improve the stability and efficiency of convolution operations by learning the convolution kernel weights for each local region and applying the appropriate normalization. Through the use of Softmax, weighting, and normalization, a normalized weight distribution mask is generated for subsequent operations. Weighting can be performed using convolution kernels of varying sizes, such as 3×3 or 5×5, enabling the preprocessing of weights for high-pass and low-pass filtering, i.e., HFIPR and LFIPR.

The FreqFPM enhances global understanding by introducing low-frequency information. The low-frequency information extracted by the LPF helps to understand the overall shape of the lane, thus identifying the lanes on a large scene scale. The LPF allows the low-frequency components of the feature map to pass through while attenuating the high-frequency components. The mask helps the filter identify and reduce high-frequency elements that may be inconsistent during the feature fusion process. The initial LPF feature is obtained via 3×3 convolution in the previous step, and this is combined with the mask generated by the LPF feature processor for further processing. To ensure the effectiveness of the convolution operation, the input is first padded. The unfold operation is then applied to convert the feature map into a sliding window. The expanded feature map is upsampled using interpolation to enlarge its size as necessary. Following this, the feature maps and masks are adjusted to the appropriate shapes, and element-wise multiplication is performed to generate the weighted feature maps. The results from all convolution kernels are then summed to produce the processed feature map.

HPF extracts high-frequency information from the image, enabling the model to accurately recognize fine structures. Incorporating high-frequency details into the feature maps at various scales preserves clarity and ensures precise lane marking identification in diverse environments, such as occlusions and blurring. The essence of HPF lies in first applying a LPF to extract the low-frequency components, and then subtracting these components from the original signal to isolate the high-frequency components. It is worth noting that the Interpolation and Upsample of the LPF do not work when the HPF is applied. Subsequently, the original signal is combined with the high-frequency-enhanced signal to obtain the high-pass filtered result. As shown in Eq. (3):

$$F_h = F + (F - F_l) \tag{3}$$

By fusing low and high-frequency information, the model can find a balance between detail recognition and global structure perception, further enhancing spatial awareness. This means that the model is not only able to accurately recognize lane lines, but also accurately locate the position of the lane in a wider area.

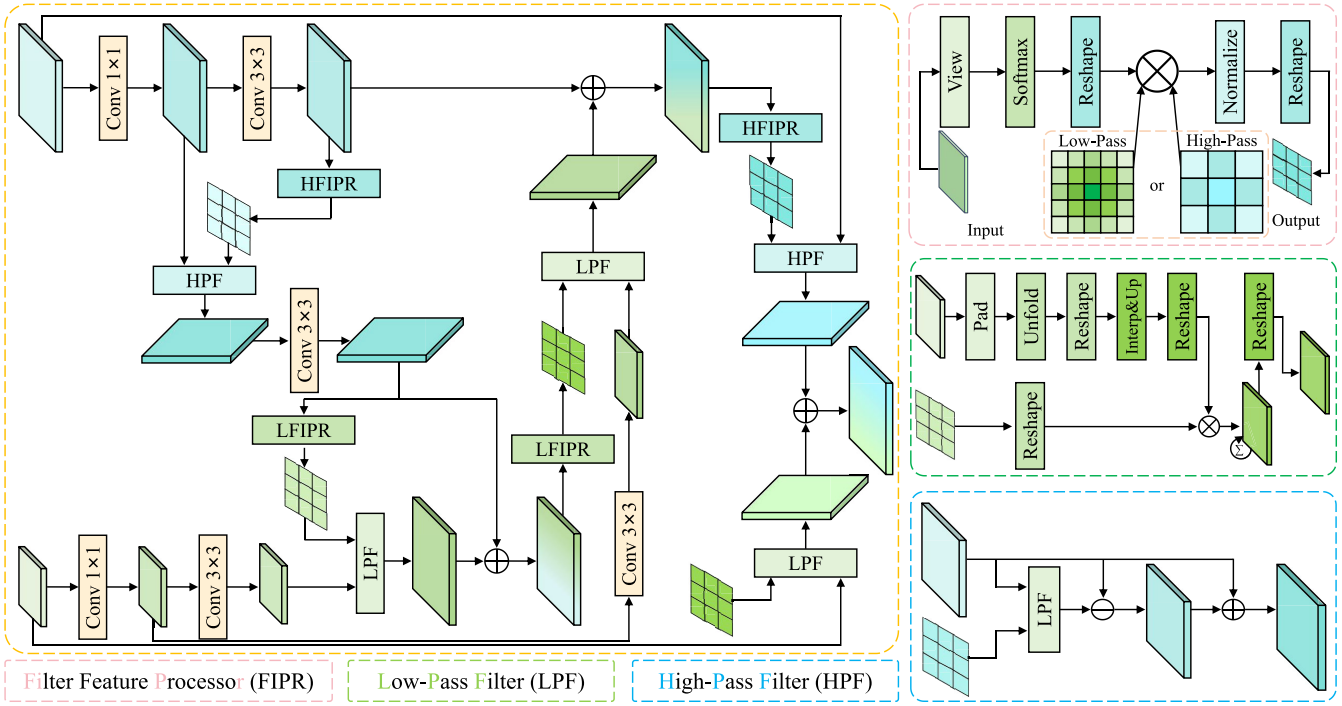

Fig. 5. The illustration of FAFF. The module is meticulously designed with a constellation of components, including dedicated filter feature processors (LFIPR and HFIPR), selective filters (LPF and HPF), and convolutional layers. These elements are interconnected through a sophisticated network of pathways, forming a complex and dynamic architecture. In this setup, the high-resolution, low-semantic high-frequency elements play a pivotal role in guiding the enhancement of semantic content within the corresponding low-resolution features. Subsequently, a feedback loop is initiated, reintegrating the enriched high-semantic data into the system, which further refines and augments the low-semantic information. This cyclical process culminates in a more robust and nuanced overall feature representation.

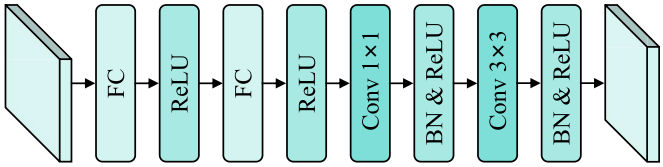

Fig. 6. The structure of the VTM module consists of two fully connected layers and two convolutional layers. This integration enables efficient feature transformation and purposeful feature processing.

D. Space Transformation Fusion Module

In STFM, FV features with varying resolutions are transformed into BEV features via the VTM. In this process, the VTM learns all the spatial location dependencies between the FV and BEV features to establish different viewpoint associations. Then, the AGFE extracts more robust and high-level BEV features. After several VTM and AGFE processing, the features at different scales are transformed into advanced BEV features with rich information. Finally, we fuse the processed BEV features, which efficiently capture the correlations between different perspectives and scales.

The VTM structure is shown in Fig. 6. It can capture the association between the FV feature and any two-pixel position in the BEV feature. The input features are processed by two fully connected layers, whose main function is to map features of different sizes into a fixed-size feature space. This operation enables the network to handle inputs of various resolutions while providing a uniform-size feature map for subsequent convolution operations. Subsequently, features are further extracted and processed by 1 × 1 and 3 × 3 convolutional layers.

AGFE fuses features at different scales using attention mechanisms to extract richer information and enhance feature representation. The input is split into two parts, each processed through separate branches. One part undergoes Group Normalization before being divided into query (Q), key(K), and value(V) components. Next, the attention score is computed using Q with K, and the attention weights are applied to V. As shown in Eq. (4):

$$F_1{}' = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{4}$$

where dk denotes the dimension of the middle layer used to compute the attention weights. The core is to calculate the inner product of the Q, K, and V vectors, and use the results as weights for different feature vectors, thereby enabling the 'focus' function of the attention mechanism. The weighted features are then fused with another part of the input. Finally, the fused features are further refined through a 1 × 1 convolutional layer to ensure optimal feature representation in the output. As shown in Fig.7.

E. Lane Representation

We use the keypoint method proposed by BEV-LaneDet [15] to represent lanes, which is simple and robust. It is a 3D lane detection head consisting of four predictive heads. The predictions include confidence, embedding for clustering,

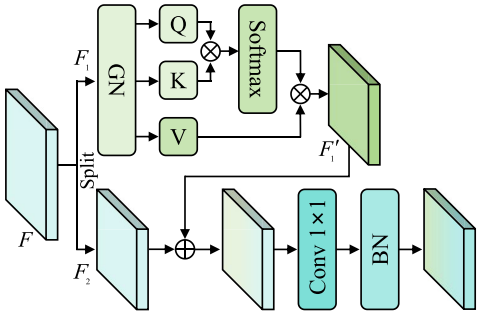

Fig. 7. Structure of AGFE. The input features are split and then fused through the attention mechanism to enhance the feature representation.

offset from the center of the cell to the lane in the y-direction, and the average height of each cell. Additionally, to further optimize the training process, a 2D prediction branch is added to calculate the lane embedding loss in the FV, thus improving the overall detection accuracy.

F. Loss Function

The loss of confidence is as in Eq. (5):

$$\begin{aligned}\mathcal{L}_c &= \sum_{i}^{H'}\sum_{j}^{W'}(BCE(p_{ij},\widehat{p}_{ij}) + IOU(p_{ij},\widehat{p}_{ij})) \\ &= \sum_{i}^{H'}\sum_{j}^{W'}(\widehat{p}_{ij}log(p_{ij}) + (1-\widehat{p}_{ij})log(1-p_{ij}) \\ &\quad + \frac{p_{ij}\widehat{p}_{ij}\alpha}{p_{ij}\alpha + \widehat{p}_{ij}\alpha + p_{ij}\widehat{p}_{ij}\alpha}),\end{aligned} \tag{5}$$

where BCE represents the binary cross-entropy loss, and IOU refers to the intersection over union loss. H' and W' denote the number of grid cells in the height and width directions of the predicted BEV plane, respectively. p denotes the confidence probability of the model's prediction, while $\widehat{p}$ represents the true confidence value. α denotes the IoU mask.

In addition, the prediction offset loss in the x-direction of the prediction plane is shown in Eq. (6):

$$\mathcal{L}_o = \sum_{i}^{H'}\sum_{j}^{W'} BCE(x_{ij}, \sigma(\widehat{x}_{ij})), \tag{6}$$

where σ denotes the sigmoid function.

BEV-LaneDet [15] enables the differentiation of channel identifiers for each pixel in the confidence branch by predicting the embedding vectors for each grid cell. We used the same embedding loss. As shown in Eq. (7):

$$\mathcal{L}_e = \lambda_{pull}\mathcal{L}_{pull} + \lambda_{push}\mathcal{L}_{push}, \tag{7}$$

where *pull* represents the mean-minimize loss of unit embeddings within the same channel, aimed at reducing feature variations among samples of the same instance. *push* refers to the variance-maximize loss of cell embeddings across different channels, designed to enhance feature differences between distinct instances.

For lane height prediction, the average height in the grid cell is used as the ground truth. We use the L1 loss to calculate the lane height loss. As shown in Eq. (8):

$$\mathcal{L}_h = \sum_{i}^{H'}\sum_{j}^{W'}\left|h_{ij} - \widehat{h}_{ij}\right|, \tag{8}$$

where h represents the predicted height and $\widehat{h}$ represents the actual ground height.

Finally, to ensure that the 2D features effectively capture lane features, we use a 2D lane detection header as an aid and introduce an auxiliary loss for 2D lane detection as shown in Eq. (9):

$$\mathcal{L}_{2d} = \mathcal{L}_c^{2d} + \mathcal{L}_c^{2d} + \mathcal{L}_e^{2d} + \mathcal{L}_e^{2d}, \tag{9}$$

where $\mathcal{L}_c$ denotes confidence loss and $\mathcal{L}_e$ denotes embedding loss.

The total loss is defined as shown in Eq. (10):

$$\mathcal{L}_{total} = \lambda_c\mathcal{L}_c + \lambda_o\mathcal{L}_o + \lambda_e\mathcal{L}_e + \lambda_h\mathcal{L}_h + \lambda_{2d}\mathcal{L}_{2d}, \tag{10}$$

where λ denotes the weight applied to each loss component.

IV. Experiments

We conduct experiments on two popular 3D lane detection benchmarks, including Apollo 3D Synthetic dataset [10] and OpenLane dataset [13].

A. Datasets and Evaluation Metrics

Apollo 3D Synthetic [10] is a synthetic dataset created using the Unity 3D engine with a variety of 10.5K virtual scene images, including highways, cities, homes, and city centers. In addition, the picture data is diverse in terms of daylight, weather conditions, traffic obstacles, and road quality. It is divided into three subsets: 1) *Balanced scenes*, 2) *Rarely observed scenes*, and 3) *Scenes with visual variations*.

OpenLane [13] is valuable content from the Waymo Open Dataset, a real-world 3D lane detection dataset containing 1000 scenarios. It includes 200,000 image frames covering more than 880,000 instance-level lanes, encompassing different weather conditions and areas (residential, urban, suburban, highway, and parking lot). In order to evaluate various driving scenarios, the test set was divided into several categories, including *Up & Down*, *Curve*, *Extreme Weather*, *Night*, *Intersection* and *Merge & Split*. For the ablation study, we used the smaller version of the OpenLane300 dataset, which contains 300 scenes.

Evaluation Metrics. For the two 3D datasets, we employ evaluation metrics from Gen-LaneNet [10]. It calculates the Euclidean distances at uniformly distributed points in the y-direction from 0 to 100 meters. If the point-by-point distance is less than a predefined threshold of 1.5 meters for at least 75% of the points, the lane prediction is considered to match. Then, according to the average distance and range, the F1 score was calculated, and the average $X_{error} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \widehat{x}_i)^2}$ and $Z_{error} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(z_i - \widehat{z}_i)^2}$ for near (0-40 m) and far (40-100 m) distances to evaluate the geometric accuracy. Specifically, this is achieved by $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$, where $Precision = \frac{N_{TP}}{N_{TP}+N_{FP}}$ and $Reacll = \frac{N_{TP}}{N_{TP}+N_{FN}}$.

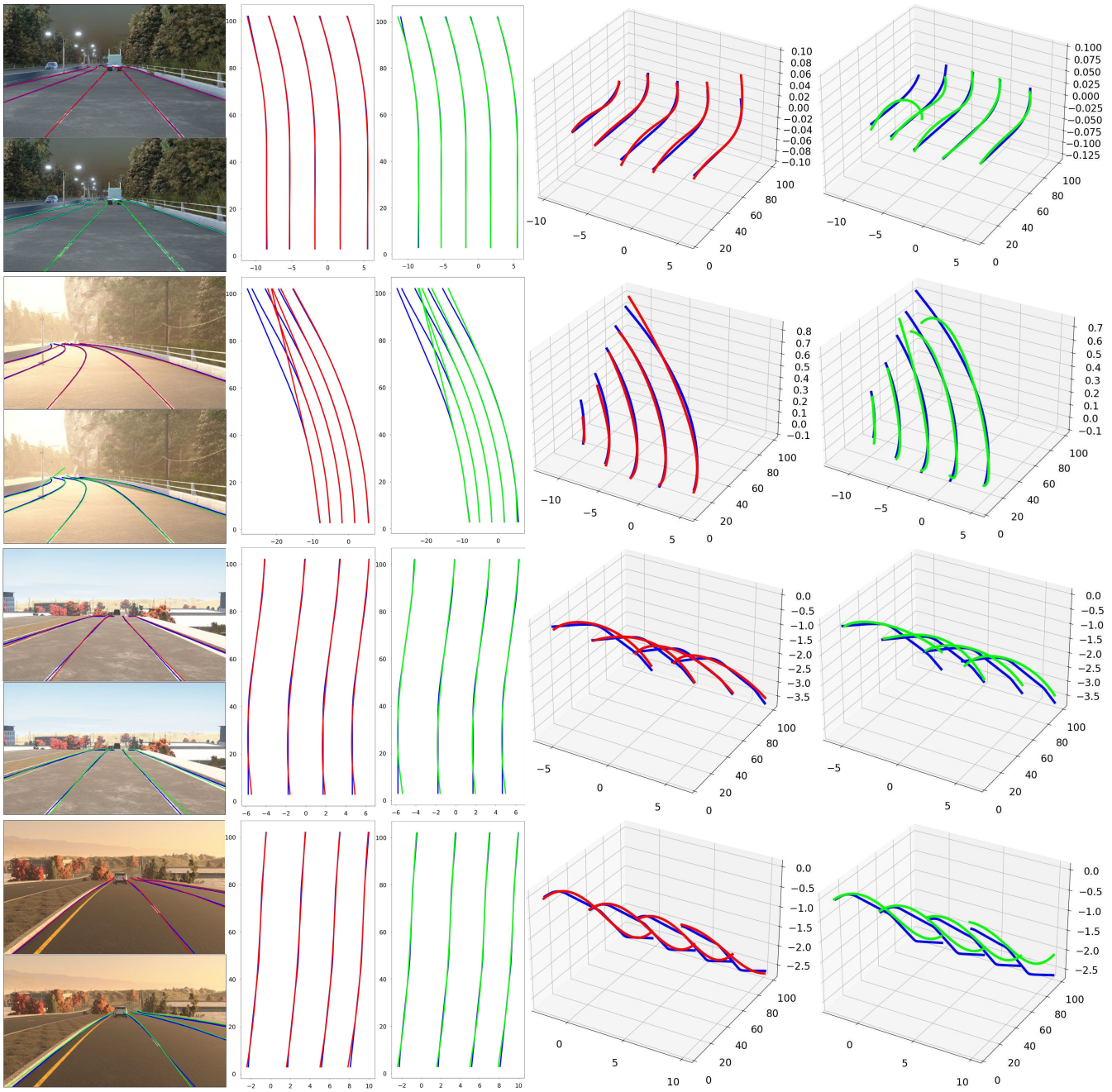

Fig. 8. Qualitative comparison on Apollo 3D Synthetic [10]. We choose BEV-LaneDet [15] as our comparison method. We visualize the results in different scenarios. Each visualization includes the FV image (left), the BEV output (middle), and the 3D lanes (right). The blue lanes are the ground truth, the red lanes are the predicted results of our method, and the green lanes represent the BEV-LaneDet output.

B. Implementation Details

We use input size 576 × 1024 and adopt ResNet-34 [45] as our backbone to extract feature maps from two scales with spatial reduction ratios of $[\frac{1}{16}, \frac{1}{32}]$. To obtain deep image features, we add additional CNN layers to generate $\frac{1}{64}$ feature map. Then, we constructed a three-level feature pyramid using the generated multi-scale features. We apply the AdamW optimizer with a weight decay of 0.01. The initial learning rate was set to 2×10^{-4} and the cosine annealing scheduler was used. We train the models for 30 epochs on OpenLane and 100 epochs on Apollo dataset with batch size 16.

C. Main Results

1) Results on ApolloSim: Table I summarises the results of our experiments on the Apollo dataset. Based on the literature [10], we evaluate our method in three different scenes and study the F1 score and X/Z errors. Despite near-saturation performance, our Freq-3DLane shows superiority in all scenarios and metrics. Specifically balanced scene +1.5% F1, rarely observed +1.4% F1, and visual variations +3.3% F1, which suggests the effectiveness of our design. Notably, our Freq-3DLane also achieves comparable or lower X/Z errors compared to previous methods. In addition, we provide qualitative comparison results, as shown in Fig. 8.

TABLE I

THERE ARE RESULTS OF DIFFERENT MODELS ON APOLLO 3D LANE SYNTHETIC [10]. RED REPRESENTS OPTIMAL, BLUE REPRESENTS SUB-OPTIMAL

Scene	Methods	F1(%)↑	X error (m)↓		Z error (m)↓	
			near	*far*	*near*	*far*
Balanced Scene	3DLaneNet [9]	86.4	0.068	0.477	0.015	0.202
	Gen-LaneNet [10]	88.1	0.061	0.496	0.012	0.214
	PersFormer [13]	92.9	0.054	0.356	0.010	0.234
	CLGO [31]	91.9	0.061	0.361	0.029	0.250
	Reconstruct from Top [46]	91.9	0.049	0.387	0.008	0.213
	CurveFormer [32]	95.8	0.078	0.326	0.018	0.219
	Anchor3DLane [34]	95.6	0.052	0.306	0.015	0.223
	BEV-LaneDet [15]	96.9	0.016	0.242	0.020	0.216
	Freq-3DLane(ours)	98.4	0.021	0.200	0.022	0.208
Rarely Observed	3DLaneNet [9]	72.0	0.166	0.855	0.039	0.521
	Gen-LaneNet [10]	78.0	0.139	0.903	0.030	0.539
	PersFormer [13]	87.5	0.107	0.782	0.024	0.602
	CLGO [31]	86.1	0.147	0.735	0.071	0.609
	Reconstruct from Top [46]	83.7	0.126	0.903	0.023	0.625
	CurveFormer [32]	95.6	0.182	0.737	0.039	0.561
	Anchor3DLane [34]	94.4	0.094	0.693	0.027	0.579
	BEV-LaneDet [15]	97.6	0.031	0.594	0.040	0.556
	Freq-3DLane(ours)	99.0	0.042	0.526	0.048	0.542
Visual Variations	3DLaneNet [9]	72.5	0.115	0.601	0.032	0.230
	Gen-LaneNet [10]	85.3	0.074	0.538	0.015	0.232
	PersFormer [13]	89.6	0.074	0.430	0.015	0.266
	CLGO [31]	87.3	0.074	0.464	0.045	0.312
	Reconstruct from Top [46]	89.9	0.060	0.460	0.011	0.235
	CurveFormer [32]	90.8	0.125	0.410	0.028	0.254
	Anchor3DLane [34]	93.6	0.068	0.367	0.020	0.232
	BEV-LaneDet [15]	95.0	0.027	0.320	0.031	0.256
	Freq-3DLane(ours)	98.3	0.021	0.191	0.022	0.198

TABLE II

COMPARISON WITH VARIOUS METHODS ON OPENLANE VALIDATION SET [13]. RED REPRESENTS OPTIMAL, BLUE REPRESENTS SUB-OPTIMAL

Methods	F1(%)↑	X error (m)↓		Z error (m)↓		FPS
		near	*far*	*near*	*far*	
3DLaneNet [9]	44.1	0.479	0.572	0.367	0.443	50
Gen-LaneNet [10]	32.3	0.591	0.684	0.411	0.521	54
PersFormer [13]	50.5	0.485	0.553	0.364	0.431	21
CurveFormer [32]	50.5	0.340	0.772	0.207	0.651	-
CurveFormer++ [47]	52.7	0.337	0.801	0.198	0.676	-
Anchor3DLane [34]	53.1	0.300	0.311	0.103	0.139	72
BEV-LaneDet [15]	58.4	0.309	0.659	0.244	0.631	102
Freq-3DLane(ours)	59.7	0.265	0.665	0.196	0.608	80

2) Results on OpenLane: We present the experimental results on OpenLane dataset in Table II. Our model significantly outperforms other methods in terms of F1 score. Compared to BEV-LaneDet [15], Freq-3DLane improves the F1 score by 1.3% and achieves lower geometric error. There is a gap in geometric error compared to Anchor 3DLane [34], but our method far exceeds the F1 score for that method. It can be seen that in real scenarios, more accurate predictions can be produced after the frequency features are fused. In addition to the quantitative results, we show the qualitative visualization results in Fig. 9.

Notably, our Freq-3Dlane method maintains a high Frame Per Second (FPS) while achieving good F1 score. This is an unexpected but pleasing result. This advantage makes Freq-3Dlane particularly well-suited for real-time scenarios. The method thus offers a strong balance of performance and efficiency, demonstrating its potential for deployment in demanding real-time applications.

Along with the results of our experiments on OpenLane, we also conducted a comprehensive evaluation across several scenarios in the OpenLane test set. As shown in Table III, previous methods were compared across six

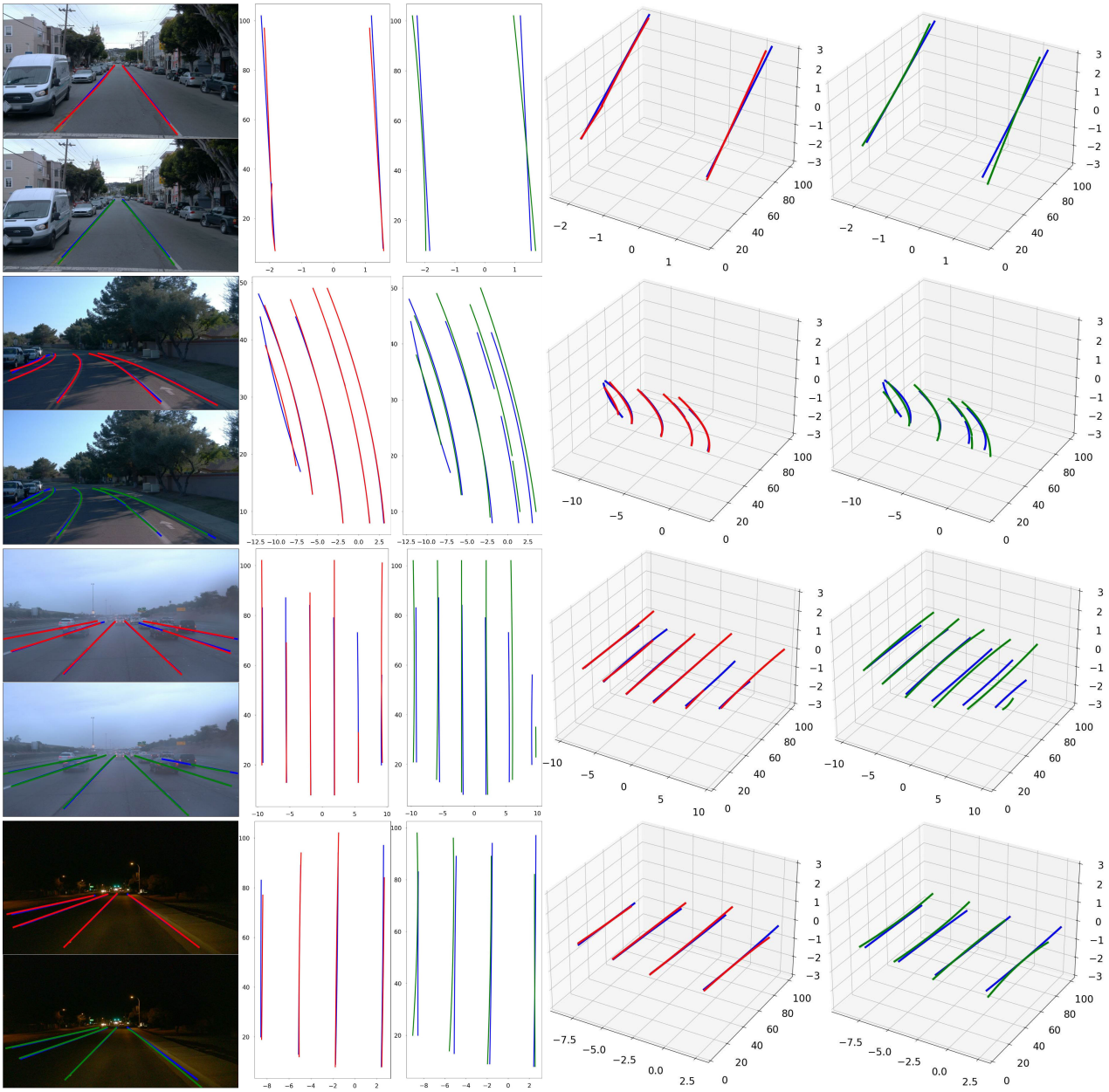

Fig. 9. Qualitative comparison results on Openlane dataset [13]. Each result includes the FV image (left), the BEV output (middle), and the 3D lanes (right). The blue lanes are the ground truth, the red lanes are the predicted results of our method, and the green lanes represent the BEV-LaneDet output.

TABLE III

COMPARISON WITH OTHER 3D LANE DETECTION METHODS ON OPENLANE DATASET UNDER DIFFERENT SCENARIOS [13]. RED REPRESENTS OPTIMAL, BLUE REPRESENTS SUB-OPTIMAL

Methods	All	Up&Down	Curve	Extreme Weather	Night	Intersection	Merge&Split
3DLaneNet [9]	44.1	40.8	46.5	47.5	41.5	32.1	41.7
Gen-LaneNet [10]	32.3	25.4	33.5	28.1	18.7	21.4	31.0
PersFormer [13]	50.5	42.4	55.6	48.6	46.6	40.0	50.7
CurveFormer [32]	50.5	45.2	56.6	49.7	49.1	42.9	45.4
CurveFormer++ [47]	52.7	48.3	59.4	50.6	48.4	45.0	48.1
Anchor3DLane [34]	53.1	45.5	56.2	51.9	47.2	44.2	50.5
BEV-LaneDet [15]	58.4	48.7	63.1	53.4	53.4	50.3	53.7
Freq-3DLane(ours)	59.7	52.0	65.5	56.5	54.5	51.8	56.4

challenging scenarios, with F1 score reported for each. Analyzing the detection scores across these scenarios reveals a significant performance improvement, highlighting the superior ability of our method to accurately capture 3D space.

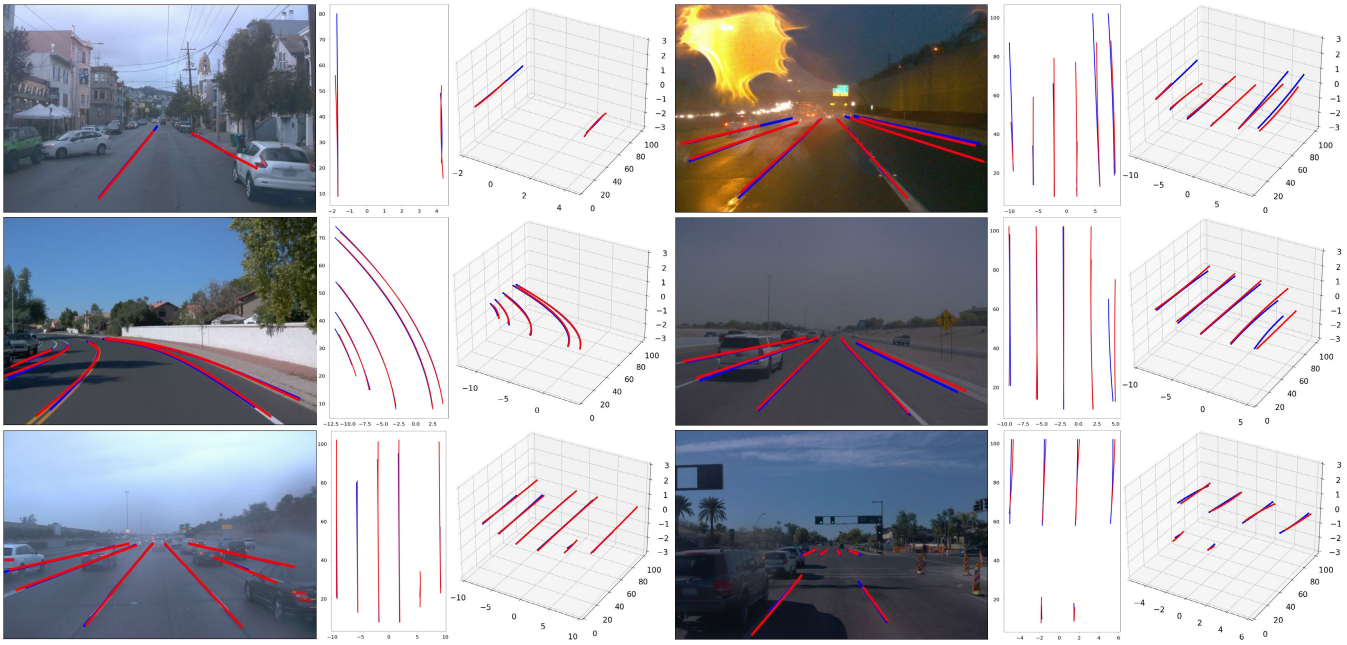

Fig. 10. Visualization results for each scenario of the Openlane dataset [13]. Each visualization includes the FV image (left), the BEV output (middle), and the 3D lanes (right). The blue lanes are the ground truth, and the red lanes are the predicted results of our method.

TABLE IV

THE EFFECT OF DIFFERENT FEATURE COMBINATIONS ON OPENLANE300 AFTER ADDING FREQUENCY INFORMATION. F_n REPRESENTS $n \times$ DOWNSAMPLING OF THE INPUT IMAGE

F_{16}	F_{32}	F_{64}	F1(%)↑	Flops(G)	Params(M)
✓			56.4	44.5	10.6
	✓		58.9	52.0	23.4
		✓	58.4	54.8	43.2
✓	✓		60.6	52.5	24.8
	✓	✓	61.7	55.2	43.3
✓		✓	60.1	55.3	43.7
✓	✓	✓	62.9	55.8	44.8

TABLE V

THE EFFECT OF DIFFERENT FEATURE COMBINATIONS ON OPENLANE300. F_n REPRESENTS $n \times$ DOWNSAMPLING OF THE INPUT IMAGE. $A\&F$ STANDS FOR USE OF FEATURE SENSITIVE DIM ALIGNER AND FREQUENCY AWARE FEATURE FUSION

FPM scale	A & F	F1(%)↑	Flops(G)	Params(M)
F_{16} + F_{32}	✓	61.9	53.4	25.6
F_{32} + F_{64}	✓	63.8	55.5	44.9
F_{16} + F_{64}	✓	61.3	56.2	45.6
F_{16} + F_{32} + F_{64}	✓	64.1	56.6	46.7

TABLE VI

PERFORMANCE GAIN FOR DIFFERENT CONTRIBUTIONS ON OPENLANE300 USING OUR NOVEL FEATURE SENSITIVE DIM ALIGNER AND FREQUENCY AWARE FEATURE FUSION. NORMAL MEANS NORMAL 1*1 CONVOLUTION, USED TO ADJUST THE NUMBER OF CHANNELS. SIMPLE REPRESENTS A SIMPLE SUM OF FEATURES

Align	Fusion	F1(%)↑	X error(m)↓	Z error(m)↓
Normal	Simple	62.9	0.328/0.603	0.230/0.491
FSDA	Simple	63.2	0.324/0.595	0.232/0.489
Normal	FAFF	63.7	0.321/0.592	0.227/0.486
FSDA	FAFF	64.1	0.315/0.589	0.226/0.483

The results of the Openlane dataset visualization for various scenarios are shown in Fig. 10.

D. Ablation Study

1) Features at Different Scales: We explore the impact of feature extraction on the network at different scales. The effectiveness and contribution of the features at each connection were evaluated by training and testing on various scales. The experimental results are shown in Table IV. Multi-scale feature extraction shows a significant performance improvement. By fusing information from different scales, the model can better capture lane features. The experimental results show that the fusion of $[\frac{1}{16}, \frac{1}{32}, \frac{1}{64}]$ original images gives the best results.

2) Frequency-aware Feature Pyramid Design: We splice FSDA with FAFF on different combinations of feature pyramids, thus introducing frequency information to construct FreqFPM. The experimental results are shown in Table V. The FreqFPM constructed after the combination can effectively improve the detection effect.

Moreover, we study the respective contributions of the submodules of the FreqFPM. The experimental results are shown in Table VI. We have found that each submodule of FreqFPM has a non-negligible contribution to the final performance. In particular, FSDA helps to improve detection precision at multiple scales, while FAFF significantly reduces the x/z error. These two modules complement each other and combine the model's advantages in multi-scale and frequency feature processing, further validating the effectiveness of FreqFPM.

3) Mechanisms for Space Transformation Fusion Module: The details of the STFM experiment are provided in Table VII.

TABLE VII
DIFFERENT DESIGN CHOICES FOR SPACE TRANSFORMATION FUSION MODULE. - REPRESENTS THE NON-USE OF VTM AND AGFE

Model	F1(%)↑	X error(m)↓	Z error(m)↓
-	64.1	0.315/0.589	0.226/0.483
VTM only	64.4	0.313/0.583	0.226/0.478
VTM + AGFE	65.6	0.306/0.576	0.221/0.472

TABLE VIII
ABLATION STUDY ON OPENLANE300 VALIDATION SET. FPM DENOTES THE FEATURE PYRAMID MODULE OF THE THREE-LAYER SET. FREQ DENOTES FREQUENCY-AWARE FEATURE PYRAMID MODULE. STFM DENOTES SPACE TRANSFORMATION FUSION MODULE

	Model	F1(%)↑	X error(m)↓	Z error(m)↓
(I)	Baseline	58.9	0.344/0.622	0.254/0.528
(II)	+ FPM	62.9	0.328/0.603	0.230/0.491
(III)	+ Freq	64.1	0.315/0.589	0.226/0.483
(IV)	+ STFM	65.6	0.306/0.576	0.221/0.472

The introduction of VTM enhanced the model's F1 score by 0.6%, demonstrating its effectiveness in mitigating information loss caused by dynamic objects and changes in viewpoint. Furthermore, incorporating AGFE led to an additional 1.2% improvement in accuracy, highlighting its ability to help the model focus more effectively on critical regions in the BEV, thereby boosting feature robustness and adaptability. In conclusion, the STF mechanism significantly strengthens multi-scale fusion and enhances the model's overall expressive power.

4) Different Components: Table VIII summarizes the impact of our different components. The first row shows our baseline. Our multi-scale FPM improves performance by 3% compared to a single scale. Further, the addition of Freq enhances the performance of our model to 64.1%, combining frequency information at different scales and enhancing spatial awareness. Moreover, STFM creates correlations between different perspectives and scales to enhance semantic and structural information, which significantly improves performance.

V. CONCLUSION

In this paper, we present Freq-3DLane, a simple yet effective end-to-end 3D lane detection framework. Freq-3DLane performs 3D lane detection by fusing multi-scale information and leveraging the frequency characteristics of features to enhance perception. Additionally, we introduce an attention-guided spatial transform fusion module to further improve detection performance. Extensive experiments demonstrate that Freq-3DLane achieves impressive results. We believe that our work has the potential to make a positive contribution to society and lay the groundwork for future research advancements.

REFERENCES

[1] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. De Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2020.

[2] E. Marti, M. A. de Miguel, F. Garcia, and J. Perez, "A review of sensor technologies for perception in automated driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 4, pp. 94–108, Winter 2019.

[3] A. Chougule, V. Chamola, A. Sam, F. R. Yu, and B. Sikdar, "A comprehensive review on limitations of autonomous driving and its impact on accidents and collisions," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 142–161, 2024.

[4] J. Bi et al., "Lane detection for autonomous driving: Comprehensive reviews, current challenges, and future predictions," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 23, 2025, doi: 10.1109/TITS.2024.3524603.

[5] N. J. Zakaria, M. I. Shapiai, R. A. Ghani, M. N. M. Yassin, M. Z. Ibrahim, and N. Wahid, "Lane detection in autonomous vehicles: A systematic review," *IEEE Access*, vol. 11, pp. 3729–3765, 2023.

[6] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–12.

[7] Y. Song, L. Sun, J. Bi, S. Quan, and X. Wang, "DRGAN: A detail recovery-based model for optical remote sensing images super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.

[8] Z. Qin, W. Huanyu, and X. Li, "Ultra fast structure-aware deep lane detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 276–291.

[9] N. Garnett, R. Cohen, T. Pe'er, R. Lahav, and D. Levi, "3D-LaneNet: End-to-end 3D multiple lane detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2019, pp. 2921–2930.

[10] Y. Guo et al., "Gen-LaneNet: A generalized and scalable approach for 3D lane detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 666–681.

[11] V. Chamola, A. Chougule, A. Sam, A. Hussain, and F. R. Yu, "Overtaking mechanisms based on augmented intelligence for autonomous driving: Data sets, methods, and challenges," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 17911–17933, May 2024.

[12] M. Pittner, J. Janai, and A. P. Condurache, "LaneCPP: Continuous 3D lane detection using physical priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10639–10648.

[13] L. Chen et al., "PersFormer: 3D lane detection via perspective transformer and the OpenLane benchmark," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jul. 2022, pp. 550–567.

[14] Z. Chen, K. Smith-Miles, B. Du, G. Qian, and M. Gong, "An efficient transformer for simultaneous learning of BEV and lane representations in 3D lane detection," 2023, *arXiv:2306.04927*.

[15] R. Wang, J. Qin, K. Li, Y. Li, D. Cao, and J. Xu, "BEV-LaneDet: An efficient 3D lane detection based on virtual camera via key-points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1002–1011.

[16] Y. Ma et al., "Vision-centric BEV perception: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10978–10997, Dec. 2024.

[17] M. Li, Y. Zhang, X. Ma, Y. Qu, and Y. Fu, "BEV-DG: Cross-modal learning under bird's-eye view for domain generalization of 3D semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11632–11642.

[18] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. V. Gool, "Towards end-to-end lane detection: An instance segmentation approach," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 286–291.

[19] T. Zheng et al., "RESA: Recurrent feature-shift aggregator for lane detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3547–3554.

[20] X. Li, J. Li, X. Hu, and J. Yang, "Line-CNN: End-to-end traffic line detection with line proposal unit," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 248–258, Jan. 2020.

[21] L. Xiao, X. Li, S. Yang, and W. Yang, "ADNet: Lane shape prediction via anchor decomposition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6381–6390.

[22] L. Tabelini, R. Berriel, T. M. Paix ao, C. Badue, A. F. D. Souza, and T. Oliveira-Santos, "Keep your eyes on the lane: Real-time attention-guided lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 294–302.

[23] T. Zheng et al., "CLRNet: Cross layer refinement network for lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 898–907.

[24] Y. Ko, Y. Lee, S. Azam, F. Munir, M. Jeon, and W. Pedrycz, "Key points estimation and point instance segmentation approach for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8949–8958, Jul. 2022.

[25] Z. Qu, H. Jin, Y. Zhou, Z. Yang, and W. Zhang, "Focus on local: Detecting lane marker from bottom up via key point," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14122–14130.

[26] J. Wang et al., "A keypoint-based global association network for lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1382–1391.

[27] S. Xu et al., "RCLane: Relay chain prediction for lane detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2022, pp. 461–477.

[28] L. Tabelini, R. Berriel, T. M. Paix ao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "PolyLaneNet: Lane estimation via deep polynomial regression," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6150–6156.

[29] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3694–3702.

[30] Z. Feng, S. Guo, X. Tan, K. Xu, M. Wang, and L. Ma, "Rethinking efficient lane detection via curve modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17062–17070.

[31] R. Liu, D. Chen, T. Liu, Z. Xiong, and Z. Yuan, "Learning to predict 3D lane shape and camera pose from a single image via geometry constraints," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 2, pp. 1765–1772.

[32] Y. Bai, Z. Chen, Z. Fu, L. Peng, P. Liang, and E. Cheng, "CurveFormer: 3D lane detection by curve propagation with curve queries and attention," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 7062–7068.

[33] F. Yan et al., "ONCE-3DLanes: Building monocular 3D lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17143–17152.

[34] S. Huang et al., "Anchor3DLane: Learning to regress 3D anchors for monocular 3D lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17451–17460.

[35] Y. Luo et al., "LATR: 3D lane detection from monocular images with transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7907–7918.

[36] Y. Luo, S. Cui, and Z. Li, "DV-3DLane: End-to-end multi-modal 3D lane detection with dual-view representation," in *Proc. The 12th Int. Conf. Learn. Represent.*, 2024, pp. 1–11.

[37] Z. Zheng et al., "PVALane: Prior-guided 3D lane detection with view-agnostic feature alignment," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 7, pp. 7597–7604.

[38] S. Zheng, Y. Xie, M. Li, C. Xie, and W. Li, "A novel strategy for global lane detection based on key-point regression and multi-scale feature fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23244–23253, Dec. 2022.

[39] Z. Qiu, J. Zhao, and S. Sun, "MFIALane: Multiscale feature information aggregator network for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24263–24275, Dec. 2022.

[40] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[41] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[42] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[43] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[44] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, "Frequency-aware feature fusion for dense image prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10763–10780, Dec. 2024.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] C. Li, J. Shi, Y. Wang, and G. Cheng, "Reconstruct from top view: A 3D lane detection approach based on geometry structure prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4369–4378.

[47] Y. Bai, Z. Chen, P. Liang, B. Song, and E. Cheng, "CurveFormer++: 3D lane detection by curve propagation with temporal curve queries and attention," 2024, *arXiv:2402.06423*.

Yongchao Song was born in Weihai, Shandong, China, in 1990. He received the B.S. and Ph.D. degrees from the School of Electronic and Control Engineering, Chang'an University, Xi'an, China, in 2015 and 2020, respectively. He is currently an Associate Professor at the School of Computer and Control Engineering, Yantai University. His current research interests include remote sensing information processing, and deep learning.

Jiping Bi was born in Weifang, Shandong, China, in 2001. He received the B.S. degree from the School of Computer and Control Engineering, Yantai University, in 2023, where he is currently pursuing the master's degree. His current research interests include image processing, traffic target detection, and automatic control.

Lijun Sun was born in Yantai, Shandong, China, in 2001. She received the B.S. degree from the School of Computer and Control Engineering, Yantai University, in 2023, where she is currently pursuing the master's degree. Her current research interests include deep learning, remote sensing image super resolution, and artificial intelligence.

Zhaowei Liu (Member, IEEE) received the Ph.D. degree from Shandong University, Jinan, China, in 2018. He is currently a Professor with Yantai University, Yantai, China. His research interests include distributed artificial intelligence, machine learning, virtual reality, and collaboration with proposed software companies in directions, including digital twin and industrial meta-universe related research. He is a member of China Computer Federation (CCF).

Yahong Jiang was born in Baoding, Hebei, China, in 1992. She is currently pursuing the Ph.D. degree in transportation planning and management with the School of Transportation Engineering, Chang'an University. Her current research interests include transportation policy evaluation and self-driving electric vehicle promotion.

Xuan Wang (Senior Member, IEEE) was born in Weihai, Shandong, China, in 1991. She received the B.S. and Ph.D. degrees in traffic information engineering and control from Chang'an University, China, in 2013 and 2018, respectively. She is currently an Associate Professor at the School of Computer Science and Control Engineering, Yantai University. Her research interests include intelligent traffic control, artificial intelligence, and computer vision.