Exploiting the Complementarity of Bilateral Domains for Fast Lane Detection

Nan Ma, *Senior Member, IEEE*, Guilin Pang, and Yiu-Ming Cheung, *Fellow, IEEE*

***Abstract*—Lane detection plays a crucial role in the visual perception system of intelligent driving, aiming to rapidly identify various lane lines embedded in complex road scenarios. However, accurately and quickly detecting lane lines remains a challenging task, especially with the limited representation capacity of spatial domain. Using frequency to guide the few-visual-clue lane detection in spatial domain can be a cure, as frequency domain effectively describes sparse lane local contexts from a complementary perspective. To achieve accurate and fast lane detection, we propose a novel network that smoothly introduces frequency space into the spatial domain. We first design two light-weight modules, i.e., the Domain Transformation Module (DTM) and the Bilateral Aggregation Module (BAM), to explicitly perceive lane features with diverse semantics in bilateral domains. Concretely, the DTM excites lane local patterns in frequency space via a parallel sub-convolutions manner, while the BAM selectively absorbs informative components from the intra- and inter-domain perspectives. We then devise a small parametric module, named Position Refinement Module (PRM), to model fine-grained lane locations. It is instantiated into the last three stages of network to reconstruct detailed positional relationships by encoding global semantics and local contexts into unified lane embeddings. Extensive experiments on two widely-used datasets show that our method significantly outperforms the state-of-the-art approaches. Especially, our method achieves a superior inference efficiency of 0.011 second per image along with a total F_1 score of 79.28% on the CULane dataset.**

***Index Terms*—Intelligent vehicles, lane detection, bilateral domains, fine-grained modeling.**

I. Introduction

EFFICIENT lane detection is pivotal in the visual perception systems of intelligent transportation. It provides a series of significantly fundamental information for intelligent driving, road scene understanding, and advanced driver assistance systems. However, in the real world, various factors (e.g., extreme lighting condition and severe occlusions) cause lane lines to exhibit sparse visual appearance signals, making networks difficult to accurately predict these lines. Another practical dilemma is the strict requirement for model inference efficiency. Therefore, accurately and quickly detecting diverse lane lines remains a key challenge.

In the literature, the classical lane detection methods ([1], [2], [3], [4]) rely on highly predefined hand-craft features (e.g., colors and structures) to detect lanes. Consequently, they frequently struggle in complex road environments, i.e., nights, traffic jams and landmark noises. Recently, advanced convolution neural networks (CNNs) have achieved impressive success in lane detection, and deep learning-based methods can be roughly divided into three categories. Firstly, most papers ([5], [6], [7], [8], [9], [10], [11], [12], [13]) treat the lane detection task as a segmentation problem, such as semantic segmentation or instance segmentation. However, the numerous pixel-level prediction operations on the entire image are time-consuming. Secondly, some studies ([14], [15], [16], [17]) make attempts at adopting plenty of predefined anchors to regress real lanes of an image. Besides that, another line of work ([18], [19]) makes efforts to rapidly locate lanes by gridding the raw image into numerous cells, thereby converting pixel-wise prediction into row-wise classification. Through these approaches are simple and fast, their overall performance is relatively inferior due to the coarse feature representations. In general, most of them are either time-consuming or struggle to achieve higher detection accuracy. Thus, it is still challenging to develop algorithms that can accurately detect lane lines while maintaining a competitive inference speed.

Another crucial yet under-explored problem in lane detection is that existing methods often struggle to address the frequency bias, a phenomenon that entails severe risks. Firstly, the frequency bias of CNN models often causes them to overlook valuable lane local contexts during the feature extraction process. That is, CNN models are more sensitive to low-frequency information than the high-frequency signals, which may lead to CNNs being trapped in local optima ([20], [21]). Specifically, high-frequency responses come from local regions with significant grayscale variations, such as edges and textures, and encode rich details crucial for accurate detection of even unforeseen lane lines. For example, previous study ([22]) found that CNN models can correctly predict low-frequency removed images, but incorrectly predict high-frequency removed images. Thus, without the delicate control of frequency-domain cues flow into the

Received 6 September 2024; revised 16 March 2025 and 17 June 2025; accepted 4 August 2025. Date of publication 13 August 2025; date of current version 3 November 2025. This work was supported in part by the NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21 and Grant 62461160309; in part by the General Research Fund of RGC under Grant 12201321, Grant 12202622, and Grant 12201323; in part by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02; in part by the National Key Research and Development Program of China under Grant 2023YFF0615800; and in part by the National Natural Science Foundation of China under Grant 62371013. The Associate Editor for this article was Z. Pu. *(Nan Ma and Guilin Pang contributed equally to this work.) (Corresponding author: Yiu-Ming Cheung.)*

Nan Ma is with the School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: manan123@bjut.edu.cn).

Guilin Pang and Yiu-Ming Cheung are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: csglpang@comp.hkbu.edu.hk; ymc@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TITS.2025.3597352

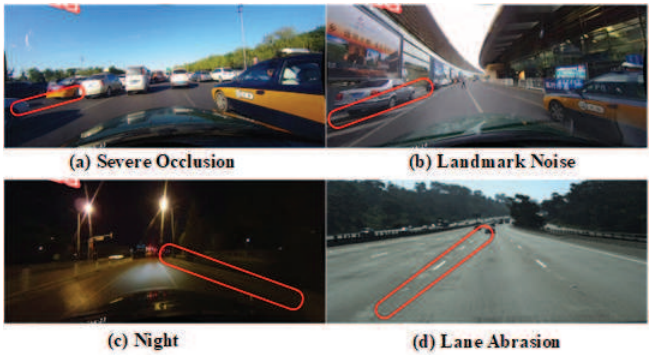

Fig. 1. Samples of few-visual-clue lane detection. In the real world, due to elongate prior shape and various factors (e.g., occlusion, extreme lighting condition and etc.), lane lines typically exhibit sparse visual appearances, resulting in subtle supervisory signals for lane detection.

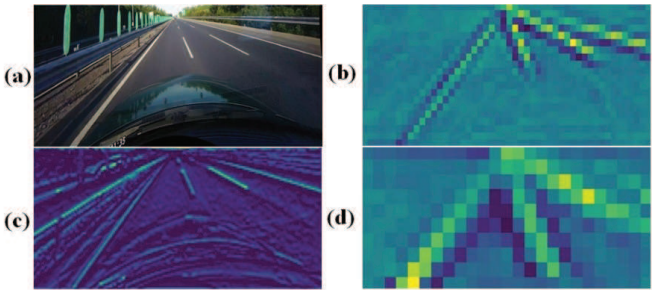

Fig. 2. Illustrations of various features with different distributions. (a) Original RGB Image. (b) Discriminative lane representations. (c) Local contexts with rich locational relations but full of noises. (d) Global information with ample semantics but lacks accurate position details.

CNN models, many discriminative high-frequency features will be missed, potentially leading to suboptimal detection performance. Secondly, global semantics suffer irreversible information loss and distortion because of the continuous down-sampling operation ([14]). As shown in Fig. 1, the visual appearances of lane lines are extremely subtle due to various factors. Furthermore, as illustrated in Fig. 2, the process of capturing global semantic information involves consecutive down-sampling operations, which significantly reduces the feature map resolution (Fig. 2 (d)) to $\frac{1}{1024}$ of the original image (Fig. 2 (a)). This substantial reduction in resolution leads to a considerable loss of detailed lane information, resulting in sparser supervisory signals and low detection accuracy ([23], [24]). In fact, inspired by human vision system, the accurate detection of few-visual-clue lane line is mainly relied on global semantics and local contexts. These two features are important to identify lane lines with elongated shape and weak visual appearance. To generate global semantics with large receptive field, down-sampling high-resolution images is a prevalent choice, but it inevitably incurs the loss of lane local contexts. All of the aforementioned lane detection methods reduce local information loss by solely reinforcing spatial domain features of RGB images. Nevertheless, it is not sufficient to consider the spatial domain only. In contrast, the frequency domain, with its inherent advantages in information preservation and edge enhancement, can better describe the local contexts of the lane. In addition, frequency spectrums come from local regions and thus encode abundant details with locational importance, which is beneficial for fine-grained lane modeling.

To tackle these problems, this paper presents a novel network for efficient lane detection by fully utilizing complementary features from bilateral domains. Specifically, to efficiently involve frequency space into spatial domain, we propose a Domain Transformation Module (DTM) and a Bilateral Aggregation Module (BAM), aiming to take full advantages of two different yet complementary clues: 1) frequency-aware local boundary, and 2) global semantics in spatial domain. To overcome the frequency bias issue and leverage neglected high-frequency details, we design the Domain Transformation Module (DTM) to explicitly activate various lane-relevant representations in frequency space. The parallel sub-convolution mechanism in the DTM can independently excite the lane-related local contexts from both low-frequency and high-frequency perspectives. By this mechanism, DTM greatly mitigates the frequency bias in networks by forcing them to focus more on valuable high-frequency signals. Then, to effectively fuse complementary information from bilateral domains while suppressing noise, we propose the Bilateral Aggregation Module (BAM). The symmetric interaction pattern in the BAM allows global semantics from the spatial domain and local contexts from the frequency space to mutually learn and collaborate for optimization. The DTM and BAM are devised as two small parametric modules to ensure superior detection speed. However, as shown in Fig. 2 (d), global information has rich semantics but lacks accurate location details, a coarse positional relation is not sufficient for accurate lane detection. Thus, to ensure high-precision localization while maintaining a lightweight design, we further introduce the Position Refinement Module (PRM), which refines coarse-grained location embeddings gradually. Particularly, features with different receptive fields are utilized to iteratively encode global semantics and local positional details into unified lane representations.

We demonstrate the effectiveness of our method on two popular datasets, i.e., CULane and TuSimple. To better compare the lane localization capabilities of diverse methods, we further report some visualization results on various complex road scenes. Ablation studies are conducted to evaluate the effectiveness of each module, it can be seen that our method can detect lane lines efficiently with the help of bilateral domains. Comprehensive experiments on two benchmarks demonstrate that ours method greatly outperforms other existing state-of-the-art methods in terms of both accuracy and efficiency. The main contributions are summarized as follows:

- We present a novel network for efficient lane detection that accurately and rapidly perceive lane representations by using frequency space to guide spatial-domain modeling process.
- We propose two small parametric modules, i.e., DTM and BAM, to explicitly excite frequency-aware local patterns as well as high-level global semantics from the bilateral domain perspectives.
- We further devise a light-weight PRM, aiming to encode features with different semantics into unified lane embeddings for fine-grained position modeling.

II. Related Work

A. Spatial-Based Lane Detection

With the development of intelligent driving techniques, the lane detection task has received significantly increasing attention. Numerous attempts have been made for spatial-based lane detection. The advantage of traditional methods is their ease of implementation, but they often encounter limited performance in challenging scenes as they heavily rely on strong prior hand-craft features of RGB images. ELDA [3] develops a light-wight algorithm based on Haar-liked features and predefined assumptions. HTPF [25] utilizes Statistical Hough Transform (SHT) followed by a Particle Filter (PF) to detect potential straight lane lines in RGB images. MHT [26] leverages multi-resolution Hough Transform to estimate the geometric structure of the lane boundaries. These highly specialized lane features work well only under ideal conditions but lack adaptive abilities to variations in road scenes. Consequently, these conventional methods suffer from a weakness of poor generalization in real-world scenarios.

Recently, convolution neural networks have manifested remarkable feature-representation capabilities across a diverse range of computer-vision tasks, involving traffic situation analysis [27], [28], [29], and lane detection. Therefore, we pay special attention to the deep learning methods, which can be broadly categorized into three groups: segmentation-based approaches, anchor-based approaches and other approaches. The first class of method tries to segment lanes from backgrounds, assigning a label to each pixel in the RGB image to indicate whether it belongs to a lane or not. Lane-Net [11] employs an instance segmentation network followed by a line fitting operation to locate lanes. FlipNet [30] developed a hierarchical feature flip fusion module (HFFF), a double-layer attention enhancement mechanism (DAEM) and a dual-pooling coordinate attention (DCA) to utilize spatial clues and aggregate global content. RESA [12] presents a recurrent feature-shift aggregator between encoder and decoder, which endows each pixel with the ability to gather global information. SAD [31] devises a knowledge distillation mechanism, enabling the model to learn from itself and earning notable performance progress. SCNN [10] proposes a message passing mechanism to generalize common layer-by-layer convolution to slice-by-slice convolution, which enhances the spatial feature extraction capacity for accurate lane detection. However, segmentation-based approaches are time-consuming as they conduct dense prediction on the high-resolution feature maps.

The anchor-based methods first predefine an extensive array of lines and subsequently perform dense regression to predict the position of real lanes. Similar to regular object detection, LaneATT [17] employs substantial anchors in the feature pooling phase. Furthermore, a anchor-based attention module is proposed to gather global information. CLRNet [14] is designed to fully utilize both high-level and low-level features in a cross-layer refinement fashion. In addition, it introduces the Line IoU loss to regress the entire lane line as a cohesive unit, thus achieving high localization accuracy. CLRerNet [13] reveals that confidence scores capable of precisely measuring the intersection over union (IoU) with ground-truths are most advantageous for accurately localizing lane positions. Consequently, it devises the LaneIoU metric. This metric is formulated by factoring in local lane angles, with the aim of enhancing the quality of confidence scores. O2SFomer [16] develops a one-to-several label assignment strategy to mitigate label semantic conflicts and enhance training efficiency of DETR (DEtection TRansformer). Although anchor-based lane representations are adequate for most real-world lanes, they may encounter challenges in demanding environments due to their close association with strong priors.

In the last category of approaches, various endeavors, such as gridding and polynomial fitting, are employed to enhance the accuracy of lane detection. Ultra-Fast [18] first devises a simple yet effective prediction formulation, which treats lane detection as row-based classification task. Benefiting from row-based global features selection, it gains extremely fast inference speed. OConv [32] proposes an oblique convolution method, including Oblique Rotation Module, Strip Spatial Attention Module and Anti-oblique Rotation Module, to break up the limitations of ordinary convolution in extracting lane information and make the network focus on the strip lanes. PGA-Net [33] utilizes a transformer-based DETR model to regress parameters of cubic polynomial function used for modeling lane shape. PolyLaneNet [34] represents each lane lines in input image by regressing polynomial parameters. Due to fewer parameters to regress, PolyLaneNet [34] is able to maintain its high efficiency. FastDraw [35] introduces a fully convolution model to directly decode lane structure without the need for post-processing operations. However, exploiting lane features solely in spatial domain is not sufficient, as the frequency bias may cause CNN models to overlook valuable high-frequency clues. All these lane detection approaches merely reinforce spatial-domain features of RGB images, which may result in sub-optimal performance.

B. Frequency-Domain Feature Learning

Frequency analysis is of great importance for digit image processing and has already been extensively used in computer vision community, encompassing tasks such as classification, semantic segmentation, and image compression and encoding. LFA [36] demonstrates that images generated through GANs (Generative Adversarial Networks) appear highly photorealistic in the RGB domain but exhibit severe artifacts in the frequency space. Thus, they perform a comprehensive analysis in frequency domain to recognition forged images. LFD [20] proposes a learning-based frequency selection strategy to removal trivial frequency signals without accuracy loss. FNN [37] captures powerful features from frequency domain for fast image classification. OFD [38] designs a powerful network to introduce frequency as an additional clue to better detect various objects from their camouflaged environments. DRL [39] designs a model conversion approach to transform spatial-domain CNN models into the frequency domain. Despite frequency analysis has made notable success in previous methods, how to utilize frequency domain for accurate lane detection is still under explored. In contrast to previous methods, we present two well-designed decoders to explicitly model both high- and low-band frequency clues,

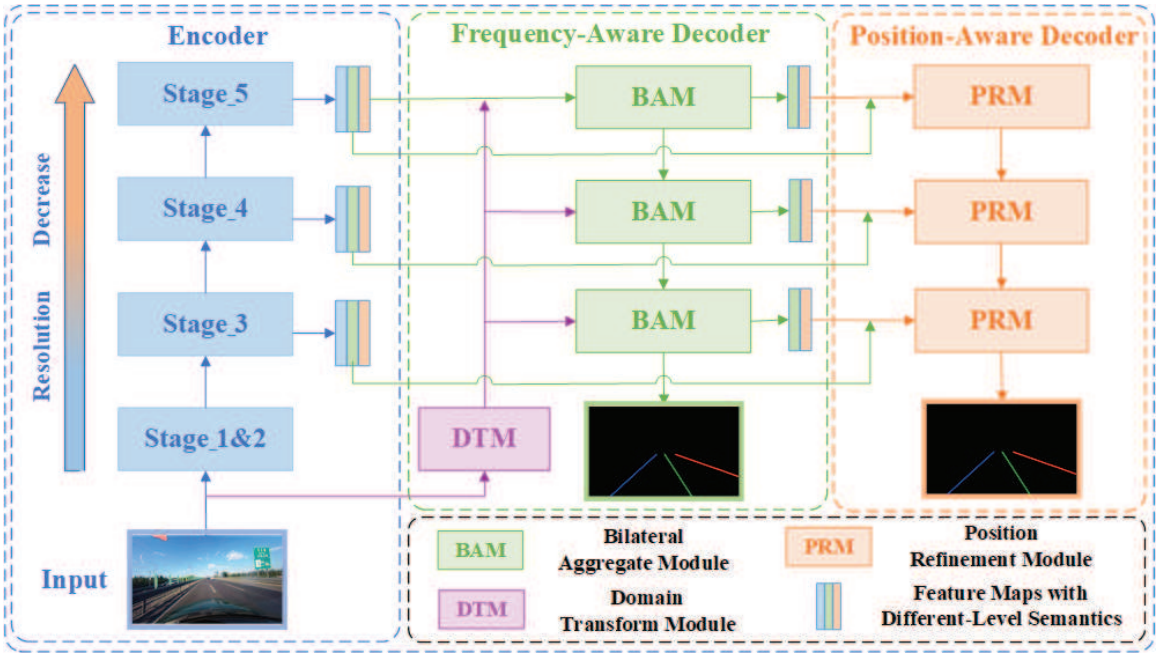

Fig. 3. Overview of our proposed method. Notably, given the demand of fast inference, the prediction generated by the Position-aware decoder serves as the sole final output throughout the inference phase. In contrast, the components, excluding the DTM and BAM, responsible for generating the output of the Frequency-aware decoder are activated solely during the training stage.

guiding the spatial-domain lane features learning. In this way, the proposed method realizes remarkable performance since it can perceive more comprehensive lane indicators from both intra- and inter-domain perspectives.

III. Proposed Method

In this section, we describe the details of our proposed Network, including Domain Transformation Module (DTM), Bilateral Aggregation Module (BAM) and Position Refinement Module (PRM). First, in Sec. III-A, we outline our method to provide a concise yet accurate definition of crucial concepts. Subsequently, three core components (DTM, BAM and PRM) are elaborated in Sec. III-B, III-C, and III-D, respectively.

A. Overview

The overall architecture of our method can be seen in Fig. 3, which comprises a shared and light-weight encoder and two small parametric decoders, i.e., frequency-aware decoder and position-aware decoder. In particular, a DTM and three BAMs are instantiated into frequency-aware decoder, while the stacked PRMs are integrated into position-aware decoder. It is an efficient lane detection network designed to cope with aforementioned frequency bias as well as irreversible information loss problems. This is achieved by utilizing frequency space signals to guide spatial-domain modeling process.

Firstly, the RGB inputs are fed separately into the spatial path of encoder and frequency path of DTM. Specifically, the former path utilizes a light-weight encoder to extract multi-level spatial-domain features from RGB inputs, i.e., $S_i (i \in \{1, 2, 3, 4, 5\})$. Meanwhile, in the latter path, the DTM is employed at the early stage to excite lane local patterns in frequency space and removal trivial noise signals. Subsequently, the BAM is built to produce more comprehensive lane representations D_i by selectively absorbing features from bilateral domains. This mechanism guides our model to prioritize high-frequency local boundaries and high-level global semantics. Then, to model fine-grained lane position information, the PRM works in a top-down manner, iteratively refining previous coarse features D_i and S_i with more detailed local contexts.

In the output of first decoder, frequency-aware decoder, we assign each pixel with a probability score to indicate whether it belongs to a lane or not. However, considering inference efficiency, this pixel-wise prediction operation is only conducted during the training phase to more effectively supervise model convergence. Meanwhile, for the output of second decoder, position-aware decoder, each lane comprises four types of elements: (1) lane confidence score c, (2) start point p_s and direction angle Θ, (3) end point p_e, and (4) the offset for each point between p_s and p_e.

B. Domain Transformation Module

High-frequency image components are strongly correlated to landmarks with intense grayscale variations like lane line edges. However, the frequency bias of CNN models causes previous methods to predominantly focus on low-frequency content in RGB image, yielding an inaccurate lane detection performance. To this end, the DTM is designed to explicitly perceive lane local contexts in frequency domain.

As shown in Fig. 4, it encompasses both a base filter and two enhanced filters. Concretely, an RGB image $x \in R^{H \times W \times 3}$ is converted to the YCbCr color space and then separated into a group of 8×8 blocks (standard block resolution in image compression). Subsequently, a base filter t_{base}, comprising the discrete cosine transform (DCT) and two convolution layers,

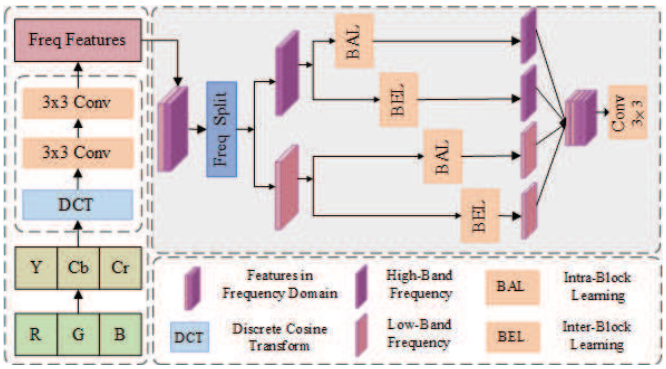

Fig. 4. Illustration of the presented Domain Transformation Module (DTM).

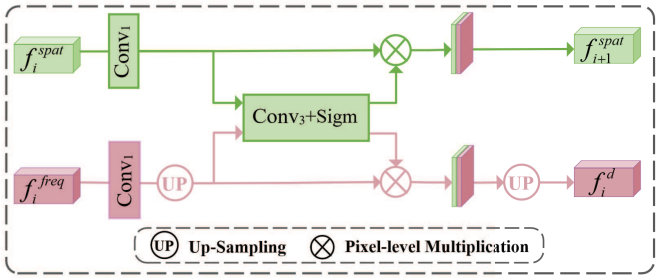

Fig. 5. The architecture of the Bilateral Aggregation Module (BAM). The symbol f^{spat} denotes the spatial-domain features that are extracted by the employed backbone, while f^{freq} represents the frequency-domain information that is obtained from DTM. Notably, the index i, where $i \in [1, 3]$, corresponds to the i^{th} BAM counted from the top in Fig. 3. Specifically, $i = 1$ refers to the top-most BAM, while $i = 3$ denotes the bottom-most one.

is performed on each block to generate the corresponding frequency spectrum $f_{i,j}, i \in \left[1, \frac{H}{8}\right], j \in \left[1, \frac{W}{8}\right]$. Therefore, the frequency domain of RGB inputs can be obtained by:

$$f_{i,j} = t_{base}(\varphi(x)), \qquad f_{i,j} \in R^{8\times8\times3} \tag{1}$$

where φ denotes color transformation and image decomposition operations. We take DCT as the first unit of base filter t_{base} due to its excellent organization of the frequency distribution. To be specific, in its frequency spectrum, the left-top corner encompasses low-frequency signals while the right-bottom corner gathers high-frequency responses. Moreover, we employ two convolution layers at the last phase of base filter t_{base} to suppress noises of frequency spectrum $f_{i,j}$. All $f_{i,j}$ are flatten and decomposed into low- and high-band frequencies (denoted by $f^{low/high}$) based on their intensities:

$$f^{low/high} = \Pi(flatten(f_x)), f^{low/high} \in R^{\frac{H}{8}\times\frac{W}{8}\times192} \tag{2}$$

where Π is frequency band partition. However, the base filter t_{base} might not be powerful enough to adaptively capture the dynamic variations of lanes in complex scenarios.

To this end, two enhanced filters t_{enh} are added to the base filter t_{base} to improve the adaptivity of frequency-domain signals. Each t_{enh} is equipped with an intra-Block Learning (BAL) unit and an inter-Block Learning (BEL) unit to comprehensively exploit contextual relationships of local regions. To excite the intra-block short-range signals, a weight matrix W_m is computed in BAL by utilizing a popular Self-Attention Excitation (SAE) mechanism ([40]). After that, a matrix product between W_m and $f^{low/high}$ is conducted to express its impact on original frequency $f^{low/high}$, i.e.,

$$\begin{aligned} W_m &= SAE(f^{low/high}) \\ f_a^{low/high} &= f^{low/high} \otimes W_m \end{aligned} \tag{3}$$

in which $\otimes$ represents multiplication. To enhance the inter-block long-range dependencies, BEL performs global average and maximum operations separately within each frequency block of $f^{low/high}$. Subsequently, the results of these operations are concatenated and processed by two convolutional layers and a sigmoid function to encode cross blocks interactions, i.e.,

$$\begin{aligned} f^{avg/max} &= Conv_3(\mathcal{C}(\rho_a(f^{low/high}), \rho_m(f^{low/high}))) \\ f_e^{low/high} &= \delta(Conv_1(f^{avg/max})) \end{aligned} \tag{4}$$

where $\mathcal{C}$ stands for concatenation, δ denotes the sigmoid function, ρ_a and ρ_m represent the global average and maximum operations, respectively. The final output f^f of DTM is obtained by injecting separated low-band frequencies $f_{a/e}^{low}$ into high-band frequencies $f_{a/e}^{high}$:

$$f^f = Conv_3(\mathcal{C}((f_a^{low} + f_e^{low}), (f_a^{high} + f_e^{high}))) \tag{5}$$

where + and $\mathcal{C}$ mean pixel-wise addition and concatenation respectively. The convolution with kernel size 3×3 is applied to eliminate the aliasing effect of aggregated frequencies.

C. Bilateral Aggregation Module

The DTM endows our model with the capacity for frequency awareness, but the information captured by the frequency path ends up having substantial background noises. For example, some landmarks (e.g., rotation arrows) always share similar structures with lane lines, making it hard to identify these lines via isolated frequency clues. In contrast, the aforementioned spatial path has captured global semantics that is crucial to distinguish lane lines from noise landmarks. Consequently, we propose the BAM to selectively absorb useful features from both spatial domain and frequency domain.

As depicted in the BAM of Fig. 5, we employ a 1×1 convolution layer $Conv_1$ to halve channels of spatial-domain feature f^{spat}. Simultaneously, in the parallel branch, a $Conv_1$ followed by an up-sampling operation is utilized to ensure that the shape of the frequency-domain feature f^{freq} becomes identical to that of f^{spat}. We then stack f^{spat} and f^{freq} along channel axis to initially gather features from different domains. The forward processes are formulated as:

$$f^{tmp} = \mathcal{C}(Conv_1(f^{spat}), \gamma Conv_1(f^{freq})) \tag{6}$$

in which γ represents bilinear interpolation. Next, to generate the cross-domain interaction matrix W_d, we perform a 3×3 convolution and a sigmoid function on aggregated features f^{tmp}, which is written as:

$$W_d = \delta(Conv_3(f^{tmp})) \tag{7}$$

where δ is defined as sigmoid function, aiming to project each value of f^{tmp} into the range between 0 and 1. To dynamically adjust the information intensity of each domain, we obtain

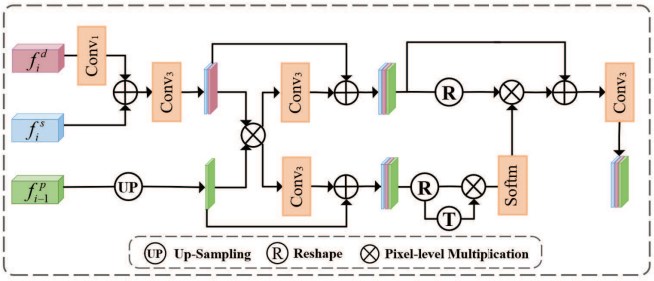

Fig. 6. The overall structure of the Position Refinement Module (PRM). The PRM takes three distinct features as inputs: the spatial feature f^s from the employed backbone, the frequency signal f^d from the proposed BAM, and the refined feature f^p from the previous PRM. The index i, where $i \in [1, 3]$, corresponds to the i^{th} PRM counted from the top in Fig. 3. Notably, in the initial PRM (i.e., $i = 1$), the pathway for f_{i-1}^p is pruned as there is no preceding PRM to provide this refined feature.

intra-domain correlation matrix $W_{1/2}$ by dividing W_d along domain dimension. After that, two pixel-wise multiplications are executed between original $f^{spat/freq}$ and $W_{1/2}$ with a lateral connection, which can be written as:

$$f_a^{spat/freq} = f^{spat/freq} \otimes \prod_d W_d \tag{8}$$

in which $\prod_d$ denotes domain decomposition. The new embedding $f_a^{spat/freq}$ can be regarded as an abrupt enhancement within each domain, overlooking the interrelationship among domains. Therefore, we reinforce cross-domain feature consistency by adding f_a^{spat} and f_a^{freq}: $f^d = f_a^{spat} + f_a^{freq}$. The design of BAM exhibits a symmetric architecture, where f^{freq} encodes its invisible frequency into f^{spat} and f^{spat} filters non-lane landmarks of f^{freq}. It enables our network to be aware of frequency-domain features in spatial-domain lane modeling, resulting in better overall performance.

D. Position Refinement Module

With the help of frequency-aware decoder, our model can already achieve a competitive performance via global semantics. However, due to sequential down-sampling, RGB inputs suffer irreversible information loss, particularly in lane details. Consequently, it fails to accurately locate lanes according to a coarse-grained position representation. As we discussed in Sec. I, global information has rich semantics but lacks accurate position details. Contrarily, local contexts are closely bound to ample location relations, which are widely distributed in frequency domain and shallowed spatial features. To facilitate this, we propose a position refinement module (PRM) to iteratively explore fine-grained lane position information.

As shown in Fig. 6, we first apply a 1×1 convolution operation to f^d and perform element-wise addition between f^s and channel-reduced f^d, followed by a 3×3 convolution as described in Eq.(9):

$$f^{sd} = Conv_3(f^s + Conv_1(f^d)) \tag{9}$$

Afterward, f_p needs to be expanded to the corresponding shape for subsequent information interaction, as f^{sd} has distinctive resolutions. Next, we use a pixel-wise multiplication to extract feature consistencies between f^p and f^{sd} and further import them into original f^p and f^{sd} via an element-wise addition and two 3×3 convolutions. It can be formulated as:

$$f_p^{sd} = Conv_3(f^{sd} + Conv_3(f^{sd} \otimes f^p)) \tag{10}$$

$$f_{sd}^p = Conv_3(f^p + Conv_3(f^p \otimes f^{sd})) \tag{11}$$

In this way, f^p and f^{sd} can gradually absorb discriminative indicators from each other to complement themselves, i.e., the redundant features of f^p are removed and the lane boundaries of f^{sd} are refined.

In addition, we further excite position-aware feature components in cross-scale semantics and frequency responses by several simple yet effective operations. We reshape the $f_{sd}^p \in R^{C \times H \times W}$ into 2-dimensional space $R^{C \times N}$, where N equals to $H \times W$. Then, we conduct matrix product between reshaped f_{sd}^p and its transpose $(f_{sd}^p)^T$ to generate position-aware weights $W_p \in R^{N \times N}$, followed by a Softmax function:

$$W_p = Softmax(r(f_{sd}^p)^T \otimes r(f_{sd}^p)) \tag{12}$$

where r and T function as matrix reshape and transpose, respectively. W_p summaries the positional importance of cross-scale features from spatial and frequency domains. To strengthen original feature, we carry out an element-wise multiplication between W_p and reshaped f_p^{sd}. Subsequently, the results are reshaped back to 3-dimensional space $R^{C \times H \times W}$ and combine with f_p^{sd} using a learnable scalar parameter a:

$$f_{final} = a * f_p^{sd} + r(r(f_p^{sd}) \otimes W_p) \tag{13}$$

As can be seen in Eq.(13), the final features f^{final} of PRM are obtained by weighted summation, which are powerful enough to encode fine-grained lane position information. By taking account of this mechanism, local contextual relationships are reinforced and the discriminability of global semantics are augmented.

IV. Experiments

This section starts introducing two widely-used lane detection benchmarks (e.g., CULane [10] and TuSimple [41]). The second subsection provides a detailed description of the implementation of our method, encompassing data preparation, model training, and evaluation metrics. Subsequently, to evaluate the effectiveness of the proposed network, we conduct comparative experiments on two challenging datasets and report corresponding results in Sec. IV-C. The two final subsections introduce the analysis of visualizations and conduct an ablation study on components, respectively.

A. Dataset Description

TuSimple: TuSimple [41] is regarded as one of the most popular benchmarks for lane detection. Due to its stable scenes and appropriate scale, this dataset is extremely suitable for swiftly verifying the performance of lane detection models. As depicted in Table I, the entire dataset is derived from a common road scene, with the goal of advancing research in lane detection, particularly on highways. Besides, TuSimple contains 6,408 raw images, each of them has 1280×720 pixels. Following the dataset's creator, we take the annotated 3,268

TABLE I
BASIC INFORMATION OF TWO POPULAR LANE DETECTION DATASETS

Dataset	Train	Validation	Test	Resolution	Scenes
TuSimple [41]	3268	358	2782	1280x720	1
CULane [10]	88880	9675	34680	1640x590	9

images for training, 358 clips for validation, and the remaining 2,782 frames for testing.

Subsequently, for TuSimple dataset, we apply official metrics (e.g., Accuracy, FP and FN) to fairly compare our algorithm with the state-of-the-art approaches. The accuracy is defined as:

$$Accuracy = \frac{C_i}{S_i} \tag{14}$$

in which C_i is the total number of correctly predicted lane line points and S_i is the total number of lane points in ground truth.

CULane: CULane [10] is a large lane detection dataset, which is collected under dynamic lighting conditions in nine different road scenarios (i.e., night, normal, crowded, arrow and etc.). As shown in Table I, CULane typically provides 133,235 frames with a resolution of 1640×590, in which 88,880 are used for training, 9,675 for validation and 34,680 for testing. Therefore, compared to TuSimple benchmark, it is more challenging in terms of scale and complexity.

Following the most lane detection researches [10], [12], [14], we take F_1-measure as the CULane metric and calculate the Intersection-over-Union (IoU) between predictions and labels. Furthermore, the True Positives (TP) is considered as predicted lanes with an IoU above 0.5. The F_1 is calculated by:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{15}$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$, FP and FN denote false positive and false negative respectively.

B. Implementation Details

Following previous works (SAD [31], CLR-Net [14], etc.), we initially crop 270 pixels along the height axis for CULane frames and 160 pixels for TuSimple images. This aims to removal the sky parts in RGB inputs, as they do not contain any lane clues. Then, we resize all cropped images of TuSimple and CULane to 320×800 to improve computational efficiency and save memory usage. Subsequently, we apply random data augmentations (e.g., horizontal flips, motion blur, etc.) on reshaped images to increase sample diversity.

For the implementation of the encoder, we first adopt light-weight pretrained ResNet-18 [47] as the backbone of spatial path to capture global semantics and enlarge receptive field iteratively. Meanwhile, in the frequency path, DTM employs Discrete Cosine Transform (DCT) and hierarchical sub-convolutions to adaptively excite frequency-aware signals. To map the feature maps of the last BAM to pixel-wise segmentation outputs, a 1×1 convolution layer followed by an up-sampling operation is employed behind the last BAM of the frequency-aware decoder. Notably, these operations are exclusively activated during the training stage, as instance segmentation is both time-consuming and computationally expensive. Finally, the position-aware decoder is constructed by stacking PRM in a one-by-one manner, followed by two fully connected layers to predict the representation of lane lines, involving 1) lane confidence scores, 2) start points and direction angles, 3) end points, 4) offsets between predictions and their ground truth.

For the frequency-aware decoder, the Negative Log-Likelihood Loss is employed to supervise the dense prediction task, which is only used for calculating the segmentation loss L_{seg}. In contrast, for the position-aware decoder, we utilize Focal Loss [48] and smooth L1 [49] to compute the classification loss L_{cla} and regression loss L_{reg}. The overall loss function consists of an equivalent sum of classification, regression and segmentation losses. That is, $L_{total} = L_{cla} + L_{reg} + L_{seg}$.

In the training process, we use Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of $1e-5$ as the model optimizer, while the initial learning rate is set to $1.4e-3$ for TuSimple and $0.8e-3$ for CULane, respectively. To dynamic adjust learning rate, we adopt 'CosineAnnealingLR [50]' strategy with the minimum learning rate set to $2e-6$. The total number of training epochs is set to 20 for CULane and 55 for TuSimple. All experiments are implemented using PyTorch 1.9 and conducted on a machine with an NVIDIA RTX 3090 (24G) GPU.

C. Comparison With the State-of-The-Art Methods

To generate convincing evaluations, we compare the proposed method with numerous existing state-of-the-art models on CULane [10] and TuSimple [41] datasets. The quantitative experimental results of our models, i.e., ResNet [47] version and ConvNext [51] version, are summarized in Table II-III. In addition, we also report the runtime of our method against other algorithms, as speed holds comparable importance with accuracy in lane detection. For the sake of fairness, the runtime is determined by averaging the inference speed over 1000 images on a single GPU.

1) Performance Comparison on TuSimple: For TuSimple benchmark, nine popular lane detection algorithms, including UltraFast-V2 [19], RESA [12], OConv [32], LaneATT [17], LaneNet [11], SAD [31], CLLD [46], FastDraw [35] and PolyLaneNet [34], are used for comparison. Using the standard metrics, we report the performance of our models in terms of Accuracy, FP and FN and summarize the results in Table III. As shown in Table III, our method outperforms the existing state-of-the-art (SOTA) ones, with the beat accuracy of 96.83%. However, small accuracy variations among different detectors strongly prove that the results of TuSimple are saturated already. This phenomenon may be caused by its elementary road scene and permissive metric. For lane detection, the total number of positive samples is far less than that of negative samples. This means that lane detectors are easy to exhibit a high accuracy but with unsatisfactory lane location results. For a comprehensive evaluation, we also verify our models via FP and FN. It is clear that our ConvNext version gains the lowest FP and FN, at 0.0217 and 0.0189 respectively. Consistently, the lower FP and FN

TABLE II

STATE-OF-THE-ART RESULTS ON CULANE [10] DATASET. IN THE CASE OF CROSS, ONLY FP IS SHOWN. NOTING THAT "RUNTIME" MEANS THE MODEL INFERENCE EFFICIENCY ON SINGLE IMAGE

Method	Normal	Crowded	Dazzle	Shadow	No line	Arrow	Curve	Cross	Night	Total	RunTime (ms)
LaneAF [42]	91.10	73.32	69.71	75.81	50.62	86.86	65.02	1844	70.90	75.63	41.7
SpinNet [43]	90.50	71.70	62.00	72.90	43.20	85.00	50.70	–	68.10	74.20	–
FastDraw [35]	85.90	63.60	57.00	59.90	40.60	79.40	65.20	7013	57.80	–	11.1
SCNN [10]	90.60	69.70	58.50	66.90	43.40	84.10	64.40	1990	66.10	71.60	116
ENet-SAD [31]	90.10	68.80	60.20	65.90	41.60	84.00	65.70	1998	66.00	70.80	13.3
CurveLane [44]	90.70	72.30	67.70	70.10	49.40	85.80	68.40	1746	68.90	74.80	–
RESA [12]	91.90	72.40	66.52	72.00	46.30	88.10	68.60	1896	69.80	74.50	22.0
UltraFast [18]	90.70	70.20	59.50	69.30	44.40	85.70	69.50	2037	66.70	72.30	**5.9**
MECNet [45]	89.60	67.10	59.90	60.30	41.80	83.00	61.40	2071	61.90	69.50	14.5
PGA-Net [33]	87.84	70.00	62.11	67.61	46.71	80.94	58.01	1700	59.02	69.86	6.9
LaneATT [17]	91.74	76.16	69.47	76.31	50.46	86.29	64.05	1264	70.81	77.02	38.5
FlipNet [30]	92.4	72.6	65.1	74.3	47.2	88.6	67.9	1432	69.8	75.0	–
Ours (Res18)	93.37	77.68	**73.60**	78.64	52.42	90.02	68.04	**1204**	74.83	79.28	10.9
Ours (ConvNext-S)	**93.47**	**78.16**	73.56	**81.88**	**54.36**	**90.66**	**70.63**	1282	**75.04**	**79.71**	16.1

TABLE III

STATE-OF-THE-ART RESULTS ON TUSIMPLE [41] DATASET. THE BEST RESULT IS SHOWN IN BOLD. IT IS NOTEWORTHY THAT ↑ INDICATES SUPERIOR PERFORMANCE FOR MODELS WITH HIGHER VALUES IN THIS COLUMN, WHILE ↓ SIGNIFIES BETTER PERFORMANCE FOR MODELS WITH LOWER SCORES IN THEIR RESPECTIVE COLUMNS

Method	Accuracy↑	FP↓	FN↓
LaneNet [11]	93.38	0.0780	0.0224
RESA [12]	96.70	0.0395	0.0283
LaneATT [17]	96.10	0.0564	0.0217
CLLD-UNet [46]	96.17	-	-
FastDraw [35]	94.90	0.0610	0.0470
ENet-SAD [31]	96.64	0.0602	0.0205
PolyLaneNet [34]	93.36	0.0942	0.0933
UltraFast-V2 [19]	95.65	0.0306	0.0461
OConv [32]	96.50	0.0875	0.0312
Ours (Res18)	96.73	0.0277	0.0195
Ours (ConvNext-T)	**96.83**	**0.0217**	**0.0189**

scores contribute to higher location precision. Nevertheless, it further demonstrates the effectiveness of our method because the proposed approach could iteratively perceive both global semantics and local contexts from bilateral domains.

2) Performance Comparison on CULane: To further evaluate the efficiency and effectiveness of our method, we compare it against twelve state-of-the-art methods, including LaneATT [17], UltraFast [18], RESA [12], PGA-Net [33], FlipNet [30], MECNet [45], CurveLane [44], SAD [31], LaneAF [42], SCNN [10], FastDraw [35], SpinNet [43]. The quantitative results of the CULane dataset are described in Table II. It can be observed from Table II that our models achieve compelling scores across nine distinctive scenarios in comparison with other counterparts, as indicated by the two metrics. It is evident that our ConvNext-Small version achieves a total F_1-measure of 79.71%, which significantly surpasses the second-best method with a 2.69% improvement. This substantiates the superior robustness and generalization of the proposed methods across various challenging environments. In addition, these compared algorithms reinforce spatial features by numerous techniques, but the overall performance is still unsatisfying. One possible explanation is that substantial lane details may be permanently lost due to long-distance downsampling and the frequency bias of CNN models. In contrast, it is worth noting that the F_1-measure of our model outperforms those of other algorithms by a large margin in night and dazzle categories, with a score improvement of 4.14% and 3.89% respectively. We attribute this remarkable improvement to the extracted frequency-domain features and the subsequent denoising operations. Notably, these features can effectively separate brightness from color information, which is known to be particularly advantageous for detecting lanes in scenarios with significantly varying environmental lighting conditions. Specifically, the frequency-domain features are extracted by successively applying a conventional color-space conversion operation, a fixed discrete cosine transform (DCT) filter, and two learnable units. Then, the bilateral aggregation module (BAM) is specifically designed to use spatial-domain features with large receptive fields to adaptively suppress the substantial background noises in frequency-domain signals, such as turn landmarks on roads, crosswalk and etc.

In addition to F_1-measure, we further compare the efficiency of our method with other state-of-the-art approaches, as model inference speed is of equal importance to location precision.

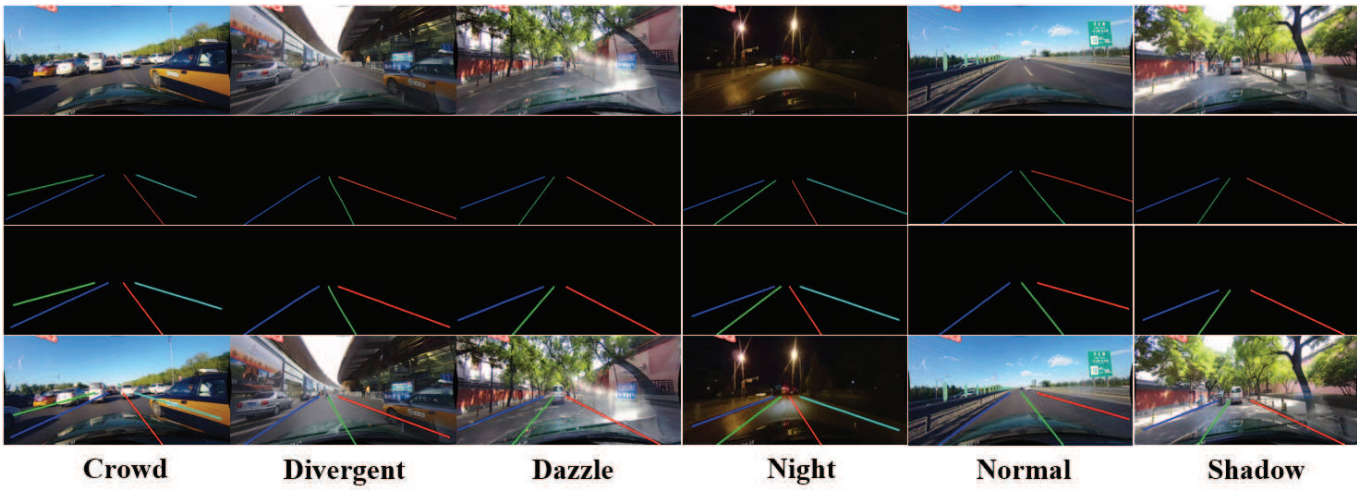

Fig. 7. Our visualization results on various challenging scenarios, i.e. crowd, night, dazzle and etc. The first three rows from top to bottom represent original images, official labels, and our results, respectively. Additionally, the last row involves mapping our results back to the original RGB inputs. In the above examples, our method achieves IoU scores of 0.72 (crowd), 0.79 (divergent), 0.71 (dazzle), 0.60 (night), 0.73 (normal), and 0.82 (shadow) across the respective challenging scenarios. Notably, following [10], [14], the IoU threshold is set to 0.5, where samples with an IoU exceeding this threshold were classified as correctly predicted.

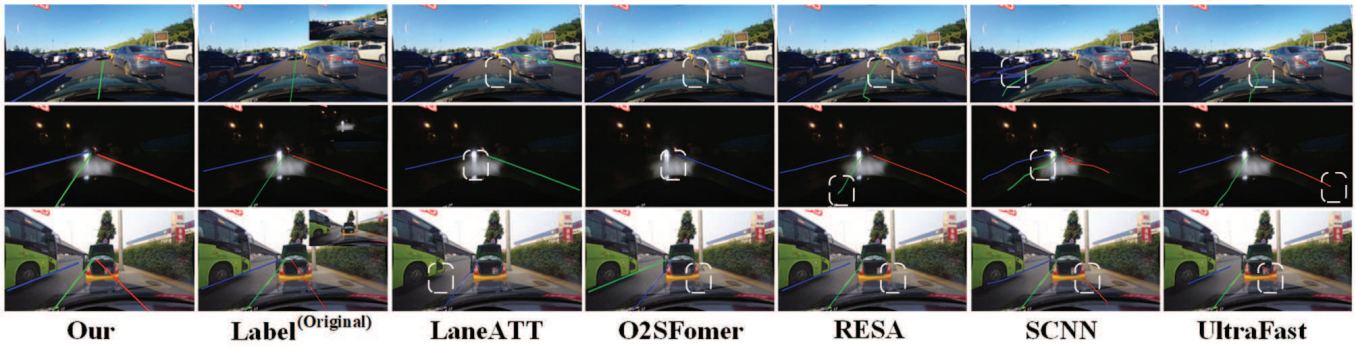

Fig. 8. Visual comparisons of our method with other five SOTA approaches, including LaneATT [17], O2SFomer [16], RESA [12], SCNN [10], UltraFast [18]. Notably, the white squares with dashed lines in the figure indicate specific instances where the existing methods fail to accurately detect the lane lines.

In Table II "RuntTime", the proposed method occupies the third-best and UltraFast is the fastest algorithm. However, by observing the F_1-measure metric, we find that UltraFast [18] only acquires a score of 68.4%, which drops behind our model (Res18 version) by a large margin of 6.98%. Moreover, SCNN attains a F_1 score of 71.60 with the run time of 116 ms. The overall performance of SCNN is not satisfactory due to its notable slow runtime. In contrast, our method obtains the best F_1 score while maintaining a relatively high inference speed (10.9 ms per image). In general, benefiting from the unified embeddings with sufficient frequency and position-aware information, our models obtain outstanding performance in terms of location precision and efficiency.

D. Visualizations

Numerous challenging examples are provided in Fig. 7. We observe that the proposed method not only has a powerful ability to accurately detect lanes in these challenging scenarios, but also well preserves their continuity and smoothness. In the example of the challenging "Crowd" scenario, our network effectively detects lanes with extremely sparse appearances, achieving an average IoU of 0.72 by leveraging global semantic cues (e.g., vehicle flow direction) and frequency-aware information. This demonstrates that the proposed method can definitely gather global semantics and local contexts by learning in both spatial and frequency domains. In general, our model performs a decent job at accurately localizing lanes, which is illustrated in Fig. 7.

To further verify the proposed approach, we compare some visualizations generated by our method and other counterparts in Fig. 8. Compared with other approaches, our algorithm shows superior performance in terms of lane integrity and smoothness. For example, in the first two rows of Fig. 8, our approach accurately captures the lane lines, whereas the lines detected by other methods exhibit either incomplete or jagged, particularly in complex scenarios with crowded vehicles and varying lighting conditions. There are two potential reasons: 1) our method can more accurately and comprehensively predict various lanes, as the frequency information contributes to a better description of local contexts; 2) the frequency-aware signals and spatial features can mutually enhance each other to form a more comprehensive lane representation, thereby

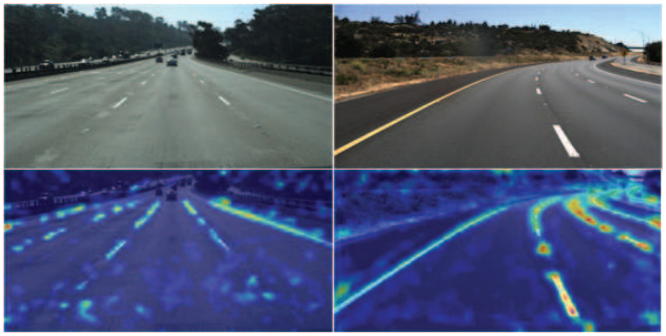

Fig. 9. Classification activation maps (Grad-CAM [52]) on TuSimple and CULane benchmarks.

TABLE IV
ABLATION STUDY ON COMPONENTS (INDICATED BY ✓) OF OUR METHOD. THE IOU THRESHOLD OF F_1 IS SET TO 50 AND 80, RESPECTIVELY

Model	DTM+BAM	PRM	CULane		TuSimple	
			F_1@50	F_1@80	Acc	Parameter (M)
M_1			78.10	49.51	95.21	+0.00
M_2	✓		78.63	50.35	96.42	+0.74
M_3	✓	✓	79.28	51.45	96.73	+1.49

endowing our model with stronger capacity to suppress cluttered noises. This is evidenced by numerous visual results in Fig. 8. In general, our model exhibits superiority on account of perceiving dynamic lanes with subtle appearance.

As shown in Fig. 9, we adopt Grad-CAM [52] to visualize feature distribution of our method, aiming at reveal the determinant evidence that impacts final lane prediction. The top row comprises original images from the TuSimple and CULane datasets, while the bottom row illustrates that Grad-CAM reversely projects the weighted heat map onto the raw images. Specially, the highlighted image components have a more pronounced impact on generating final decision. It is observed in Fig. 9 that our method predominantly focuses on the valuable high-frequency regions, such as lane lines, rather than background noises. This strongly proves that frequency can effectively guide the model to suppress the redundant information, forcing it to focus on discriminative signals such as lane edges.

E. Ablation Study

In this subsection, we analyze each component of our approach, i.e., Domain Transformation Module (DTM), Bilateral Aggregation Module (BAM) and Position Refinement Module (PRM), and discuss their advantages respectively. To verify the impact of each component, we conduct detailed ablation studies. Subsequently, we analyze the effects of different decoding formulations on the overall performance.

1) Analysis of Each Module: To investigate the importance of FAD and PAD, we sequentially incorporate them into the baseline M_1 one by one. We take ResNet-34 as the backbone of the model M_1. Especially, the FAD comprises a DTM and three BAMs while three PRMs are instantiated into PAD. The results of ablation studies are summarized in Table IV. We first detect lanes by baseline, which concentrates on spatial domain only. It is clear that the performance of our baseline M_1 is unsatisfactory, achieving 78.10% (IoU set to 50) and 49.51% (IoU set to 80) on CULane and 95.21% on TuSimple, respectively. This indicates that detecting lanes in the isolated spatial domain is not sufficient. Consequently, we introduce the DTM as an auxiliary path based on baseline M_1 to explicitly encode frequency-aware lane clues. Then, the frequency-aware clues and spatial features are fed into the BAM, which boosts the F_1 (IoU set as 80) from 49.51% to 50.35%. This demonstrates that using frequency to guide spatial-domain encoding process is capable of producing more comprehensive and discriminative lane representations. This is further evidenced by consistent performance gains on TuSimple, with an accuracy improvement of 1.21%. Notably, it only increases the model size by 0.74M, which is important to maintain high inference efficiency. The additional number of parameters introduced by our proposed method remains the same across both the CULane and TuSimple datasets.

In addition, fine-grained position modeling is of great importance for accurate lane detection. The DTM and BAM endows our model M_2 with a superior capacity to excite frequency-aware signals and global semantics from both intra- and inter-domain perspective. However, they fails to model fine-grained lane position information and thus we apply PRM on M_2 to address this limitation. The PRM is specially designed to refine lane locations iteratively and increases the F_1 (IoU set as 80) by a large margin to 51.45%. This is reasonable because PRM provides a more detailed contexts from local frequency regions and multi-level spatial features. Another noteworthy phenomenon we should notice is that the PAD achieves more performance gains on CULane than on TuSimple. A possible reason is that CULane comprises more complex sample space, mitigating the risk of overfitting. In general, each module contributes to the improved performance while incurring only a minor increase in the number of parameters.

2) Analysis of Fine-Grained Position Embedding: The proposed method takes the ResNet-50 as the backbone and captures cross-scale lane representations via a global-to-local fashion. It normally perceives global semantics in the deep spatial path to excite high-level features with large receptive field. In contrast, local contexts from frequency domain and shallow spatial path is further introduced to model fine-grained lane locations, since low-level signals excel in depicting position details. To this end, the BAM-PRM pair can be placed in different stages to effectively detect dynamic lanes with hierarchical domain receptive field.

We compared four model variations of our approach and summarized the results in Table V. For localization, we look at the recall value, which is the ratio of the number of correctly localized lanes to the total number of labels. For classification, we look at the precision of the correctly predicted lanes. From the version M_4, we can find that the BAM-PRM pair is positioned at the last stage and thus obtain a sub-optimal F_1 score, with a F_1@60 of 74.05. It is obvious that the global semantics from deepest stage help strengthen lane classification capacity but performs poor in terms of lane localization. A

TABLE V
ABLATION STUDY ON NUMBER OF STACKED "BAM-PAM" PAIRS

Model	Number	CULane		
		F_1@60	Precision	Recall
M_4	1	74.05	81.09	68.14
M_5	2	74.91	81.15	69.58
M_6	3	75.15	81.41	69.79
M_7	4	75.02	81.19	69.72

TABLE VI
ABLATION STUDY ON THE COLOR SPACE TRANSFORMATION OPERATION IN OUR PROPOSED DTM. HERE, ↑ AND ↓ DENOTE BETTER PERFORMANCE FOR HIGHER AND LOWER VALUES, RESPECTIVELY

Color Space	CULane		TuSimple		
	F_1@55↑	Precision↑	Acc↑	FP↓	FN↓
RGB	69.97	77.36	94.57	0.0429	0.0592
YCbCr	70.88	78.95	95.02	0.0419	0.0553

potential reason is that embeddings from deepest stage encounter seriously unrecoverable information loss. Benefiting from the rich details of shallow layer, M_5 significantly enhances its localization capacity. This is evidenced by a notable recall value improvement, from 68.14 to 69.58. Additionally, from the results in line 3, it is observed that M_6 further increases the performance by placing the BAM-PRM pairs at last three stages. This indicates that the proposed approach achieves the best localization performance with the help of local contexts from both high-band frequency and shallow spatial domain. In contrast, when placing the BAM-PRM pairs at the last four stages, the precision drops by a percentage of 0.13. This indicates that some redundant features (e.g., background noises from low-level layers) are introduced and lead to performance degradation.

3) Analysis of Color Space Transformation in DTM: Different color spaces have been investigated in order to find a better one for accurate lane detection. Notably, to underscore the effects of different color spaces, we perform this ablation experiment without pre-trained weights, data augmentation techniques, or the proposed PRM. The experimental results are summarized in Table VI. It can be observed that the performance metrics for both CULane and TuSimple datasets show improvements when using YCbCr color space compared to RGB. Specifically, for CULane, the F_1@55 score increased from 69.97% with RGB to 70.88% with YCbCr. Additionally, the precision improved from 77.36% to 78.95%, indicating a more reliable detection performance. Similar trends can also be observed in the TuSimple dataset.

These results suggest that the transformation to YCbCr color space provides a consistent enhancement in the performance of our method across different datasets. It is reasonable because the YCbCr color space is known to separate brightness from color information, which can be beneficial for reliably separating different-band frequencies.

TABLE VII
ABLATION STUDY ON COMPONENTS (INDICATED BY ✓) OF OUR METHOD. THE IOU THRESHOLD OF F_1 IS SET TO 50. FOR "CROSS" CATEGORY, ONLY FALSE POSITIVES (FP) ARE SHOWN. ↑ DENOTES BETTER PERFORMANCE WITH HIGHER VALUES, WHILE ↓ INDICATES BETTER PERFORMANCE WITH LOWER VALUES

Model	DTM+BAM	PRM	Night↑	Crowded↑	Cross↓	Total↑
M_8			74.24	77.18	1638	78.57
M_9	✓		74.94	77.29	1323	79.04
M_{10}	✓	✓	75.04	78.16	1282	79.71

4) Analysis of Each Module Under Challenging Scenarios: To investigate the effect of each module under challenging conditions of the CULane dataset such as nighttime and occlusions, we conduct additional ablation studies using ConvNext-small as the backbone of our models. The experimental results are summarized in Table VII. The M_8 serves as the baseline model, which achieves F_1 scores of 74.24 for nighttime, 77.18 for occlusions, and a Total F_1 score of 78.57. By incorporating the "DTM+BAM" component into M_8, our M_9 remarkably improves the F_1 score of nighttime to 74.94, occlusions to 77.29, and largely decreases the false positive (FP) of "Cross" from 1638 to 1323. These performance gains demonstrate that our DTM+BAM module (1) adaptively perceives both high- and low-frequency signals crucial for capturing local lane details in few-visual-clue scenarios (e.g., nighttime and occlusions); (2) leverages global semantics with a large receptive field to effectively suppress non-lane noises (e.g., road arrows) in the frequency domain; and (3) generates more discriminative lane representations by effectively encoding lane features from both spatial and frequency domains.

The addition of PRM to the M_9 (resulting in our M_{10}) yields a significant performance improvement. Specifically, our M_{10} outperforms the variant M_9 by 0.87% in the occlusions scenario, rising from 77.29 to 78.16, and by 0.67% in the Total F_1 metric, advancing from 79.04 to 79.71. Additionally, the FP metric of "Cross", which measures performance with lower values being better, shows a slight decrease from 1323 to 1282. These incremental gains highlight the positive impact of the PRM in modeling the fine-grained lane locations. Specifically, in the proposed PRM, we employ a position-aware gating operation to construct the relationship of different pixels within the integrated features, which adaptively excite pixel-wise location responses when detecting lane lines in different scenarios.

V. CONCLUSION

In this paper, we have proposed a novel network, which utilizes the complementarity of bilateral domains to achieve accurate and fast lane detection. By considering the frequency bias of CNN models, we have designed a shallow and small DTM to explicitly excite frequency-aware lane local signals. It excels in perceiving valuable high-frequency boundaries and preserving. To better distinguish lane lines from other non-target landmarks, the BAM is further presented to adaptively absorb global semantics of spatial domain into lane local

signals in frequency space. Furthermore, to model fine-grained lane locations, we have constructed three cascaded PRMs, which iteratively refine coarse lane positions via more detailed local contexts. The final features from both bilateral domains further upgrades the generalization capacity on real-world lane detection systems. Benefiting from taking advantages of powerful bilateral-domain clues and small parametric modules, the proposed method achieves consistent superior performance in challenging scenarios in terms of accuracy and efficiency.

REFERENCES

[1] M. Aly, "Real time detection of lane markers in urban streets," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2008, pp. 7–12.

[2] Z. Kim, "Robust lane detection and tracking in challenging scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 16–26, Mar. 2008.

[3] H. Jung, J. Min, and J. Kim, "An efficient lane detection algorithm for lane departure detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2013, pp. 976–981.

[4] Y. Wang, N. Dahnoun, and A. Achim, "A novel system for robust lane detection and tracking," *Signal Process.*, vol. 92, no. 2, pp. 319–334, Feb. 2012.

[5] G. Pang, B. Zhang, Z. Teng, N. Ma, and J. Fan, "Fast-HBNet: Hybrid branch network for fast lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15673–15683, Sep. 2022.

[6] 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.

[7] P. Lu, C. Cui, S. Xu, H. Peng, and F. Wang, "SUPER: A novel lane detection system," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 3, pp. 583–593, Sep. 2021.

[8] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, and C. C. Loy, "Inter-region affinity distillation for road marking segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12483–12492.

[9] S. Lee et al., "VPGNet: Vanishing point guided network for lane and road marking detection and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1947–1955.

[10] 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.*, 2018, pp. 7276–7283.

[11] 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.

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

[13] H. Honda and Y. Uchida, "CLRerNet: Improving confidence of lane detection with LaneIoU," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 1176–1185.

[14] 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.

[15] 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.

[16] K. Zhou and R. Zhou, "End-to-end lane detection with one-to-several transformer," 2023, *arXiv:2305.00675*.

[17] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De 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.

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

[19] Z. Qin, P. Zhang, and X. Li, "Ultra fast deep lane detection with hybrid anchor driven ordinal classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2555–2568, May 2024.

[20] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1740–1749.

[21] Z. Lin, Y. Gao, and J. Sang, "Investigating and explaining the frequency bias in image classification," in *Proc. Thirty-First Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 717–723, doi: 10.24963/ijcai.2022/101.

[22] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8684–8694.

[23] Q. Chang and Y. Tong, "A hybrid global-local perception network for lane detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 2, pp. 981–989.

[24] X. Xu, T. Yu, X. Hu, W. W. Y. Ng, and P.-A. Heng, "SALMNet: A structure-aware lane marking detection network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4986–4997, Aug. 2021.

[25] G. Liu, F. Wörgötter, and I. Markelic, "Combining statistical Hough transform and particle filter for robust lane detection and tracking," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 993–997.

[26] B. Yu and A. K. Jain, "Lane boundary detection using a multiresolution Hough transform," in *Proc. Int. Conf. Image Process.*, Oct. 1997, pp. 748–751.

[27] X. Chen, J. Zheng, C. Li, B. Wu, H. Wu, and J. Montewka, "Maritime traffic situation awareness analysis via high-fidelity ship imaging trajectory," *Multimedia Tools Appl.*, vol. 83, no. 16, pp. 48907–48923, Nov. 2023.

[28] Y. Li et al., "Airborne transmission of virus-laden droplets in an aircraft cabin," *Transp. Saf. Environ.*, vol. 5, no. 4, p. 079, Sep. 2023.

[29] X. Chen, D. Ma, and R. W. Liu, "Application of artificial intelligence in maritime transportation," *J. Mar. Sci. Eng.*, vol. 12, no. 3, p. 439, Mar. 2024.

[30] Y. Wen, Y. Yin, and H. Ran, "FlipNet: An attention-enhanced hierarchical feature flip fusion network for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 1–10, Aug. 2024.

[31] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1013–1021.

[32] X. Zhang et al., "Oblique convolution: A novel convolution idea for redefining lane detection," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 2, pp. 4025–4039, Feb. 2024.

[33] Q. Li et al., "PGA-net: Polynomial global attention network with mean curvature loss for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 417–429, Jan. 2024.

[34] L. Tabelini, R. Berriel, T. M. Paixao, 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.

[35] J. Philion, "FastDraw: Addressing the long tail of lane detection by adapting a sequential prediction network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11582–11591.

[36] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.

[37] L. Gueguen, A. E. Sergeev, B. Kadlec, R. Liu, and J. Yosinski, "Faster neural networks straight from JPEG," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 3933–3944.

[38] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4504–4513.

[39] M. Ehrlich and L. Davis, "Deep residual learning in the JPEG transform domain," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3484–3493.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[41] (2017). *Tusimple Benchmark*. [Online]. Available: http://benchmark.tusimple.ai/

[42] H. Abualsaud, S. Liu, D. B. Lu, K. Situ, A. Rangesh, and M. M. Trivedi, "LaneAF: Robust multi-lane detection with affinity fields," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7477–7484, Oct. 2021.

[43] R. Fan, X. Wang, Q. Hou, H. Liu, and T.-J. Mu, "SpinNet: Spinning convolutional network for lane boundary detection," *Comput. Vis. Media*, vol. 5, no. 4, pp. 417–428, Dec. 2019.

[44] H. Xu, S. Wang, X. Cai, W. Zhang, X. Liang, and Z. Li, "CurveLane-NAS: Unifying lane-sensitive architecture search and adaptive point blending," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 689–704.

[45] X. Yao, Y. Wang, Y. Wu, G. He, and S. Luo, "MLP-based efficient convolutional neural network for lane detection," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 12602–12614, Oct. 2023.

[46] A. Zoljodi, S. Abadijou, M. Alibeigi, and M. Daneshtalab, "Contrastive learning for lane detection via cross-similarity," 2023, *arXiv:2308.08242*.

[47] 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.
[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
[50] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[51] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Nan Ma (Senior Member, IEEE) received the Ph.D. degree in computer application from the University of Science Technology Beijing, Beijing, China, in 2013. She is currently a Full Professor with the Faculty of Information Technology, Beijing University of Technology, Beijing. Her current research interests include interactive cognition, intelligent driving, multimedia content analysis, and computer vision.

Guilin Pang received the B.S. degree in software engineering from Beijing Union University, Beijing, China, in 2020, and the M.S. degree in software engineering from Beijing Jiaotong University, Beijing, in 2023. He is currently a Research Assistant with the Department of Computer Science, Hong Kong Baptist University. His research interests include theory and applications of lane detection.

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Chair Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning and visual computing and their applications in data science, pattern recognition, multi-objective optimization, and information security. He is a member of European Academy of Sciences and Arts and a fellow of AAAS, IAPR, IET, and BCS. He serves as the Editor-in-Chief for IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He also serving/served as an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, *ACM Transactions on Intelligent Systems and Technology*, and *Pattern Recognition*. More details can be found at: https://www.comp.hkbu.edu.hk/ymc.