International Journal of Basic and Applied Sciences, 14 (8) (2025) 263-275

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/9x67sq37
Research paper

MPDDNet: A Multi-Scale Parallel Network with Deformable Non-Local and Direction-Aware Fusion for Lane Detection in Complex Environments

Shiling Huang [1, 2] *, Nur Ariffin Mohd Zin [2], Mohd Hamdi Irwan Hamzah [2]

[1] *Intelligent Manufacturing College, Nanning University, Nanning, 530200, China*
[2] *Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, 86400, Malaysia*
**Corresponding author E-mail: huangshiling@unn.edu.cn*

Received: October 18, 2025, Accepted: November 20, 2025, Published: December 15, 2025

Abstract

Lane detection in complex driving environments remains challenging due to issues such as poor visibility, occlusions, and intricate lane topologies. To address these challenges, this paper proposes MPDDNet, a novel multi-scale parallel network that integrates two key components: a Direction-aware Adaptive Multi-scale Feature Fusion (DAM-FF) module and a Deformable Non-local (DF-NL) module. The DAM-FF module explicitly embeds directional priors into multi-scale feature fusion through adaptive weighting and direction-aware spatial attention, significantly enhancing detailed feature representation in challenging scenarios. The DF-NL module combines multi-scale feature fusion with deformable attention mechanisms, enabling efficient global context modeling while implicitly incorporating structural priors of lane geometry. Through parallel integration, these modules achieve synergistic optimization of local details and global semantics. Extensive experiments on three benchmarks demonstrate that MPDDNet establishes new state-of-the-art performance, achieving 83.03% F1 score and 63.54% mF1 score on the CULane dataset. Our method also achieves remarkable results on the LLAMAS and TuSimple benchmarks, with 97.80% F1@50 and 98.40% F1 score, respectively. The consistent superiority across all three datasets and various challenging scenarios validates our approach's robustness and generalization capability, providing an effective solution for lane detection in complex environments.

Keywords: *Lane Detection; Multi-Scale Feature Fusion; Global Context Modeling; Direction-Aware Attention; Deformable Non-Local.*

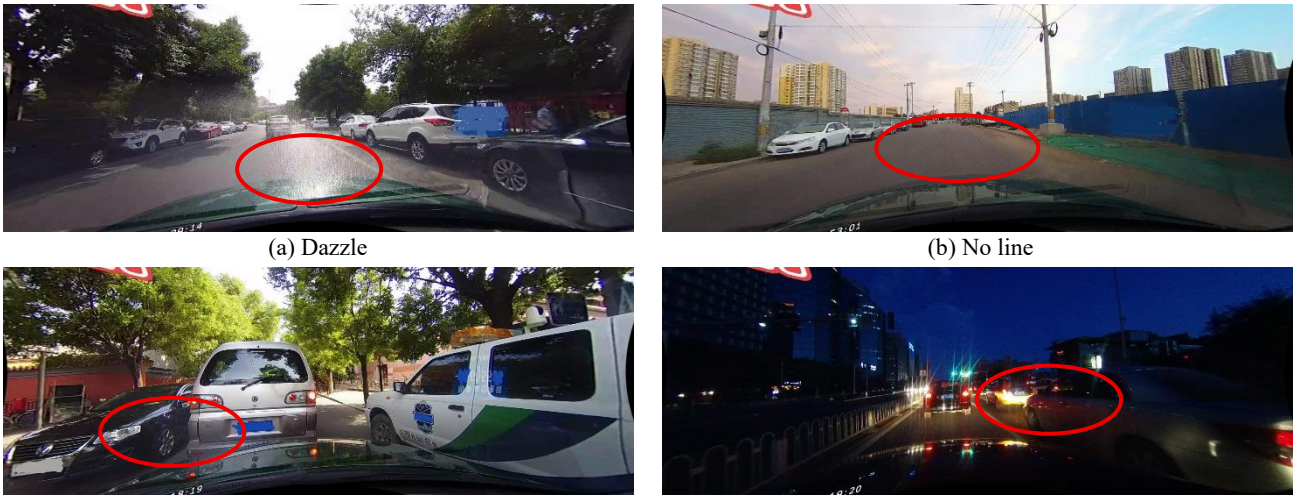
1. Introduction

The rapid development of autonomous driving technology puts forward extremely high requirements for environmental perception systems. As a key basis for autonomous vehicle positioning, path planning, and decision-making, the accuracy and robustness of Lane Detection are very important [1 - 3]. Despite the remarkable progress of deep learning-based detection methods in recent years, it is still a great challenge to achieve high-precision lane line perception in complex, dynamic environments in the real world [1].

A large number of studies [4 - 7] have shown that for lane detection in complex environments, the current mainstream models mainly follow two technical routes to improve the robustness of the model. First, it is committed to optimizing Multi-scale Feature Fusion. For example, the CLRNet [4] introduces a cross-layer recursive refinement mechanism to deeply fuse deep semantic features and shallow detail features, thus achieving excellent performance. Second, they focus on enhancing Global Context Modeling. For example, HGLNet [8] uses a Deformable Attention mechanism [9] to capture long-distance dependencies in images adaptively. To better understand the overall topology of complex scenes. These methods perform well on datasets with a single environment (e.g., TuSimple [10]), but generally face two key types of challenges when dealing with complex real-world scenarios:

1) Challenges of detail perception and multi-scale fusion: Lane lines in complex environments often become blurred or broken due to occlusion, dramatic illumination changes, and other factors (see Fig. 1 (a, b)). Although the mainstream methods [4], [8], [11], [12], [14], [15] generally adopt multi-scale feature fusion strategies (such as FPN[15] and its variants), the fusion process mostly relies on predefined or static weight mechanisms, which makes it difficult to adaptively enhance the subtle features of weakly visible regions and suppress background interference at the same time. Although previous studies have attempted to improve this problem by introducing recursive refinement, feature recalibration, or a lightweight pyramid structure [4], [16], most of these methods fail to explicitly incorporate the inherent directional and structural priors of lane lines, resulting in a bottleneck in the perception and recovery ability of slender lane structures in extreme scenarios.
2) Challenges of global semantic modeling and structured reasoning: In scenes such as intersections and severe occlusion (see Fig. 1 (c, d)), the complex topology of lane lines is easy to confuse with visually similar distractors, which requires a strong global context reasoning ability of the model. Existing methods [17 - 20] typically leverage mechanisms such as proposal-based context aggregation

[4] or deformable attention [8] to capture long-range dependencies. However, the former heavily relies on the accuracy of the initial proposal. At the same time, the latter lacks directional constraints on lane geometry (such as vertical continuity and horizontal parallelism) due to its isotropic search mechanism, which is prone to capturing irrelevant noise, and there is still room for improvement in inference efficiency and accuracy. Although some studies have explored the introduction of topological constraints or BEV transformation, they rely on additional data or complex calculations, and are difficult to be widely applied.



(a) Dazzle (b) No line

(c) Crowded (c)Night add Crowded

Fig. 1: Illustrations of Hard Cases for Lane Detection. (a) Dazzle—Where Bright Light Sources Cause Glare, Making It Difficult to Detect Lane Markings; (b) No Line—Where Lane Markings Are Absent Or Worn Out, Causing A Lack of Reference for Detection; (c) Crowded—Where Dense Traffic Or Surrounding Objects Obscure Lane Boundaries; and (d) Night and Crowded—A Combined Challenge Where Low-Light Conditions and Congestion Together Complicate the Detection Process.

To address the above challenges, this study aims to address two key issues: (1) How to construct an adaptive multi-scale feature fusion mechanism to accurately enhance lane features and suppress background noise in complex degraded environments? (2) How to design an efficient global context modeling framework to accurately reason about lane topology by explicitly embedding directional structural priors, thus significantly reducing false detections?

To this end, this paper proposes a novel Direction-aware Lane Attention Network (MPDDNet) to cope with the above challenges. The core innovation of this work is embodied in the design of two closely integrated modules:

1) Direction-aware Adaptive Multi-scale Feature Fusion Module (DAM-FF): By integrating the channel and spatial attention mechanisms and introducing the direction-aware convolution, the module realizes the adaptive weighted fusion of multi-scale features to effectively enhance the detailed feature representation of the lane structure in a complex environment, while suppressing the interference of background noise.
2) Deformable Non-Local Module (DF-NL): This module captures multi-scale context using convolutions with different dilation rates

By embedding DAM-FF and DF-NL into a unified network architecture, MPDDNet realizes the collaborative enhancement from local detail awareness to global semantic reasoning. Experiments on several public benchmarks show that the comprehensive performance of the proposed method is significantly better than the existing advanced models, especially in challenging scenes (e.g., occlusion, glare, intersection, night, curve, etc.), showing better robustness and accuracy.

The main contributions of this paper can be summarized as follows:

1) We propose a novel Multi-scale Parallel Network with Deformable Non-local and Direction-aware Fusion (MPDDNet). The network features a parallel architecture comprising two dedicated modules: the Direction-aware Adaptive Multi-scale Feature Fusion (DAM-FF) module for capturing fine-grained local details and structural cues, and the Deformable Non-local (DF-NL) module for modeling long-range global dependencies. This design provides a unified and efficient solution for lane detection in complex environments.
2) We design a direction-aware adaptive Multi-scale feature fusion module (DAM-FF), which adaptively enhances the detailed feature representation of lane lines through dynamic weight allocation and direction-aware convolution, and effectively improves the localization accuracy of the model in challenging scenes such as occlusion and illumination change.
3) We develop a deformable nonlocal module (DF-NL) that integrates multi-scale feature fusion with deformable attention mechanisms. By employing dilated convolutions with different receptive fields to predict sampling offsets, the module implicitly incorporates directional priors of lane structures while maintaining computational efficiency through sparse sampling. This design enables effective global context modeling specifically tailored for lane topology reasoning in complex environments.

The rest of this paper is organized as follows: Section II reviews the related research work; The third part introduces the overall framework and core modules of MPDDNet in detail. Section 4 gives the experimental results and analysis. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Representative methods for lane detection

According to the representation form of lane lines adopted by deep learning models, existing methods can be roughly divided into three categories: segmentation-based methods, anchor-based methods, and parametric curve-based methods. Various methods have their unique advantages and limitations, which jointly promote the development of lane detection technology.

Segment-based methods treat lane detection as a pixel-level classification task, aiming to predict for each pixel whether it belongs to a lane line or not. SCNN [19] is the pioneering work in this field, which captures the long-distance spatial relationship of lane lines through the information transfer mechanism between slices, and significantly improves the performance. However, its high computational overhead makes it difficult to meet the requirements of real-time applications. On this basis, RESA [21] proposed a circular feature shift aggregator,

which realized the extraction and integration of global features with higher efficiency. Although these methods can achieve pixel-level high-accuracy localization, their inherent disadvantages, such as high computational complexity and cumbersome post-processing steps (such as clustering pixels into complete lane instances), limit their application. More importantly, the lane line is regarded as a set of discrete pixels rather than an entity with an overall structure, which makes it difficult to ensure the coherence of the lane line in scenarios such as occlusion.

Anchor-based methods regress the exact lane line position through preset anchors, which can be mainly divided into two categories: row anchors and line anchors. Row anchor methods (such as UFLD [22]) define the detection problem as selecting cells that may contain lane points on a predetermined row, and their structure is simple and fast. However, the strong vertical prior constraint limits their flexibility, and the row-level prediction is difficult to effectively capture the long-distance global context. CondLaneNet [23] introduces conditional convolution, which first locates the lane starting point and then detects it based on line anchors, which improves the processing ability of complex topological structures, but its performance is affected in scenarios where the starting point is difficult to accurately identify. Line anchor methods, such as LaneATT [18], use predefined line segments as anchors to predict the lane lines by regression offset. CLRNet [4] is the current representative model with leading performance, which deeply fuses multi-scale features by a cross-layer recursive refinement mechanism and aggregates global context information by ROIGather operation to achieve excellent detection accuracy. However, its multi-scale fusion strategy is relatively naive, and the efficiency of the ROIGather operation highly depends on the quality of the initial proposal. If the proposal is inaccurate, the subsequent refinement process is difficult to recover from the error.

The method based on a parametric curve uses parametric models (such as polynomials and Bezier curves) to directly represent the overall shape of the lane line, and completes the detection by regression model parameters. PolyLaneNet [24] uses polynomial regression and achieves efficient inference speed. LSTR [25] combines the Transformer architecture and uses the self-attention mechanism [26] to capture global features for regression parameters. The advantage of this class of methods is that the output representation is compact, the inference speed is fast, and the continuity of the lane is naturally guaranteed. However, the performance of the proposed method is very sensitive to parameter prediction errors, especially the high-order coefficients. A small deviation may lead to significant distortion of the lane shape, so that the stability and accuracy of the proposed method are still lower than those of the previous two methods under complex topology structures (such as road intersections).

2.2. Multi-scale feature fusion in lane detection

Multi-scale feature fusion is a key technology to improve the performance of dense prediction tasks, such as object detection and semantic segmentation [15], [27]. The core of multi-scale feature fusion is to effectively integrate feature maps from different depths of the network: shallow features contain rich detail and location information, while deep features carry high-level semantic information. In the field of general computer vision, Feature Pyramid Network (FPN) [15] and its subsequent improved schemes (such as PANet [27]) have become the classical paradigm for multi-scale feature fusion. FPN builds a multi-scale feature pyramid with strong semantic information by laterally connecting deep semantic features with shallow high-resolution features through a top-down path. PANet, on the other hand, adds bottom-up path enhancement to FPN to improve the flow of bottom features to top features. However, the limitations of these generic paradigms emerge when they are directly applied to the lane line detection task. Lane lines have unique visual properties: they are usually slender, continuous, and strongly directional structures. General-purpose multi-scale fusion strategies (e.g., simple element-by-element addition or lane concatenation) are inherently task-agnostic, failing to embed a targeted design for structural priors such as lane lines. For example, the lane line features become extremely weak in scenes with partial occlusion, shadows, or resolution degradation. The general fusion mechanism is difficult to enhance these weak but critical lane features adaptively, while effectively suppressing the texture interference in the background (such as tire indentation, road repair marks). Although some advanced lane detection models (such as CLRNet [4]) deepen feature utilization by introducing a cross-layer recursive refinement mechanism, the basis of their multi-scale fusion is still relatively naive, and they fail to explicitly guide the network to pay attention to the structural characteristics of lane lines, thus limiting the detail recovery ability in extreme scenarios. Therefore, the existing multi-scale fusion methods often fail to achieve pixel-level accurate localization while maintaining high semantic understanding in the face of complex environments. This fully demonstrates that designing an adaptive fusion mechanism that can sense the direction and structure of the lane lines is crucial to improving the robustness of the model in challenging scenarios. The direction-aware Adaptive Multi-scale Feature Fusion module (DAM-FF) proposed in this paper is designed to address this challenge.

2.3. Global context modeling in lane detection

In complex driving scenarios such as intersections and heavy occlusion, lane detection faces two fundamental challenges: accurately reasoning about lane topology and recovering occluded visual cues. Traditional Convolutional Neural Networks (CNNs), constrained by their inherent local receptive fields, struggle to establish direct long-range dependencies between pixels, which is crucial for understanding slender structures like lane lines [19], [7], [28].

2.3.1. Relationship between global context and long-range dependencies

In lane detection, global context modeling and the establishment of long-range dependencies are fundamentally complementary concepts: the former represents the objective, while the latter serves as the key mechanism to achieve this goal [29]. Although traditional Convolutional Neural Networks (CNNs) can progressively expand their receptive fields through hierarchical stacking, enabling higher-level features to contain rich semantic information, this expansion of receptive fields remains local and gradual [30], [31]. This inherent limitation makes it difficult to establish the direct long-range associations required for lane structures. For instance, when partial occlusion occurs in lane markings, relying solely on local features proves inadequate for inferring the trajectory of obscured sections, necessitating the establishment of direct connections with distant visible lane segments. The geometric characteristics of lane lines dictate that their detection must consider global spatial relationships. Specifically, as spatially extended structures, the coherence between local segments of lane lines requires long-range connections to ensure structural continuity. Furthermore, in complex intersection scenarios, the convergence and divergence relationships among multiple lane lines can only be accurately resolved from a global perspective. Additionally, under partial occlusion conditions, recovering complete lane markings necessitates full utilization of contextual information from unobstructed regions.

2.3.2. Technical development pathways

Architectural innovations based on self-attention [26] mechanisms have introduced new paradigms for global context modeling. The self-attention mechanism achieves long-range dependency modeling by computing global correlations among sequence elements. Wang et al. [32] introduced this concept to the visual domain through the Non-local module, which directly computes pairwise relationships between spatial positions. Dosovitskiy et al. [33] further demonstrated the potential of pure Transformer architectures in visual tasks through Vision Transformer. To enhance efficiency, Zhu et al. [9] proposed Deformable DETR, which employs deformable attention to achieve sparse sampling.

In the field of lane detection, these technical advancements have primarily manifested through two pathways:

1) End-to-End Transformer Architectures (e.g., LSTR [25], O2SFormer [34]) formulate lane detection as a sequence prediction problem, leveraging the Transformer's global modeling capability to directly output lane parameters. While these methods exhibit powerful global reasoning capabilities, they demand substantial data resources and show limitations in detail recovery.
2) Hybrid Enhancement Architectures incorporate attention modules while maintaining CNN backbones. SCNN [19] achieves information propagation along row and column directions through slice-wise convolution; CLRNet [4] employs ROI attention to refine lane proposals; LaneFormer [9] designs row-column dual-axis attention; while CondFormer [35] and HGLNet [8] explore the integration of conditional convolutions with attention mechanisms.

In recent years, emerging trends that fuse parametric representations with hybrid architectures have further pushed the performance boundary. For instance, LaneFormer [9] achieves more efficient long-range dependency modeling within a hybrid architecture through its novel row-column dual-axis attention mechanism. These advancements collectively illustrate a converging trend in lane detection technology towards greater efficiency and accuracy.

2.3.3. Current challenges and improvement directions

Current approaches in lane detection continue to face two fundamental limitations that hinder their performance in complex scenarios. First, the post-hoc correction dependency prevalent in methods like CLRNet restricts global modeling to the detection head stage, where effectiveness becomes heavily reliant on the output quality of preceding network layers, making it particularly challenging to rectify initial errors propagated through the network. Second, existing methods generally lack explicit modeling of inherent structural priors such as the strong directionality and continuity characteristics of lane lines, consequently requiring extensive training data to implicitly learn these geometrically constrained patterns. To address these critical issues, this paper proposes the Deformable Non-local Module (DF-NL), which strategically performs global interaction on mid-level backbone features rather than postponing it to later stages. By guiding deformable offset learning through direction-aware constraints and integrating multi-scale feature fusion, the DF-NL module enables more targeted global context modeling that explicitly incorporates lane structural priors. This design not only facilitates early error correction in the feature extraction process but also significantly enhances the reliability of lane topology reasoning in challenging environments such as severe occlusions and complex intersections, thereby providing a more robust foundation for lane structure understanding.

3. Method

To address the challenges of lane detection in complex environments, such as loss of details, insufficient global context modeling, and scale change, this paper proposes MPDDNet, a multi-scale parallel network that integrates the Direction-aware Adaptive Multi-scale Feature Fusion module (DAM-FF) and the Deformable Non-local module (DF-NL), and the structure diagram is shown in Figure 2. MPDDNet adopts a segmentation-based lane line representation method, optimizes local details and global semantics through a parallel dual-path architecture, and uses a dynamic fusion strategy to realize scene-adaptive feature integration, aiming to improve the robustness and accuracy of the model in various challenging scenarios.

3.1. Overall architecture

Many existing lane detection models often struggle to strike the best balance between preserving the fine details of the lane and understanding the global topology when dealing with complex scenes. To address the challenge of balancing fine-grained lane details with global topological understanding in complex scenes, we propose MPDDNet with a novel parallel dual-path architecture, as illustrated in Fig. 2. The network employs ResNet [36] as the backbone to extract multi-scale feature maps {C2, C3, C4, C5}. These features are then fed into two core modules for parallel processing:

1) DAM-FF pathway: focusing on enhancing local details and directional structures. This module applies a carefully designed attention mechanism to the features of each scale independently, and then performs upsampling fusion to preserve the fine geometric information of the lane to the maximum extent.
2) DF-NL pathway: focuses on capturing long-range dependencies and global context. The module receives features at all scales, uses a deformable attention mechanism to efficiently model the global relationship between pixels, and is particularly good at reasoning about the topology of occluded lanes.

Finally, the Adaptive Dynamic Feature Fusion (AD-FF) module adaptively calculated the weights according to the specific content of the input image, and weighted fused the "detailed features" and "global features" output by the two pathways. The fused features are finally fed into a lightweight decoder to generate pixel-wise lane line segmentation maps. This parallel design clearly divides the labor, avoids the possible information loss caused by the serial structure, and enables the network to flexibly respond to the needs of different scenarios.

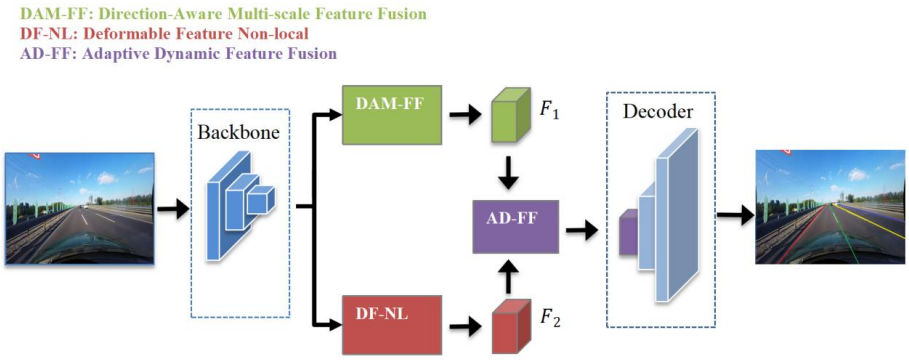


Fig. 2: Overall Architecture Diagram of MPDDNet.

3.2. Direction-aware adaptive multi-scale feature fusion module (DAM-FF)

3.2.1. Motivation and problem analysis

The traditional Feature Pyramid Network (FPN) and its variants have obvious limitations in lane detection. FPN and its related variants employ a top-down approach for feature fusion, integrating different-scale features through simple element-by-element addition or concatenation. However, this general fusion mechanism fails to fully consider the unique geometric properties of lane lines, which are slender, continuous, and strongly directional. In challenging scenes with partial occlusion, illumination changes, or resolution degradation, weak lane features are easy to be submerged by background noise in the fusion process, resulting in loss of detailed information. The DAM-FF module is proposed to overcome the above "invisible" problem. The core idea is to explicitly embed the directional prior into the multi-scale fusion process, so that the network can adaptively enhance the lane line structure response while suppressing irrelevant background interference.

3.2.2. Module structure design

The overall architecture of DAM-FF is shown in Fig. 3 and contains three key components: Efficient Channel Attention (ECA), Direction Aware Spatial Attention (DASA), and an Adaptive Multi-Scale Weight Fusion (AMWF) mechanism. The DAM-FF first independently refines the multi-scale features (C2-C5) with channel and direction-aware spatial attention. The refined features are then up-sampled to a common scale and fused by the Adaptive Multi-scale Weighting Fusion (AMWF) mechanism.

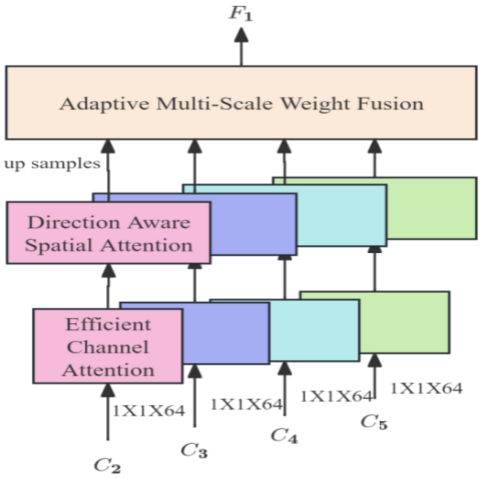


Fig. 3: Illustration of the DAM-FF Module.

The core design principle of the DAM-FF module follows an "optimize-before-fuse" strategy, where each multi-scale feature map $F_i \in R^{C\times H_i\times W_i}$(with $i \in \{2,3,4,5\}$) extracted from the backbone network undergoes sequential channel and spatial augmentation before fusion.

a) Efficient Channel Attention (ECA)

Firstly, we designed an ECA module with reference to ECA-Net [37]. The channel dimension calibration is performed on the features of each scale independently. The specific method is to perform a global average pooling operation on each channel in space, and the Sigmoid activation function is used to generate the channel weight, which is multiplied by F_i, and the important feature channels at each scale are enhanced through cross-channel interaction. Channel feature enhancement of the ECA module is computed as in (1).

$$F_i^{eca} = F_i \otimes \text{Sigmoid}\left(\text{Conv1D}_k\left(\text{GAP}(F_i)\right)\right) \quad (1)$$

Where GAP represents the global average pooling operation, Conv1D_k is the one-dimensional convolution with kernel size k, and $\otimes$ represents the broadcast multiplication of channel direction.

b) Direction Aware Spatial Attention (DASA)

The direction aware spatial attention mechanism serves as a core component of the DAM-FF module, with its design motivated by the strong directional structural characteristics inherent to lane lines in real-world scenarios. This mechanism effectively captures spatial

context information of lane markings through multi-scale convolution operations, significantly enhancing the model's perception capability for slender lane structures.

For the channel-calibrated features F_i^{eca}, vertical and horizontal convolutions are employed to capture directional priors of lane lines, formulated as shown in Equation (2):

$$F_i^{att} = F_i^{eca} \otimes \left(\alpha_i \cdot \sigma\left(Conv_{1\times K}(F_i^{eca})\right) + \beta_i \cdot \sigma\left(Conv_{K\times 1}(F_i^{eca})\right)\right) \quad (2)$$

Where $Conv_{1\times K}$ and $Conv_{K\times 1}$ represent one-dimensional convolutions in vertical and horizontal directions, $K \in \{3,7,7\}$, The parameters α_i and β_i are learnable balance coefficients, and F_i^{att} denotes the output feature after direction-aware enhancement, where σ denotes the sigmoid function.

c) Adaptive Multi-scale Weight Fusion (AMWF)

Traditional multi-scale feature fusion methods usually use simple feature concatenation or element-wise addition operations, which implicitly assume that all scale features contribute equally to the final task. However, in the practical application of lane detection, the information carried by feature maps at different scales exhibits significant differences: shallow features contain rich spatial details but are more prone to noise. In contrast, deep features possess strong semantic information but lack spatial resolution. Simple fusion strategies cannot dynamically adjust the relative importance of each scale feature according to the specific content of the input image. Inspired by the success of attention mechanisms in feature selection, we propose the Adaptive Multi-scale Weight Fusion (AMWF) module. The core innovation of this module is to introduce a lightweight weight generation network, which can automatically learn the optimal fusion weight of each scale feature according to the current input features, to replace the traditional hand-designed fusion strategy.

Given a direction-aware enhanced multi-scale feature F_i^{att} firstly, the AMWF module unifies all scale features to the maximum scale by bilinear interpolation upsampling operation, and the upsampled feature map is denoted as $\tilde{F}_i$. Then, a lightweight weight generation network Φ is designed to automatically calculate the weight W of the importance of each scale based on the input features, as defined in Equation (3):

$$W = \mathrm{Softmax}\left(\Phi\left(\tilde{F}_i\right)\right) \quad (3)$$

Subsequently, the fused feature F_1 is obtained by performing a weighted aggregation of all upsampled features according to the learned weights, formulated in Equation (4):

$$F_1 = \sum_{i=2}^{5} w_i \cdot \tilde{F}_i \quad (4)$$

Where $w_i \in W$ denotes the weight for the i-th scale feature.

3.3. Multi-scale deformable nonlocal module (DF-NL)

3.3.1. Design motivation and problem analysis

Lane detection in complex driving environments demands effective modeling of long-range dependencies to accurately infer lane topology under challenging conditions such as severe occlusion and complex intersections. While attention-based methods have demonstrated remarkable capability in capturing global contextual relationships between pixels, their direct application to lane detection faces significant limitations [30], [38]. The fundamental challenge lies in the computational complexity of standard self-attention mechanisms. As these operations compute pairwise relationships across all spatial positions, their complexity scales quadratically with feature map size, reaching $O((HW)^2)$, which makes it difficult to be used in practical applications when dealing with high-resolution feature maps.

Our approach draws inspiration from the deformable attention paradigm, which addresses the computational bottleneck through learnable sampling offsets that enable sparse, content-aware feature aggregation. However, while deformable mechanisms significantly improve computational efficiency, they exhibit a critical limitation in lane detection applications: the lack of explicit directional priors. Lane structures possess inherent geometric properties characterized by strong directional continuity, yet standard deformable attention requires extensive training data to implicitly learn these structural patterns.

3.3.2. DF-NL module structure

To address these limitations, we propose the Multi-scale Deformable Non-local Module (DF-NL), as illustrated in Fig. 4, which fundamentally improves upon standard Non-local modules by replacing their computationally expensive dense global interactions with an efficient deformable sparse sampling mechanism. This architectural innovation effectively overcomes the quadratic complexity bottleneck of conventional global modeling while preserving expressive power. Specifically, DF-NL introduces two key enhancements: multi-scale feature fusion with adaptive dilation strategies and direction-aware deformable attention. The module first integrates features across multiple scales and then employs a multi-scale dilation strategy for offset prediction that implicitly incorporates directional priors inherent in lane geometry. This synergistic combination enables precisely tailored global context modeling for lane structures, capturing long-range dependencies while maintaining sensitivity to directional characteristics.

The DF-NL module receives the multi-scale feature map $F_i \in R^{C\times H_i\times W_i}$ (where $i \in \{2,3,4,5\}$) from the backbone network. Firstly, the size of the F_i feature map is uniformly adjusted to be the same as the F_5 feature map through the bilinear interpolation downsampling operation to ensure that all feature maps are consistent in the spatial dimension. Subsequently, all the adjusted feature maps are concatenated in the channel dimension to form a fused feature containing multi-scale information. Finally, a 1×1 convolutional layer, batch normalization layer, and ReLU activation function were used to fuse and reduce the dimension of the concatenated features, generating a unified feature representation F_f for subsequent processing in the DF-NL module.

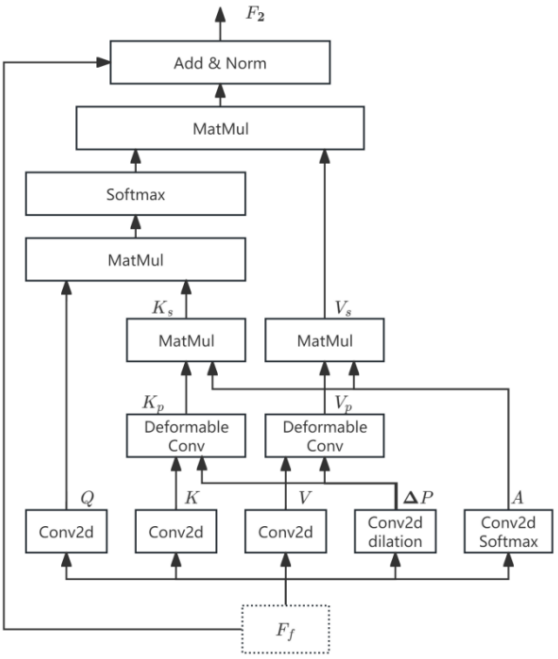


Fig. 4: Illustration of the DF-NL Module.

The core innovation of DF-NL lies in its deformable attention mechanism with implicit directional awareness. Given the input feature F_f, the module first projects it into three distinct representations through linear transformations: query (Q), key (K), and value (V), following the standard attention formulation [26]. In this mechanism, the Q represents the current focus that seeks relevant information, the K serves as an identifier for the stored information, and the V contains the actual content to be aggregated. These projections are formulated as (5), (6), and (7):

$$Q = W_q \cdot F_f \tag{5}$$

$$K = W_k \cdot F_f \tag{6}$$

$$V = W_v \cdot F_f \tag{7}$$

Where W_q, W_k, W_v are learnable projection matrices.

Unlike standard deformable attention mechanisms [9] that typically employ single-scale convolutions for offset prediction, DF-NL introduces a multi-scale dilation strategy to enhance its perception of lane structures. This approach enables the module to capture both local details and longer-range spatial relationships critical for lane geometry understanding.

Formula (8) defines the computation of the multi-scale offset field ΔP as follows:

$$\Delta P = \mathrm{GroupNorm}\left(\mathrm{Conv2d}_{1\times1\times N_p}\left(\mathrm{Concat}\left(\mathrm{Conv2d}_{3\times3,\mathrm{dilation}=1}\,(F_f), \mathrm{Conv2d}_{3\times3,\mathrm{dilation}=2}\,(F_f)\right)\right)\right) \tag{8}$$

Where $\Delta P \in R^{H\times W\times N_p\times 2}$ represents dynamic sampling location offsets in the spatial dimensions of the input feature map F_f, where H and W correspond to the spatial height and width of the feature map, respectively. For each spatial position q in the feature map F_f (i.e., each coordinate point (i,j) where i=1,..., H and j=1,..., W), ΔP provides N_p two-dimensional coordinate offsets (Δx, Δy). Each offset vector represents a displacement from the original position q, pointing to context sampling points that exhibit the strongest semantic relevance to q. This design enables the model to adaptively generate N_p sampling points for each feature map position, breaking through the limitations of traditional fixed-grid sampling and thereby more effectively capturing feature information with long-range dependencies and complex geometric structures such as lane lines. $\mathrm{Conv2d}_{k\times k,\mathrm{dilation}=d}$ represents a 2D convolution operation with kernel size k × k and dilation rate d, which controls the receptive field; Concat indicates the channel-wise concatenation of feature maps; $\mathrm{Conv2d}_{1\times1}$is a 1×1 convolution; and GroupNorm refers to Group Normalization, which stabilizes activation distributions during training. The process begins by generating two distinct offset components. First, via a 3×3 convolution with dilation rate 1, preserving fine-grained local patterns and immediate spatial relationships, and through a 3×3 convolution with dilation rate 2, which captures broader contextual information and longer-range dependencies along potential lane directions. These complementary offset fields are then concatenated and processed by $\mathrm{Conv2d}_{1\times1\times N_p}$ that learns to fuse the multi-scale spatial information, followed by Group Normalization to stabilize the training process. The resulting unified offset field ΔP inherently incorporates directional priors of lane structures through this multi-scale design, enabling the deformable attention to sample features in a geometry-aware manner without requiring explicit directional supervision.

The strategic use of dilation rates 1 and 2 serves a specific purpose:

1) Dilation=1 captures local contextual information with a standard receptive field.
2) Dilation=2 expands the receptive field to capture longer-range spatial relationships along potential lane directions.

By fusing the offset fields generated from these complementary receptive fields, the network learns to predict sampling locations that are not merely content-aware but also geometry-aware. This allows the DF-NL module to inherently favor sampling along the dominant directional axes (vertical or horizontal) of lane structures without relying on explicit, hand-crafted supervision, thereby providing a powerful structural prior that is learned directly from data.

Simultaneously, the module predicts attention weights for each sampling point to determine their relative importance in feature aggregation. The attention weight matrix A is generated through a dedicated convolutional layer followed by normalization, as defined in Equation (9):

$$A = \text{Softmax}\left(\text{Conv2d}_{3\times3\times N_p}(F_f)\right) \tag{9}$$

Where $A \in R^{H\times W\times N_p}$ represents the attention weight tensor, with N_p denotes both the number of output channels from the convolutional layer and the number of sampling points in the deformable attention mechanism. The 3×3 convolutional layer first processes the input feature F_f to produce initial attention scores, which are then normalized across all sampling points using the Softmax function. This ensures that the attention weights form a valid probability distribution, enabling the model to adaptively focus on the most relevant spatial locations during feature aggregation. The predicted weights play a crucial role in the subsequent deformable sampling process, allowing the module to emphasize features from geometrically meaningful regions while suppressing irrelevant background information.

Based on the predicted offsets ΔP, the DF-NL module performs deformable sampling on both the K and V features to aggregate spatially adaptive features. This process is formally expressed in Equations (10) to (11), which define the aggregation of the sampled key features K_s and value features V_s, respectively.

$$K_s = \sum_{k=1}^{N_p} A^{(k)} \odot \text{BilinearSample}\left(K, \Delta P^{(k)}\right) \tag{10}$$

$$V_s = \sum_{k=1}^{N_p} A^{(k)} \odot \text{BilinearSample}\left(V, \Delta P^{(k)}\right) \tag{11}$$

Where BilinearSample denotes the bilinear sampling operation, N_p is the number of sampling points, $\Delta P^{(k)}$ is the offset of the k-th sampling point, $A^{(k)}$ is the attention weight of the k-th sampling point, $\odot$ represents element-wise multiplication.

The computation in Equations (10) and (11) aggregates the sampled key (K_s) and value (V_s) features through three steps for each of the Np sampling points:

1) Bilinear Sampling: The BilinearSample operation uses the offset $\Delta P^{(k)}$ to fetch a feature vector from the input K (or V) at a precise, potentially sub-pixel location. It does this by computing a weighted average of the four nearest pixels, ensuring the process is differentiable.
2) Weighting: The sampled feature is then element-wise multiplied ($\odot$) by its corresponding attention weight $A^{(k)}$, which scales the feature based on its importance.
3) Summation: The weighted features from all N_p points are summed together.

This process results in K_s and V_s, which are dynamically aggregated feature maps. They concentrate information from the most relevant spatial contexts as determined by the learned offsets and attention weights, enabling the model to focus on geometrically meaningful regions like lane lines.

The final output of the DF-NL module integrates the attention-weighted features through a residual connection to preserve original feature information while enhancing it with globally contextualized representations. First, the global context representation Y is computed using the scaled dot-product self-attention mechanism introduced in [26], which enables the model to focus on the most relevant spatial locations by measuring compatibility between queries and keys. This process is formally defined in Equation (12):

Based on the self-attention mechanism [26], the output representation Y is computed by measuring the compatibility between queries Q and K_s, then using the resulting attention weights to aggregate value features V_s. This process is formally defined in Equation (12):

$$Y = \text{Softmax}\left(\frac{QK_s^{\,T}}{\sqrt{d_k}}V_s\right) \tag{12}$$

Where Q denotes the query matrix projected from the input feature, K_s and V_s represent the aggregated key and value features from Equations (10) and (11) respectively, and d_k is the dimensionality of the key features. The scaling factor $\frac{1}{\sqrt{d_k}}$ stabilizes the gradient during training by preventing the dot products from growing excessively large. The Softmax function normalizes the attention scores to form a probability distribution, ensuring that the output Y effectively summarizes the global contextual information from the value features based on the attention-weighted spatial relationships.

The final output of the DF-NL module integrates the attention-weighted features through a residual connection to preserve original feature information while enhancing it with globally contextualized representations. The output feature F_2 is generated by combining the attention output Y with the original input feature F_f through a gated residual connection, as formulated in Equation (13):

The DF-NL module produces the output feature F_2 through a gated residual connection [36] that adaptively fuses the global context Y with the original input feature F_f, as defined in Equation (13):

$$F_2 = \gamma Y + F_f \tag{13}$$

Where γ is a learnable scalar parameter that allows the network to automatically learn the optimal blending ratio between the global contextual information generated by the DF-NL module and the original features. The residual connection ensures that critical original feature information is preserved while augmenting it with structurally reasoned representations, thereby enhancing the robustness of lane topology modeling in diverse environments.

3.4. Adaptive dynamic fusion (AD-FF) and decoder

The Adaptive Dynamic Feature Fusion (AD-FF) module serves as the critical integration point between the two parallel pathways. As shown in Fig. 2, this module takes the enhanced features from both DAM-FF and DF-NL branches and performs content-aware fusion through a scene-adaptive weighting mechanism. The fusion process begins with channel-wise concatenation of both feature sets, followed by global average pooling to capture scene characteristics. A lightweight convolutional network then generates spatial attention weights that dynamically balance the contribution of detailed local features from DAM-FF against global contextual features from DF-NL based

on the specific input content. This adaptive weighting enables the network to emphasize local details in scenarios with clear lane markings while prioritizing global context in challenging conditions such as occlusions or poor visibility.
The formal implementation of the fusion process can be described as in (14) and (15):

$$W = \text{Softmax}(\text{MLP}(\text{GAP}(\text{Concat}(F_1, F_2)))) \tag{14}$$

$$F_{\text{fused}} = W[0] \odot F_1 + W[1] \odot F_2 \tag{15}$$

Where F_1 denotes the detail-enhanced features from the DAM-FF module, F_2 represents the globally-aware features from the DF-NL module, W is the adaptive weight tensor generated by Softmax normalization with dimensions $R^{H\times W\times 2}$, and $W[0]$ and $W[1]$ are the respective weight components for F_1 and F_2 derived from slicing the first and second channels of W, while F_{fused} is the resulting fused feature. The Adaptive Dynamic Fusion (AD-FF) process begins by concatenating the local detail features F_1 from DAM-FF and the global context features F_2 from DF-NL along the channel dimension to form a comprehensive feature representation. This combined representation undergoes global average pooling (GAP) to extract channel-wise statistics, which are then processed by a multi-layer perceptron (MLP) to produce initial weighting coefficients. These coefficients are normalized through the Softmax function to generate the spatial-aware weight tensor W, ensuring a valid probability distribution across all spatial locations. The final fused feature F_{fused} is obtained through element-wise multiplication of $W[0]$ with F_1 and $W[1]$ with F_2, followed by their summation, enabling the network to dynamically balance fine-grained details from DAM-FF and global contextual information from DF-NL according to the specific requirements of each spatial location in the input scene.
Following feature fusion, a lightweight decoder transforms the integrated features into the final lane segmentation map. The decoder employs a simple yet effective architecture consisting of two sequential bilinear upsampling operations with 2× scaling factors, restoring the spatial resolution to match the original input dimensions. A final 1×1 convolution layer projects the high-dimensional features to the target number of lane classes while preserving the learned spatial relationships. This streamlined design ensures computational efficiency while maintaining the integrity of the fused feature representations throughout the decoding process.

4. Experiments

4.1. Dataset and evaluation metrics

We evaluate our proposed MPDDNet on three widely used lane detection benchmarks: CULane [19], TuSimple [10], and LLAMAS [39]. CULane is a large-scale lane detection dataset containing 88,880 training images and 34,680 testing images captured in diverse driving scenarios. Following the standard practice established in prior work [19], we adopt the same data split, where the original training set is divided into 79,205 images (approximately 89.1%) for training and 9,675 images (approximately 10.9%) for validation, while the official test set of 34,680 images (28.1% of the total dataset) is used for evaluation. This standardized split ensures a fair comparison with existing methods. The test set is categorized into nine challenging scenarios, including crowded, night, no line, shadow, arrow, curve, crossroad, and dazzle light, which comprehensively evaluate model robustness under various conditions. All images have a resolution of 1640 × 590 pixels.
LLAMAS provides over 100,000 highway images with precise lane annotations. The dataset follows a standardized split with approximately 58,269 images for training (58.3%), 20,844 for validation (20.8%), and 20,929 for testing (20.9%). Since its test labels are not publicly available, evaluation requires submitting predictions to the official server, while the validation set enables local performance assessment during development.
TuSimple is a highway dataset comprising 3,626 training images (56.6%), 358 validation images (5.6%), and 2,782 testing images (43.4%), totaling 6,406 annotated frames with 720 × 1280 resolution. The dataset maintains this predefined split to ensure fair comparison across different methods, focusing on highway lane detection under relatively stable lighting conditions.
Evaluation Metrics for CULane and LLAMAS, we employ the F1-measure as the primary evaluation metric, which is based on Intersection-over-union (IoU) between predicted and ground truth lanes. A predicted lane is considered a true positive (TP) when the IoU exceeds a predefined threshold (typically 0.5); otherwise, it is classified as a false positive (FP) if no matched ground truth exists, or false negative (FN) if a ground truth lane lacks a corresponding prediction. The F1-score is defined as in (16):

$$F_1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

Where $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$.

Following COCO [40] detection metric, we also report mF1 to better compare the localization performance of algorithms. It is defined as in (17):

$$mF_1 = (F_{1@50} + F_{1@55} + \cdots + F_{1@95})/10 \tag{17}$$

Where $F_{1@50}$, $F_{1@55}$, …, $F_{1@95}$ are F1 metrics when IoU thresholds are 0.5, 0.55, …, 0.95 respectively. This metric rewards detectors with better localization performance across varying matching criteria.
For the TuSimple dataset, the official evaluation metric is accuracy, formulated as in (18):

$$\text{Accuracy} = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}} \tag{18}$$

Where C_{clip} represents the number of correctly predicted points, and S_{clip} denotes the total number of ground truth points. The ground truth lane width is 20 pixels. For a predicted lane to be considered a true positive (TP), the ratio of its correctly predicted points must exceed 85%. Otherwise, it is classified as either a false positive (FP) or a false negative (FN). These classifications are subsequently used to calculate the F1-score.

4.2. Implementation details

Training Configuration: On the CULane dataset, we employ ResNet-18, ResNet-34, and ResNet-101 as our backbones, initialized with ImageNet pre-trained weights. The model is optimized using AdamW with an initial learning rate of 2e-4 and a weight decay of 0.01. A cosine annealing scheduler (T_max=100, eta_min=1e-5) is applied. For data augmentation, we utilize random affine transformations (translation, rotation, scaling) and horizontal flipping. All models are trained on an NVIDIA RTX 3060 GPU with a batch size of 6.
Dataset-specific Settings: On the CULane dataset, the model is trained for 15 epochs. All input images are resized to 320 × 800 pixels for both training and testing. For a fair comparison of computational efficiency on CULane, all reported FPS (Frames Per Second) are measured with this 320×800 resolution at a batch size of 1, using PyTorch 1.12.1 and CUDA 11.8 on the same NVIDIA RTX 3060 GPU. For completeness, the training epochs for TuSimple and LLAMAS are set to 20 each, following common practice.
Architectural Hyperparameters: In the DF-NL module, we use 8 attention heads with 9 sampling points per head. The DAM-FF module processes multi-scale features from ResNet stages C2 to C5, producing output features with 64 channels.

4.3. Comparison with the state-of-the-art approach

4.3.1. Performance on CULane

Our proposed MPDDNet was evaluated on the challenging CULane benchmark and compared with state-of-the-art lane detection methods. As shown in Table 1, our MPDDNet establishes new state-of-the-art performance on the CULane benchmark. The ResNet-34 variant achieves the highest overall F1 score of 83.03% and mF1 score of 63.54%, significantly outperforming all existing methods. Notably, MPDDNet demonstrates consistent and superior performance across almost all challenging scenarios. It achieves remarkable results in difficult conditions such as No line (58.87%), Curve (77.15%), and Night (78.66%), highlighting its robustness in handling faint markings, complex geometries, and poor illumination.
The performance variations observed in our MPDDNet model across different backbone networks primarily stem from the alignment between their architectural characteristics and task requirements. ResNet-34 achieves the optimal balance in overall performance, where its moderate network depth and complexity enable it to simultaneously accommodate both detailed features and semantic information, thus delivering stable performance in most scenarios. In contrast, ResNet-101, with its deeper network structure, performs best in the most challenging No line and Dazzle light scenarios, indicating that its stronger feature extraction capabilities provide advantages in extreme situations, allowing it to extract more discriminative features from complex visual information. However, due to the increased parameter count, ResNet-101 may exhibit slight overfitting in certain scenarios, resulting in slightly lower performance in curve detection compared to the ResNet-34 variant. On the other hand, the ResNet-18 variant, with its lightweight architecture, achieves an inference speed of 89 FPS. Although its accuracy is slightly lower than the deeper networks, it still significantly outperforms other segmentation-based methods such as SCNN (7.5 FPS) and RESA (45.5 FPS), providing a viable solution for real-time applications.
Although segmentation-based methods generally achieve lower FPS than detection-based approaches, our method not only leads in performance among segmentation-based paradigms but also comfortably exceeds the 30 FPS requirement for real-time applications. This makes MPDDNet particularly suitable for practical autonomous driving systems, where both accuracy and real-time performance are crucial. The choice of different backbone networks offers flexibility for practical deployment: ResNet-18 is suitable for applications with extremely high-speed requirements, ResNet-34 achieves the optimal balance between accuracy and speed, while ResNet-101 provides a solution for specific complex scenarios where detection accuracy is paramount.

Table 1: State-of-the-Art Results on Culane. The Evaluation Metric for All Scenarios Is the F1 Score with an IoU Threshold of 0.5. for the Cross Scenario, Only False Positives are Shown. FPS Is Measured Based on the Pytorch Framework

Method	mF1	F1@50	Normal	Crowd	Dazzle	Shadow	No line	Arrow	Curve	Cross↓	Night	FPS
SCNN(VGG16)	38.84	71.60	90.60	69.70	58.50	66.90	43.40	84.10	64.40	1990	66.10	7.5
RESA(ResNet-34)	-	74.50	91.90	72.40	66.50	72.00	46.30	88.10	68.60	1896	69.80	45.5
RESA(ResNet-50)	47.86	75.30	92.10	73.10	69.20	72.80	47.70	88.30	70.30	1503	69.90	35.7
UFLD(ResNet-18)	38.94	68.40	87.70	66.00	58.40	62.80	40.20	81.00	57.90	1743	62.10	282
UFLD(ResNet-34)	-	72.30	90.70	70.20	59.50	69.30	44.40	85.70	69.50	2037	66.70	170
LaneATT(ResNet-18)	-	75.09	91.11	72.96	65.72	70.91	48.35	85.49	63.37	1170	68.95	250
LaneATT(ResNet-34)	-	76.68	92.14	75.03	66.47	78.15	49.39	88.38	67.72	1330	70.72	171
CLRNet(ResNet-18)	55.23	79.58	93.30	78.33	73.71	79.66	53.14	90.25	71.56	1321	75.11	119
CLRNet(ResNet-34)	55.14	79.73	93.49	78.06	74.57	79.92	54.01	90.59	72.77	1216	75.02	103
CLRNet(ResNet-101)	55.64	80.47	93.73	79.59	75.30	82.51	54.58	90.62	74.13	1155	75.37	94
HGLNet(ResNet-18)	55.83	80.65	93.48	78.31	75.13	81.75	53.74	89.98	73.27	959	75.06	116
HGLNet(ResNet-34)	56.07	81.23	93.76	78.89	75.29	82.21	54.95	90.43	74.95	1023	75.47	133
HGLNet(ResNet-101)	56.24	81.40	93.74	79.91	75.81	83.34	55.61	90.78	75.65	1240	76.01	58
MPDDNet(ResNet-18)	61.73	82.92	95.69	82.44	78.25	84.02	58.01	92.07	76.73	1278	78.66	89
MPDDNet(ResNet-34)	63.54	83.03	95.97	82.14	77.24	83.81	56.11	92.06	77.15	981	78.66	83
MPDDNet(ResNet-101)	61.66	82.80	95.90	81.52	79.88	83.09	58.87	91.92	75.42	1369	78.67	42

Fig. 5 presents a comprehensive visual comparison of lane detection performance under six challenging scenarios on the CULane dataset, featuring qualitative results from RESA, LaneATT, CLRNet, HGLNet, and our MPDDNet (all using ResNet-34 backbone). The visual results demonstrate that our MPDDNet achieves the closest alignment with ground truth (GT) annotations, particularly excelling in Dazzle, Shadow, and the most challenging No-line conditions. While other models exhibit incomplete predictions, fragmented outputs, or even complete failure in detecting lanes—especially in No-line scenarios where markings are absent or severely degraded—our method effectively handles strong glare, lighting variations, and missing markings through global context modeling and direction-aware feature fusion. These comparisons substantiate that MPDDNet maintains superior lane continuity and completeness where conventional methods struggle, demonstrating remarkable robustness in adverse driving environments.

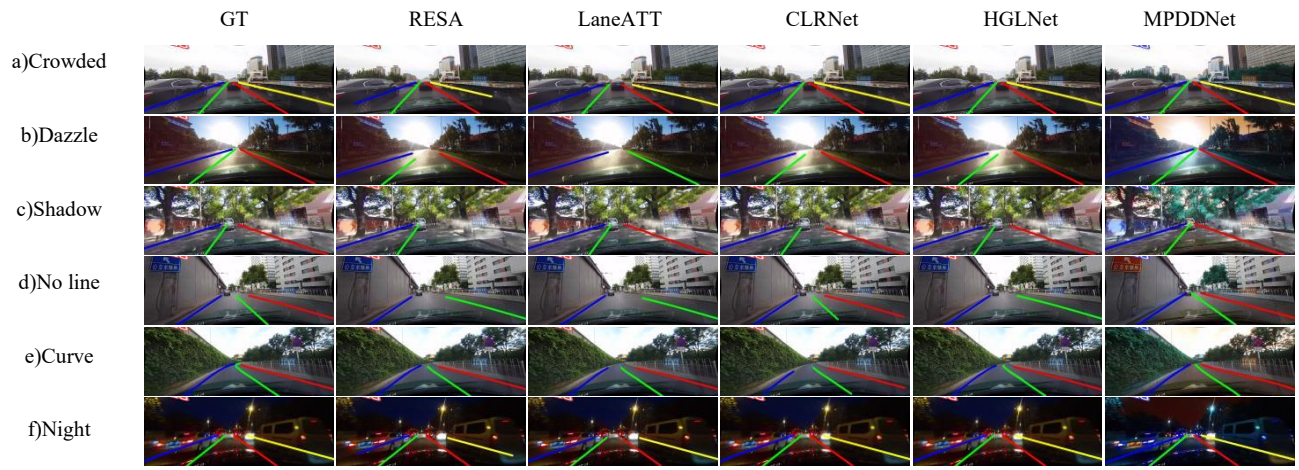


Fig. 5: Presents A Visual Comparison of Lane Detection Results Under Six Challenging Scenarios on the Culane Dataset, Featuring Outputs from RESA, Laneatt, Clrnet, Hglnet, and Our Proposed Method.

4.3.2. Performance on LLAMAS

The evaluation on the LLAMAS benchmark further validates the superior performance of our MPDDNet, as detailed in Table 2. Our MPDDNet achieves state-of-the-art performance on the LLAMAS benchmark. On the validation set, our ResNet-34 variant establishes new records with 72.40% mF1 and 97.80% F1@50, significantly outperforming HGLNet (71.66% mF1, 97.98% F1@50) and CLRNet (71.61% mF1, 97.16% F1@50). This performance advantage is consistently maintained on the test set, where our method achieves 96.95% F1@50 with a ResNet-34 backbone. The superior mF1 scores across all our model variants demonstrate enhanced localization precision, attributable to our direction-aware fusion mechanism and deformable non-local attention that effectively capture both structural details and global context. Notably, even our ResNet-18 configuration surpasses all existing methods in overall performance metrics, confirming the efficiency of our architecture design. These results validate the strong generalization capability of MPDDNet in structured highway environments and its robustness in precise lane localization tasks.

Table 2: Comparison with Popular Methods on LLAMAS

Method	Backbone	valid			test
		mF1	F1@50	F1@75	F1@50
PolyLaneNet	EfficientnetB0	48.82	90.2	45.40	88.40
LaneATT	ResNet18	69.22	94.64	82.36	93.46
LaneATT	ResNet34	69.63	94.96	82.79	93.74
LaneATT	ResNet122	70.8	95.17	84.01	93.54
LaneAF	DLA34	69.31	96.90	84.71	96.07
CLRNet	ResNet18	71.61	96.96	85.59	96.00
CLRNet	DLA34	71.21	97.16	85.33	96.12
HGLNet	ResNet18	71.46	96.74	85.57	95.99
HGLNet	DLA34	71.66	97.98	87.10	96.20
MPDDNet	ResNet18	72.15	97.45	86.25	96.65
MPDDNet	ResNet34	72.40	97.80	86.60	96.95
MPDDNet	ResNet101	72.35	97.15	86.10	96.35

4.3.3. Performance on tusimple

As shown in Table 3, the performance difference between different methods on this dataset is minimal, indicating that the accuracy on TuSimple appears to be nearly saturated. Despite this, our method achieves a new state-of-the-art with a 98.40% F1 score and surpasses the previous best method by 0.51% F1 score. This significant improvement demonstrates the effectiveness of our approach. Meanwhile, all versions of our method achieve lower FP and FN rates compared to other methods, which firmly demonstrates that MPDDNet can reliably predict lane lines even in challenging scenarios while maintaining superior detection accuracy.

Table 3: Comparison with Popular Methods on TuSimple

Method	Backbone	F1 (%)	Acc (%)	FP (%)	FN (%)
SCNN	VGG16	95.97	96.53	6.17	1.80
RESA	ResNet34	96.93	96.82	3.63	2.48
PolyLaneNet	EfficientNetB0	90.62	93.36	9.42	9.33
UFLD	ResNet34	88.02	95.86	18.91	3.75
LaneATT	ResNet122	96.06	96.10	5.64	2.17
CondLaneNet	ResNet101	97.24	96.54	2.01	3.50
CLRNet	ResNet18	97.89	96.84	2.28	1.92
HGLNet	ResNet101	97.82	96.74	1.81	2.57
MPDDNet	ResNet-18	98.25	97.05	1.65	1.45
MPDDNet	ResNet-34	98.40	97.18	1.52	1.38
MPDDNet	ResNet-101	98.35	97.02	1.62	1.46

4.4. Ablation studies

We conduct detailed ablation experiments to systematically evaluate the contributions of the proposed DAM-FF and DF-NL modules. The following analyses are based on the results shown in Table 4, where all models are built upon the ResNet-34 backbone.

Table 4: Effects of Each Component in our Method. Results are Reported on CULane.

Baseline	DAM-FF	DF-NL	mF1	F1	Shadow	No line	Cross	Curve	Dazzle
✓			58.50	79.40	76.90	48.40	2050	72.10	70.60
✓	✓		61.80	81.90	84.70	58.30	2200	75.90	76.10
✓	✓	✓	63.54	83.03	83.81	56.11	981	77.15	77.24

Effectiveness of DAM-FF. The introduction of the DAM-FF module leads to a notable enhancement in detection performance, elevating the mF1 score from 58.50% to 61.80%. This module proves particularly beneficial in low-visibility conditions, where it achieves F1 score gains of 7.80 and 9.90 in Shadow and No-line scenarios, respectively. These improvements confirm that the direction-aware multi-scale fusion mechanism successfully mitigates the effects of adverse lighting conditions and reconstructs deteriorated lane markings through effective integration of contextual information across multiple scales and orientations.

Impact of DF-NL Integration. The addition of the DF-NL module further advances the model's capability in handling structurally complex environments. This integration boosts the mF1 score by an additional 1.74 points, reaching 63.54%. A particularly remarkable improvement is observed in the Cross scenario, where the metric (lower values indicate better performance) shows a substantial reduction from 2200 to 981. Simultaneously, the model achieves its optimal performance in Curve detection at 77.15. Although minor performance decreases are observed in Shadow and No-line scenarios compared to the DAM-FF-only configuration, the substantial gains in structurally demanding situations underscore DF-NL's effectiveness in capturing global contextual relationships and modeling long-range dependencies.

The integrated MPDDNet framework, combining both DAM-FF and DF-NL, delivers the most balanced and robust performance across diverse challenging conditions. This outcome validates two key design aspects: first, DAM-FF successfully strengthens local feature representation to address illumination changes and lane visibility issues; second, DF-NL provides complementary global structural understanding for complex road geometries. The synergistic operation of these modules establishes a comprehensive solution for reliable lane detection in varied driving environments.

4.5. Future work

While the proposed MPDDNet demonstrates compelling performance, several avenues remain for further exploration. First, we plan to conduct a more granular performance analysis, including per-class confusion studies (e.g., curve vs. intersection) and a detailed latency breakdown across the backbone and key modules. Second, a comprehensive comparison with state-of-the-art real-time multi-task models (e.g., YOLO, HybridNets) will be pursued to better situate our method's efficiency-accuracy trade-off. Looking forward, we aim to extend the core DF-NL concept into 3D Bird's-Eye View space for unified perception, integrate temporal modeling for consistency in video sequences, and ultimately deploy and optimize the framework on embedded platforms like NVIDIA Jetson to validate its practicality in real-world autonomous driving systems.

5. Conclusion

This paper presents MPDDNet, a novel multi-scale parallel network for lane detection in complex driving environments. Our method effectively addresses two key challenges: robust feature representation under degraded conditions through the Direction-aware Adaptive Multi-scale Feature Fusion (DAM-FF) module, and effective global context modeling via the Deformable Non-local (DF-NL) module.

The proposed DAM-FF module explicitly embeds directional priors into multi-scale feature fusion, significantly enhancing detailed feature representation in challenging scenarios. The DF-NL module enables efficient global context modeling while implicitly incorporating structural priors of lane geometry. Through parallel integration of these complementary modules, MPDDNet achieves simultaneous optimization of local details and global semantics.

Extensive experiments demonstrate that our method establishes new state-of-the-art performance on multiple benchmarks. On CULane, our MPDDNet achieves the highest overall F1 score of 83.03% and mF1 score of 63.54%, with particularly strong performance in challenging scenarios. The consistent superiority across LLAMAS and TuSimple benchmarks further validates our approach's generalization capability.

While the current implementation achieves a good balance between accuracy and efficiency, future work will explore more light-weight implementations, extend the direction-aware paradigm to other structural perception tasks, and further investigate cross-dataset generalization, 3D lane detection, and uncertainty estimation. Our work provides valuable insights for lane detection in complex environments and offers a solid foundation for future research in autonomous driving perception systems.

Acknowledgement

This work is funded by the Research Fund of Nanning University.

References

[1] X. He et al., “Monocular Lane Detection Based on Deep Learning: A Survey,” Dec. 11, 2024, arXiv: arXiv:2411.16316.
[2] J. Gamerdinger, S. Teufel, and O. Bringmann, “Datasets for Lane Detection in Autonomous Driving: A Comprehensive Review,” Apr. 11, 2025, arXiv: arXiv:2504.08540.
[3] Y. Zhang, Z. Tu, and F. Lyu, “A Review of Lane Detection Based on Deep Learning Methods,” Mech. Eng. Sci., vol. 5, no. 2, May 2024, https://doi.org/10.33142/mes.v5i2.12721.
[4] T. Zheng et al., “CLRNet: Cross Layer Refinement Network for Lane Detection,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 888–897. https://doi.org/10.1109/CVPR52688.2022.00097
[5] W. Hao, “Review on lane detection and related methods,” Cognitive Robotics, vol. 3, pp. 135–141, 2023, https://doi.org/10.1016/j.cogr.2023.05.004.

[6] H. Honda and Y. Uchida, “CLRerNet: Improving Confidence of Lane Detection with LaneIoU,” May 15, 2023, arXiv: arXiv:2305.08366. Accessed: May 17, 2024. [Online]. Available: http://arxiv.org/abs/2305.08366.

[7] Z. Chen, Y. Liu, M. Gong, B. Du, G. Qian, and K. Smith-Miles, “Generating Dynamic Kernels via Transformers for Lane Detection,” in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 6812–6821. https://doi.org/10.1109/ICCV51070.2023.00629.

[8] Q. Chang and Y. Tong, “A Hybrid Global-Local Perception Network for Lane Detection,” Proc. AAAI Conf. Artif. Intell., vol. 38, no. 2, pp. 981–989, Mar. 2024, https://doi.org/10.1609/aaai.v38i2.27858.

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," Mar. 17, 2021, arXiv: arXiv:2010.04159. Accessed: Jun. 15, 2024. [Online]. Available: http://arxiv.org/abs/2010.04159.

[10] TuSimple/tusimple-benchmark. (Oct. 04, 2025). Jupyter Notebook. TuSimple. Accessed: Oct. 18, 2025. [Online]. Available: https://github.com/TuSimple/tusimple-benchmark.

[11] Z. Cheng, G. Zhang, C. Wang, and W. Zhou, “DILane: Dynamic Instance-Aware Network for Lane Detection,” in Computer Vision – ACCV 2022, vol. 13842, L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa, Eds., in Lecture Notes in Computer Science, vol. 13842. , Cham: Springer Nature Switzerland, 2023, pp. 124–140. https://doi.org/10.1007/978-3-031-26284-5_8.

[12] Q. Qiu, H. Gao, W. Hua, G. Huang, and X. He, “PriorLane: A Prior Knowledge Enhanced Lane Detection Approach Based on Transformer,” Feb. 07, 2023, arXiv: arXiv:2209.06994. Accessed: Jul. 25, 2024. [Online]. Available: http://arxiv.org/abs/2209.06994.

[13] J.-Q. Zhang, H.-B. Duan, J.-L. Chen, A. Shamir, and M. Wang, "HoughLaneNet: Lane Detection with Deep Hough Transform and Dynamic Convolution," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, 2023, pp. 19154-19164.

[14] C. Chen, J. Liu, C. Zhou, J. Tang, and G. Wu, “Sketch and Refine: Towards Fast and Accurate Lane Detection,” Jan. 26, 2024, arXiv: arXiv:2401.14729. Accessed: Jul. 25, 2024. [Online]. Available: http://arxiv.org/abs/2401.14729.

[15] C. Chen, J. Liu, C. Zhou, J. Tang, and G. Wu, "Sketch and Refine: Towards Fast and Accurate Lane Detection," Jan. 26, 2024, arXiv: arXiv:2401.14729. Accessed: Jul. 25, 2024. [Online]. Available: http://arxiv.org/abs/2401.14729L. Gao, Z. Qu, S. Wang, and S. Xia, “A Lightweight Neural Network Model of Feature Pyramid and Attention Mechanism for Traffic Object Detection,” IEEE Trans. Intell. Veh., vol. 9, no. 2, pp. 3422–3435, 2024, https://doi.org/10.1109/TIV.2023.3345271.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, 2017, pp. 936-944. https://doi.org/10.1109/CVPR.2017.106.

[17] Z. Qin, P. Zhang, and X. Li, “Ultra Fast Deep Lane Detection with Hybrid Anchor Driven Ordinal Classification,” Jun. 15, 2022, arXiv: arXiv:2206.07389. Accessed: Jul. 31, 2024. [Online]. Available: http://arxiv.org/abs/2206.07389.

[18] L. Tabelini, R. Berriel, T. M. Paixão, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Keep your Eyes on the Lane: Real-time Attention-guided Lane Detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, 2021, pp. 294-303. https://doi.org/10.1109/CVPR46437.2021.00036

[19] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial As Deep: Spatial CNN for Traffic Scene Understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018. https://doi.org/10.1609/aaai.v32i1.12301

[20] J. Su, C. Chen, K. Zhang, J. Luo, and X. Wei, "Structure Guided Lane Detection," in Proceedings of the 30th ACM International Conference on Multimedia (MM '22), New York, NY, USA: ACM, 2022, pp. 1933–1941. https://doi.org/10.1145/3503161.3547915.

[21] T. Zheng et al., "RESA: Recurrent Feature-Shift Aggregator for Lane Detection," in *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021, pp. 3547-3555. https://doi.org/10.1609/aaai.v35i4.16469

[22] Z. Qin, H. Wang, and X. Li, "Ultra Fast Structure-aware Deep Lane Detection," in Computer Vision – ECCV 2020, vol. 12369, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 276–291. https://doi.org/10.1007/978-3-030-58586-0_17

[23] L. Liu, X. Chen, S. Zhu, and P. Tan, "CondLaneNet: a Top-to-down Lane Detection Framework Based on Conditional Convolution," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, 2021, pp. 3753-3762. https://doi.org/10.1109/ICCV48922.2021.00375.

[24] L. Tabelini, R. Berriel, T. M. Paixão, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "PolyLaneNet: Lane Estimation via Deep Polynomial Regression," in 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy: IEEE, 2021, pp. 6150-6156. https://doi.org/10.1109/ICPR48806.2021.9412201.

[25] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, “End-to-end Lane Shape Prediction with Transformers,” in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 3693–3701. https://doi.org/10.1109/WACV48630.2021.00374.

[26] A. Vaswani et al., “Attention is All You Need,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Jun. 26, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[27] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path Aggregation Network for Instance Segmentation,” Sep. 18, 2018, arXiv: arXiv:1803.01534. Accessed: Jun. 15, 2024. [Online]. Available: http://arxiv.org/abs/1803.01534.

[28] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT: IEEE, 2018, pp. 8759-8768. https://doi.org/10.1109/CVPR.2018.00913.

[29] L. Chen et al., "PersFormer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, 2022, pp. 1051-1060.

[30] B. Zhang, L. Zhang, T. Wang, Y. Wei, Z. Chen, and B. Cao, “Omni-Refinement Attention Network for Lane Detection,” Sensors, vol. 25, no. 19, p. 6150, Oct. 2025, https://doi.org/10.3390/s25196150.

[31] 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, 2024, https://doi.org/10.1109/TITS.2023.3309948.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local Neural Networks,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 7794–7803. https://doi.org/10.1109/CVPR.2018.00813.

[33] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in International Conference on Learning Representations (ICLR), 2021.

[34] K. Zhou and R. Zhou, “End-to-End Lane detection with One-to-Several Transformer,” May 13, 2023, arXiv: arXiv:2305.00675. Accessed: May 17, 2024. [Online]. Available: http://arxiv.org/abs/2305.00675.

[35] L. Zhuang, T. Jiang, M. Qiu, A. Wang, and Z. Huang, “Transformer Generates Conditional Convolution Kernels for End-to-End Lane Detection,” IEEE Sens. J., vol. 24, no. 17, pp. 28383–28396, Sep. 2024, https://doi.org/10.1109/JSEN.2024.3430234.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, 2020, pp. 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155.

[38] Z. Yang et al., “LDTR: Transformer-based Lane Detection with Anchor-chain Representation,” Mar. 21, 2024, arXiv: arXiv:2403.14354. Accessed: Jul. 25, 2024. [Online]. Available: http://arxiv.org/abs/2403.14354.

[39] K. Behrendt and R. Soussan, “Unsupervised Labeled Lane Markers Using Maps,” in 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South): IEEE, Oct. 2019, pp. 832–839. https://doi.org/10.1109/ICCVW.2019.00111.

[40] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755.