Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **9**, Number **4**, November **2024**

A Real-time Semantic Segmentation Model for Lane Detection

Chen-Xu Ma, Jing-Ang Li, Yong-Hua Han*, Yu-Meng Wang

School of Information Science and Engineering,
Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China
chenxv832@gmail.com, lja19816897163@163.com, han_yong_huahan@zstu.edu.cn
2387035786@qq.com

Hai-Bo Mu

Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou 310051, P. R. China
muhaibo@hikvision.com

Lu-Rong Jiang

School of Information Science and Engineering,
Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China
jianglurong@zstu.edu.cn

*Corresponding author: Yong-Hua Han
Received November 16, 2023, revised February 17, 2024, accepted June 1, 2024.

Abstract. *To bolster the safety of autonomous and assisted driving systems, the imperative of achieving a synergy between real-time processing and high accuracy in lane detection cannot be overstated. Addressing the challenges posed by the intricate nature of lane detection algorithms and the concomitant degradation of accuracy due to the loss of information on small-scale targets, this study introduces an enhanced lane detection model predicated on the DeeplabV3+ framework. The model integrates the lightweight MobilenetV2 as the foundational backbone network to meet the exigencies of real-time operation. In parallel, the incorporation of the Multi-scale Feature Extraction Enhancement Module is meticulously designed to counter the heterogeneous distribution of lane dimensions, thereby bolstering the model's capability to accurately predict diminutive targets, including marginal lanes and those at extended distances. In an innovative stride, this research proposes the Convolutional Block Weighted Attention Module, meticulously devised to refine the distribution of attentional resources across both channels and spatial dimensions, which in turn augments the model's efficacy in processing clusters of pixel points within homogenous semantic classifications. The Feature Fusion Module is judiciously engineered to produce semantically enriched feature maps. By implementing skip connections at strategic junctures between the encoding and decoding layers, the model achieves an efficacious fusion of features across varying depths, culminating in a marked enhancement of segmentation performance.Empirical analysis conducted on a representative dataset corroborates the model's prowess, as evidenced by an impressive 99.48% Accuracy and an 88.22% mIoU, all while maintaining a brisk prediction latency of merely 35.12 ms per image. These findings underscore the proposed model's exceptional capacity to deliver real-time performance without compromising on accuracy, setting a new benchmark in the domain of lane detection.*
Keywords: real-time semantic segmentation; lane detection; attention module

1. **Introduction.** Lane detection, a fundamental component of autonomous driving technology, is imperative for augmenting road safety, facilitating the secure maneuvering of vehicles, and averting lane deviations. Its significance extends beyond the domain of autonomous driving to encompass urban traffic planning and administration, where it equips policymakers with essential insights into road utilization and traffic patterns [1].

The introduction of deep learning has significantly enhanced lane detection capabilities by automating feature extraction, diminishing the dependence on human expertise, reducing error frequencies, and refining operational workflows [2]. Through the deployment of multi-layer neural networks, a more nuanced abstraction of features is attained, enabling a thorough semantic analysis of the imagery, which in turn, bolsters the overall efficacy [3]. Additionally, deep learning models exhibit exceptional proficiency in fitting, generalizing, and processing in parallel when managing voluminous datasets [4]. Notwithstanding the manifold benefits conferred by deep learning, lane detection technologies encounter a plethora of challenges, prompted by the dynamic nature of actual driving surroundings. To depict the intricacies of lane detection across diverse environments more graphically, FIGURE 1 displays a selection of representative lane marking scenarios.

In the dynamic and unpredictable environment of real-world driving, sustaining both the precision and immediacy of detection remains a pivotal area of research emphasis. Contemporary studies are largely focused on augmenting the accuracy of lane detection, with a particular emphasis on performance within intricate scenarios. As an exemplar, the joint learning algorithm predicated on attention-FCN, as proposed by Wang et al., has significantly ameliorated segmentation accuracy, notably yielding substantial advancements in the detection of diminutive targets [5]. This research dovetails with the aims of the present study, which endeavors to formulate an accurate and efficient deep learning model predicated on the characteristics of lane markings, through the integration of advanced techniques such as attention mechanisms.

In the domain of lane detection, the imperative to reconcile high precision with minimal inference latency is paramount. Hence, the design of deep learning models mandates a judicious balance between the quantity of parameters and the complexity of the model within a logical framework, to ensure parity between the accuracy of lane detection and the speed of inference.

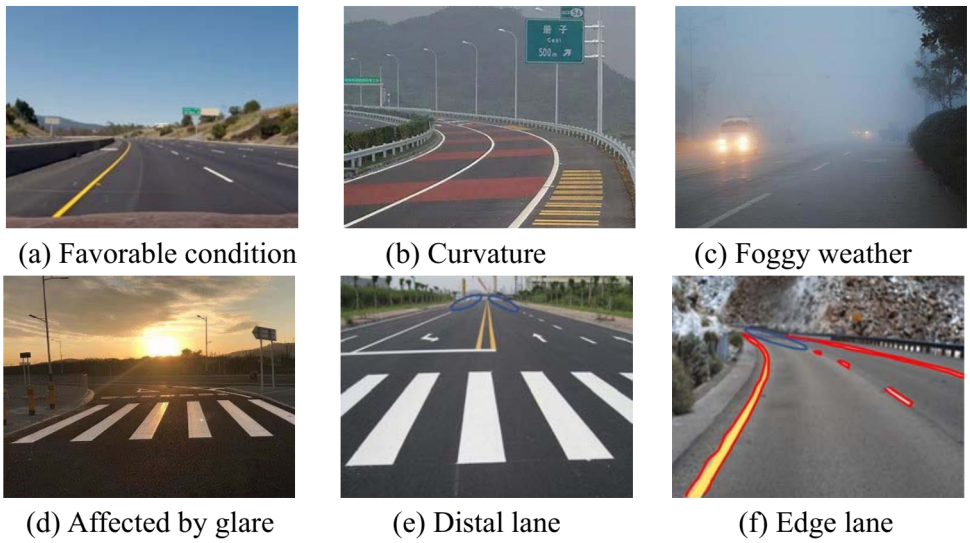

Figure 1. Typical lane scenario image

In FIGURE 1, Panel (a) presents an image of lane markings captured under optimal weather conditions; panel (b) depicts lane markings characterized by pronounced curvature; panel (c) displays an image of lane markings under foggy conditions; panel (d) illustrates lane markings affected by glare; panel (e) identifies distant lane markings, specifically the aggregation of pixel points at the terminus of the lane markings, as highlighted by the blue annotations within the figure; and panel (f) portrays edge lane markings, which are the collective pixel points situated at the periphery of the lane markings, as demarcated by the red annotations in the figure.

Acknowledging these constraints, a suite of lightweight networks and other novel solutions have been advanced to ameliorate lane detection performance. Initial deep learning constructs for lane detection were predicated on frameworks such as FCN [6], UNet [7], PSPNet [8], and SCNN [9]. The FCN is capable of processing inputs of arbitrary dimensions and employs skip connections to fortify accuracy, yet it does not encompass multi-scale features, thereby circumscribing its precision. UNet contemplates multi-scale features but is encumbered by a symmetrical architecture that imposes significant computational demands, thus impinging upon the real-time capabilities of lane detection. PSPNet amalgamates global contextual information and incorporates a pyramid design to address the challenge of recognizing small targets at a singular scale, albeit with compromised accuracy. The SCNN model harnesses spatial convolutions to facilitate the learning of lane marking features across disparate spatial locales, yet it necessitates considerable computational outlay during training and does not sufficiently cater to the detection of smaller targets. These progresses intimate that, notwithstanding the extant challenges, further enhancements in lane detection technology are feasible through the pursuit of innovative methodologies.

In response to the intricacy and computational exigencies of such models, lightweight networks have garnered extensive application. Quintessential lightweight networks include MobileNet [10], ShuffleNet [11], and EfficientNet [12]. While these archetypal lightweight networks are commendable for their real-time execution, their precision in detection and generalization capabilities are somewhat lacking. Researchers have proffered a plethora of remedial strategies to counteract these deficiencies. Hou et al. have introduced Self-Attention Distillation (SAD), which has been corroborated on architectures such as ENet and has demonstrated efficacy in congested conditions and under poor lighting, yet it has not been exhaustively evaluated in complex settings such as lanes with significant curvature [13]. Chen et al. have advocated for a lightweight UNet model that offers promising real-time performance, but it does not adequately address scenarios with variable lighting conditions [14]. Yao et al. have developed a dual-branch real-time lane detection model that captures both global and local detail features, showing proficiency in detecting both straight and undulating lane markings, but it falls short in addressing complex conditions like low illumination [15]. Song et al. have proposed the LLSS-Net, which exhibits exceptional detection capabilities in low-light scenarios, but its research does not extend to other multifaceted environments [16]. Our proposed methodology for lane detection aspires to rectify these enumerated inadequacies.

Within the complex and ever-evolving realm of driving environments, lane detection technology confronts several challenges, which can be itemized as follows: (1) The necessity for augmented accuracy, given that intricate scenarios incorporate visual complications such as suboptimal lighting, glare, low levels of illumination, and the presence of confounding elements including road markings, stains, wear, and obstructions. The adaptability of neural network models is paramount [17]; (2) The procurement and labeling of datasets entail substantial expenditure [18], complicating the comprehensive representation of diverse complex scenarios, which, in turn, impinges upon the model's

generalizability and precision; (3) An imbalance between positive and negative samples predisposes the model to a predilection for background prediction [19]; (4) The simultaneous attainment of accuracy and real-time performance presents a significant challenge.

To address challenges (1) and (4), we propose a lane detection model predicated on DeeplabV3+, which incorporates several novel features: (1) An efficacious Multi-scale Feature Extraction Enhancement Module, which amalgamates depthwise separable dilated convolutions across a spectrum of sampling rates, thereby harvesting an extensive suite of multi-scale feature information and enhancing the model's discernment of peripheral and remote lane markings; (2) A Feature Fusion Module (FFM) that employs a series of compact convolutional kernels to refine edges, effectively ameliorating the loss of detail for diminutive targets; (3) The deployment of a Convolutional Block Weighted Attention Module (CBWAM), which appraises the significance of discrete blocks to eliminate analogous interferences, thereby bolstering the comprehensive segmentation accuracy of lane markings; (4) The adoption of skip connections to progressively amalgamate superficial features, remedying the suboptimal exploitation of the encoding layer's output features. In addressing challenge (2), the extant dataset has been enriched with images featuring lane markings of pronounced curvature, alongside the utilization of image enhancement techniques to emulate variations in outdoor illumination. Regarding challenge (3), Dice Loss has been incorporated within the loss function, encompassing the congruity between the predicted and actual values, thus effectuating an equilibrium between positive and negative samples.

In summation, this manuscript delineates an enhanced model based on DeeplabV3+, designed to surmount the obstacles inherent in lane detection amidst the variable and intricate conditions typical of real-world driving. This endeavor not only refines theoretical constructs within the domain of deep learning architectures but also empirically substantiates the efficacy of the model across a gamut of complex scenarios. The findings of this investigation are anticipated to significantly contribute to the safety and dependability of forthcoming autonomous driving systems and to underpin the evolution of intelligent traffic management infrastructures.

The remaining sections of the paper are structured as follows: Section two explicates the comprehensive design of the lane detection model predicated on the refined DeeplabV3+. Section three evaluates the performance of the proposed methodology in relation to alternative models and exhibits corresponding detection imagery. Finally, section four encapsulates the discourse of the entire document.

2. **Methodology.**

2.1. **General Network Architecture.**

Lane detection, as an integral element of autonomous and assisted driving systems, is vital for vehicle navigation and safety. Nonetheless, the accuracy of such detection is frequently undermined by a plethora of complex environmental factors, including uneven lighting conditions, degradation of lane markings, and the presence of diminutive and elusive targets like edge lanes and distant lanes. These impediments can precipitate a degradation in the efficacy of conventional models in lane segmentation. In an effort to surmount these limitations, we have instituted targeted enhancements to the DeeplabV3+ model, with the objective of amplifying both the precision and real-time processing capabilities of lane detection. The refined network architecture is illustrated in FIGURE 2.

The revised design primarily modifies the encoder configuration of DeeplabV3+. The upgraded encoder is composed of four distinct modules: the backbone network, the Multi-scale Feature Extraction Enhancement Module, the Feature Fusion Module (FFM), and

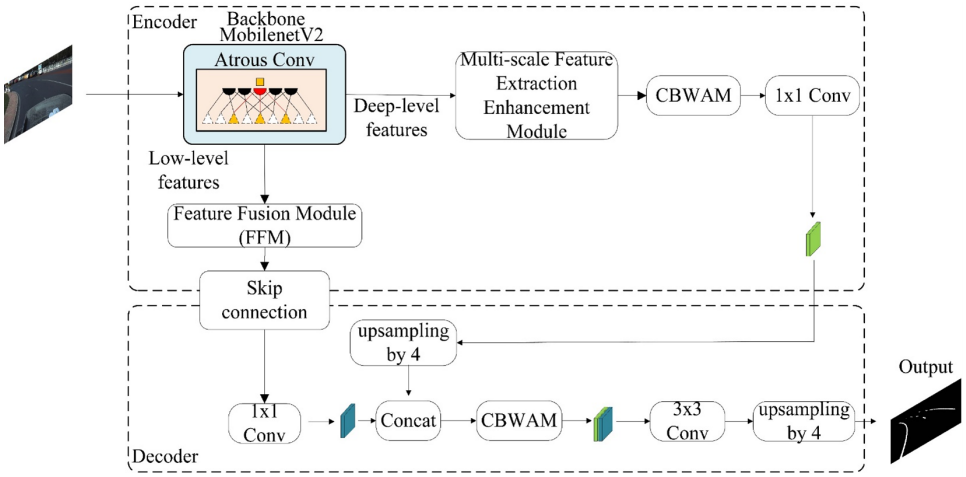

Figure 2. Overall network architecture diagram

skip connection. The backbone network's role is to distill primary features, such as color and texture, from the input lane imagery. To optimize real-time performance, the highly efficient MobilenetV2 serves as the backbone network. Following this, the Multi-scale Feature Extraction Enhancement Module is deployed to harvest lane characteristics across multiple scales, thereby enriching the model's depiction of lanes. Subsequently, the Feature Fusion Module (FFM) is meticulously crafted to refine the model's detection of intricate details, with a focus on edge lanes and distant lanes. Lastly, the Convolutional Block Weighted Attention Module (CBWAM) is integrated, further delineating between lane-containing regions and their counterparts, thereby fortifying the model's resilience to disruptions in multifaceted settings. The incorporation of skip connection throughout the encoder and decoder layers facilitates a synergistic fusion of features across various strata, providing the model with a superior feature representation capacity and, in turn, significantly advancing lane detection performance.

2.2. **Network Design Module Details.**

2.2.1. *Multi-scale Feature Extraction Enhancement Module.* In the domain of lane detection, lane morphology and distribution exhibit significant variability across different roadway segments, engendering a model's detection performance that is prone to perturbation by road type and environmental factors. The Atrous Spatial Pyramid Pooling (ASPP) framework, tasked with harvesting multi-scale feature information, is adept at securing features across disparate scales. However, it faces two salient challenges when tasked with processing diminutive lane targets: First, the conventional ASPP employs parallel dilated convolution layers with unduly broad sampling rate intervals, which may lead to a forfeiture of critical edge information in the detection of small targets; second, the profusion of parallel convolution layers burgeons the model's parameter count, impeding the development of streamlined lane detection models.

To mitigate these issues, the present study introduces an enhancement to the original ASPP framework, resulting in the formulation of a Multi-scale Feature Extraction Enhancement Module. The architecture of this module is delineated in FIGURE 3. This refined ASPP framework embodies two pivotal innovations: initially, the parallel dilated convolution branches are re-engineered with refined sampling rate combinations. Extending the foundational ASPP structure, additional parallel dilated convolution layers are incorporated, and the sampling rates are meticulously calibrated to encompass more

granular intervals, establishing a configuration of 3, 6, 9, 12, 15, 18, and 24. This arrangement empowers lower sampling rates, such as 3, 6, and 9, to apprehend a greater quantum of local feature information, whereas higher rates, such as 12, 15, 18, and 24, proffer an enlarged receptive field conducive to the assimilation of extensive global contextual information. Given that the 3×3 convolution at sampling rates of 18 and 24 is less parametrically dense, thus suboptimal for exhaustive small target feature extraction, it is equivalently transmuted into a 5×5 convolution at rates of 9 and 12. This conversion preserves the receptive field while diminishing parameter volume, concurrently amplifying the model's proficiency in segmenting diminutive targets, including edge lanes and remote lanes.

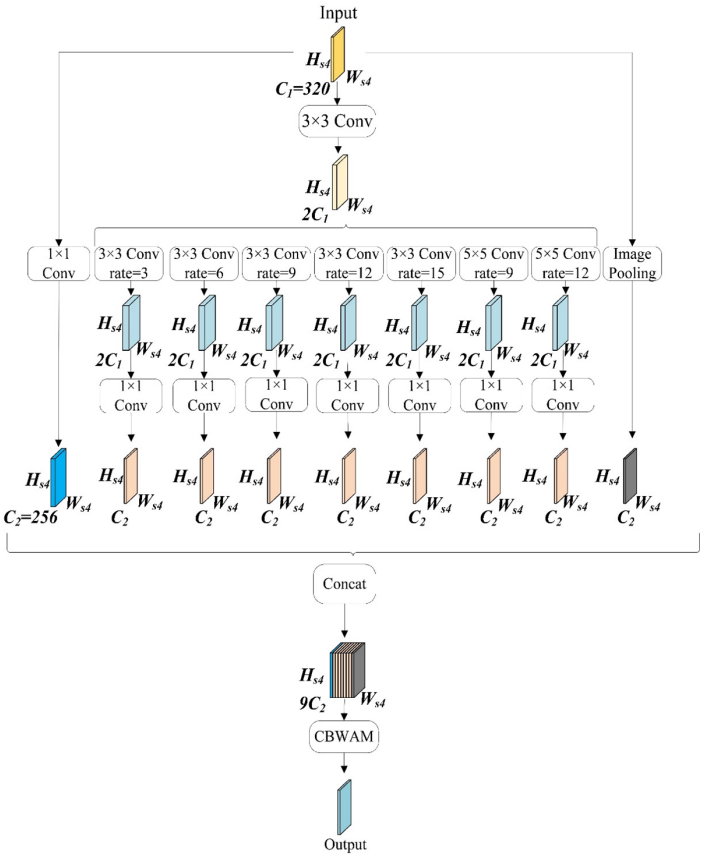

Figure 3. Multi-scale Feature Extraction Module

Secondarily, the study integrates Dilated Depthwise Separable Convolution as a surrogate for traditional dilated convolution layers. This convolution variant markedly trims the model's parameter bulk while preserving the efficacy of feature extraction. Dilated Depthwise Separable Convolution bifurcates the convolution process into two discrete stages: depthwise convolution and pointwise convolution. Within the depthwise convolution phase, each convolution kernel independently interacts with a singular channel of the input feature map, thereby conserving the channel count of the feature map. The ensuing 1×1 pointwise convolution modulates the channel quantity and forges inter-channel linkages. This bifurcated approach not only curtails parameter volume but also elevates computational efficiency. These methodological advancements render the Multi-scale Feature Extraction Enhancement Module not only more adept at meticulously capturing lane features but also significantly streamline the model's architecture, thus paving the way for real-time lane detection. By virtue of this module, our model attains enhanced proficiency

in extracting thoroughfares comprising small-sized lanes and in accurately demarcating lanes, thereby furnishing autonomous driving systems with more dependable visual intelligence.

2.2.2. *Feature Fusion Module(FFM).* Within the architecture of the backbone network, the capacity for feature representation of diminutive targets is markedly compromised following successive downsampling processes. This is particularly deleterious for the identification of nuanced targets, such as edge lanes and remote lanes. To augment the ability of the lane detection model to delineate features of small targets, we have integrated the principle of super-resolution upsampling into the design of our module, culminating in the development of a Feature Fusion Module (FFM). This module enhances the representational strength of the deep features harvested by the backbone network and mitigates the diminution of small target information resulting from serial downsampling.

As depicted in FIGURE 4, the engineered Feature Fusion Module consists of three core components: a Residual Module, Hybrid Dilated Convolution, and Super-resolution Upsampling. The Residual Module is initially deployed to refine the edge details within the deep features, establishing a more precise groundwork for the restoration of fine details. The Hybrid Dilated Convolution, through its employment of dilated convolutions with assorted sampling rates, amalgamates local and global contextual information, thereby amplifying feature expressiveness. The final component, Super-resolution Upsampling, employs sub-pixel convolution techniques to enhance the resolution of deep feature maps, reinstating the lost fine details whilst maintaining potent semantic information. The incorporation of the Feature Fusion Module not only elevates the precision of lane detection but also reinforces the model's proficiency in discerning small targets, thus increasing the model's robustness and adaptability across a spectrum of intricate roadway scenarios.

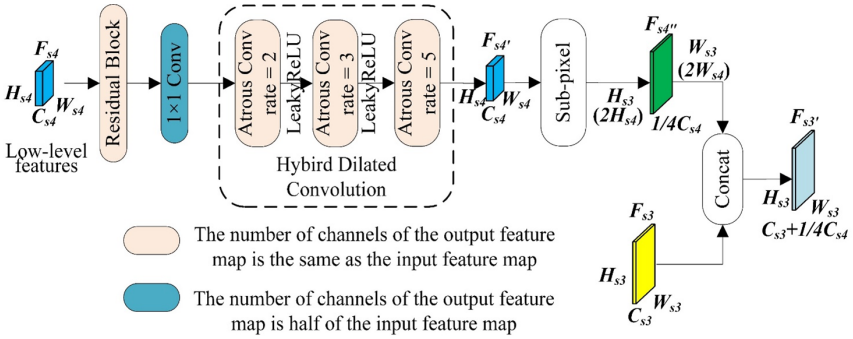

Figure 4. Block Weighted Attention Module

More specifically, the module capitalizes on the deep features, F_{s4}, derived from the 16-fold downsampling executed by the backbone network. Post sub-pixel convolution processing, it procures super-resolution features teeming with lane information, designated as F'_{s4}. These features are then disentangled to recapture the minutiae of the target features. The emergent F'_{s4} is integrated with the feature layer F_{s3}, sourced from 8-fold downsampling by the backbone network, producing a feature layer F'_{s3} that is replete with copious local feature information, consequently bolstering the accuracy of lane detection. This fusion approach not only refines the hierarchical structuring of features but also significantly enhances the detection precision of diminutive lane targets. Herein, F_{si} denotes the feature map acquired post $2i$-fold downsampling by the backbone network, where H and W represent the height and width, respectively, of the image input to the backbone network; each downsampling iteration reduces the feature map dimensions by

half. C_{si} signifies the channel quantity of the feature map following $2i$-fold downsampling by the backbone network, with i indicating the ith instance of downsampling by the backbone network.

2.2.2.1. Residual module. In the training of deep neural networks, particularly those with an extensive number of layers, practitioners frequently confront the phenomena of vanishing gradients. Such phenomena can complicate the training process and may precipitate a decline in network efficacy, which in turn impairs the propagation of features within the network. The deployment of a Residual Module can efficaciously mitigate this issue by leveraging a shortcut connection to sustain the flow of information, thereby facilitating more direct backward propagation of gradients. The architecture of this module is illustrated in FIGURE 5.

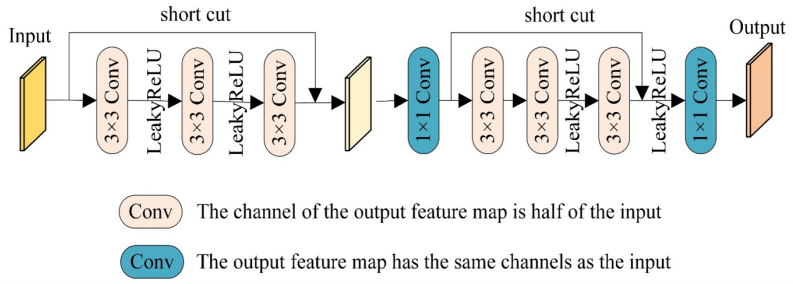

Figure 5. Residuals Module

Central to the Residual Module are two consecutive 3×3 convolutional layers, with each layer succeeded by a LeakyReLU activation function and a Batch Normalization (BN) layer. The LeakyReLU activation function, a modification of the ReLU function, permits the passage of slight negative gradient values, thereby resolving the issue of neuron inactivation, colloquially termed "neuron death," which occurs under the ReLU function when inputs are negative. The BN layer is utilized to normalize the features entering each layer, diminishing internal covariate shift, expediting the training process, and bolstering the model's stability.

The module's output, denoted as $H(x)$, integrates the original input, x, with the output of the non-linear transformation, $F(x)$. The residual connection facilitates the conveyance of the input feature, x, directly to the output without subjecting it to non-linear transformations, effectively creating a shortcut that lessens the risk of gradient vanishing during network training. The expression of the Residual Module's output, $H(x)$, is presented in Equation (1):

$$H(x) = F(x) + x \tag{1}$$

Here, x signifies the network's input, and $F(x)$ represents the output subsequent to the convolutional layers. The incorporation of the Residual Module enables our model to more efficiently preserve the gradient flow while processing profound information, circumventing information loss during training, and thereby augmenting the model's capacity for learning and generalization. This enhancement significantly improves the accuracy of detecting small-scale lane targets, including edge lanes and distant lanes.

2.2.2.2. Hybrid Dilated Convolution. In the architecture of neural networks, the expanse of the receptive field is of paramount importance for the assimilation of global information. This holds particularly true for the task of lane detection, wherein an expansive receptive field is instrumental in facilitating the model's interpretation of both the comprehensive layout and the local interconnections of lane demarcations within an image. To this

end, the present study introduces the concept of Hybrid Dilated Convolution (HDC), which seeks to augment the receptive field, thereby enabling the extraction of global information pertaining to lane markings from the input feature maps, while concurrently circumventing the gridding effect.

The gridding effect is characterized by the occurrence of unconvolved pixels within the feature map when dilated convolutions with larger dilation rates are employed, potentially leading to the omission of critical information. The HDC approach was formulated to overcome this challenge, predicated on a strategic combination of dilation rates for convolutional kernels, thus ensuring the comprehensive engagement of each pixel in the feature map.

Consider a series of N dilated convolutional layers, each of dimension $K \times K$, and with a progressive sequence of dilation rates $r = [r_1, r_2, r_3, \ldots, r_n]$. The application of the HDC paradigm ensures the maximal exploitation of pixel information in the feature map following the initial dilated convolutional kernel, thereby eliminating the presence of dormant pixels within the feature map. The maximal interspace between two nonzero elements within the ith layer is articulated in Equation (2):

$$M_i = \max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i] \tag{2}$$

Here, M_i represents the maximal interspace between two nonzero elements within the ith layer, and r_i denotes the dilation rate for the ith layer's dilated convolution, with the stipulation that $M_n = r_n$, signifying that the maximal interspace in the concluding layer is equivalent to the dilation rate of that layer. Adherence to the HDC principle is maintained as long as $M_2 \leq K$, implying that the maximal interspace between two nonzero elements in the second layer does not exceed the convolutional kernel's dimension. For instance, a configuration of three sequential 3×3 dilated convolutional layers with dilation rates $r = [2, 3, 5]$ is considered compliant with the HDC principle, as the values calculated from Equation (3) satisfy the stipulated threshold of being less than or equal to 3.

$$M_2 = \max[M_3 - 2r_1, M_3 - 2(M_3 - r_2), r_2] = \max[5 - 2, 5 - 2 \times (5 - 3), 3] = 3 \tag{3}$$

By virtue of this design, the hybrid dilated convolution is endowed with the capability to effectively capture global information of lane markings with diverse configurations, without compromising pixel information.

2.2.2.3. Super-resolution upsampling. The characteristic features of diminutive lane markings endure significant attrition during the successive downsampling stages of the backbone network, necessitating a restorative mechanism to salvage these features, thereby enhancing the network's proficiency in delineating the features of minor lane markings. Super-resolution image reconstruction is predicated on the restoration of high-resolution imagery from low-resolution inputs, with a concerted effort to preserve maximal image detail. Sub-pixel convolutional layers have gained widespread adoption for their efficacy in elevating image resolution while retaining intricate image details, as delineated in Equation (4).

$$P_{SR} = f^L(P_{LR}) = PS(W_L * f^{L-1}(P_{LR}) + b_L) \tag{4}$$

P_{SR} symbolizes the resultant high-resolution image, P_{LR} denotes the low-resolution image input, and f_L constitutes the transformation function from input to output. $W_L * f^{L-1}(P_{LR}) + b_L$ represents the feature map at the Lth layer, derived from convolving the $(L-1)$th layer. PS signifies the periodic shuffling operator, which reconstitutes pixels from identical locations across $n \times n$ low-resolution images into an $n \times n$ feature map, subsequently serving as the corresponding segment within the super-resolution image. This

procedure is systematically applied to all pixels, culminating in the construction of the super-resolution image.

The feature map $F_{S4'}$, emanating from the hybrid dilated convolution, serves as the precursor for the sub-pixel convolutional layer, engendering a high-resolution feature map $F_{S4''}$ with dimensions twice the magnitude of the original. This high-resolution feature map is then amalgamated with the feature map $F_{S3''}$, which is procured from the 8-fold downsampling of the backbone network, through a channel-wise fusion process, thereby enriching the detail information.

The strategic implementation of the super-resolution upsampling layer markedly fortifies the model's capacity to articulate the features of lesser targets, such as peripheral lanes and distant lanes, thus empowering the model to more adeptly navigate a plethora of complex roadway scenarios and enhance segmentation precision.

2.2.3. *Convolutional Block Weighted Attention Module (CBWAM).* In semantic segmentation for lane detection, delineating the input image into discrete regions uncovers that certain sectors are inundated with extraneous elements such as the sky, lanes devoid of markings, roadside vegetation, and vehicle fronts. These segments are notably less pertinent compared to those that include lane demarcations. Furthermore, these areas may contain objects akin to lanes, such as white barriers and signage, which pose a risk of confounding the detection process. Informed by this understanding, the current study introduces the Convolutional Block Weighted Attention Module (CBWAM), depicted in FIGURE 6.

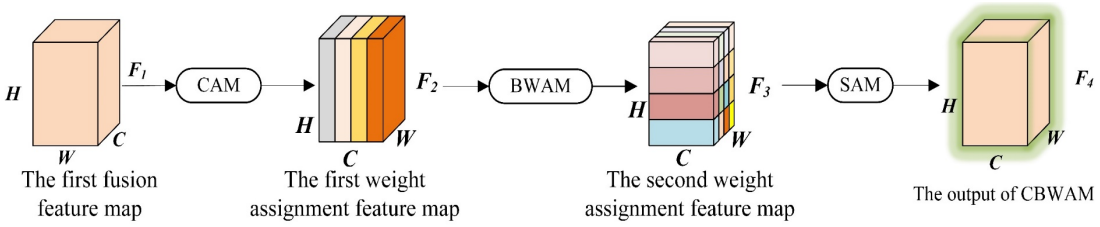

Figure 6. Convolutional Block Weighted Attention Module

The CBWAM encompasses three distinct components: the Channel Attention Module (CAM), the Block Weighted Attention Module (BWAM), and the Spatial Attention Module (SAM).CAM assigns weights to the feature channels of the initial fusion feature map, generating the first weight assignment feature map. BWAM applies block weight division to the first weight assignment map, yielding the second weight assignment feature map. Lastly, SAM performs pixel weight division on the second weight assignment feature map, producing the output of CBWAM.

2.2.3.1. Channel Attention Module(CAM). Within the realm of lane detection, the CAM strategically assigns varying weights to the features across different channels, thus amplifying the model's sensitivity to pixels pertaining to lanes while concurrently attenuating the prominence of background features. This enhancement is pivotal for bolstering model performance amidst complex and variable roadway conditions. FIGURE 7 illustrates the architecture of the CAM.

The initial fusion feature map, F_1, is concurrently subjected to maximum pooling and global average pooling. The aggregated outputs, denoted as M_1 and A_1 respectively, are then synergized and activated via a Sigmoid function to derive the channel-specific weights V_1. These weights are subsequently applied to the initial feature map to yield the first attention-modulated feature map, F_2, effectively orchestrating the allocation of

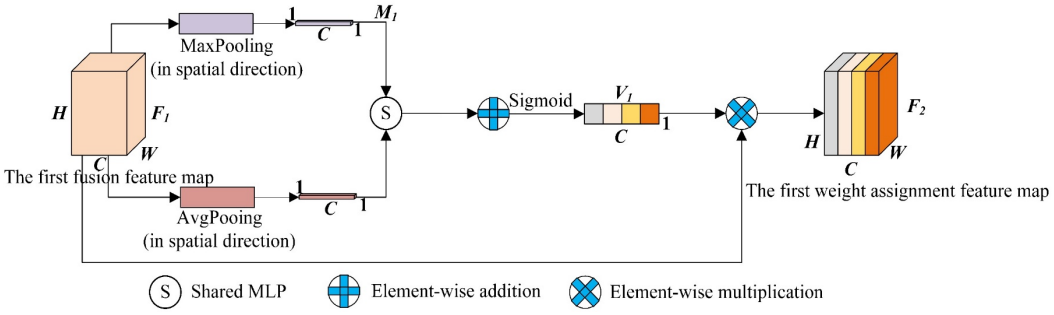

Figure 7. Channel Attention Module

attentional resources among channels. In this context, H, W, and C signify the height, width, and channel count of the feature map fed into the CBWAM.

Equations (5) and (6) represent the results of average pooling (M_1) and maximum pooling (A_1) performed along the spatial direction.

$$A_{1c} = H_{AP}(F_1) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{5}$$

$$M_1 = H_{MP}(F_1) = \sum_{i=1}^{H} \sum_{j=1}^{W} \max(x_c(i,j)) \tag{6}$$

Where H_{AP} and H_{MP} represent the functions of average pooling and maximum pooling, respectively. $x_c(i,j)$ represents the value of the Cth channel at the position (i,j). Equation (7) represents the final result F_2 obtained from the channel attention mechanism CAM.

$$F_2 = F_1 \otimes \delta(S(M_1) \oplus S(A_1)) \tag{7}$$

Where the sigmoid function is represented by δ, and S denotes the Shared MLP operation. The symbols $\oplus$ and $\otimes$ represent element-wise addition and element-wise multiplication, respectively.

2.2.3.2. Block Weighted Attention Module(BWAM). The task of lane detection necessitates recognizing that the salience of feature information is not uniform across channels nor across spatial dimensions. To adeptly mitigate the influence of non-essential elements on lane detection, the BWAM is contrived to partition the domain into multiple zones, endowing each with distinct weights to coarsely recalibrate spatial attention. FIGURE 8 delineates the BWAM's schematic representation.

Initially, BWAM conducts average pooling on the CAM's output, condensing it to a singular channel. This output is then segmented into 16 zones, each denoted as B_i ($i = 1, 2, 3, \ldots, 16$), and subjected to both maximum pooling and average pooling. The collective pooling results, M_2 and A_2, are subsequently amalgamated and processed via a series of three cascading 3×3 convolutional kernels, culminating in the derivation of regional weights V_2 via the Sigmoid function. These weights are then applied to the initial image to effectuate block-based weighting, producing the second attention-augmented feature map, F_3. This map is further refined by appending a Spatial Attention Module (SAM), which intricately adjusts the weights to generate the CBWAM-modulated output feature map.

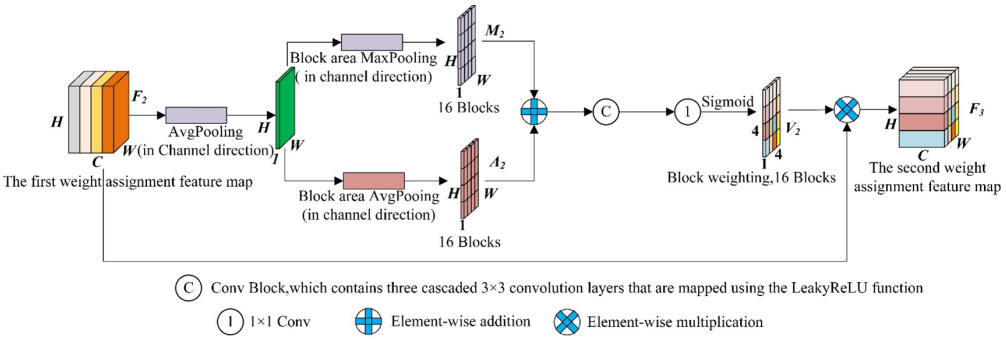

Figure 8. Block Weighted Attention Module

Equation (8) represents the result of average pooling after channel-wise blocking as M_2, while Equation (9) denotes the result of maximum pooling after channel-wise blocking as A_2.

$$A_{2,p} = \frac{1}{\frac{1}{4}H \times \frac{1}{4}W} \sum_{i=1}^{\frac{H}{4}} \sum_{j=1}^{\frac{W}{4}} x_a(i,j,p) \tag{8}$$

$$M_{2,pq} = \frac{1}{\frac{1}{4}H \times \frac{1}{4}W} \max_{i \in [1,\frac{H}{4}], j \in [1,\frac{W}{4}]} x_a(i,j,p) \tag{9}$$

Where $A_{2,pq}$ represents the pixel value of the pth block obtained after applying average pooling to the blocked region. The pixel value $x_a(i,j,p)$ corresponds to the pth block at the position (i,j) after average pooling, while $M_{2,pq}$ denotes the pixel value of the pth block obtained after applying maximum pooling to the blocked region.

2.2.3.3. Spatial Attention Module(SAM). The SAM is a granular spatial weighting system, meticulously designed to confer precise weights upon each pixel within the feature map, thereby elevating the fidelity of lane detection. As illustrated in FIGURE 9, the SAM complements the BWAM by executing nuanced adjustments to the spatial attention distribution across the entire image.

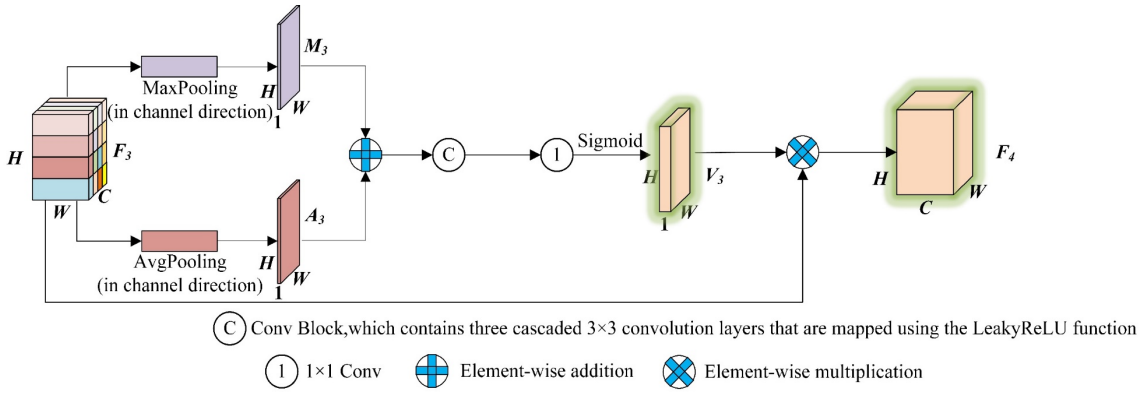

Figure 9. Spatial Attention Module

The SAM engenders two distinct feature maps, M_3 and A_3, via global average pooling and global max pooling, respectively. These maps are merged and subjected to feature extraction through a triad of cascading 3×3 convolutional kernels. Utilizing the Sigmoid function, pixel-specific weights V_3 are generated and integrated with the feature map emanating from BWAM, resulting in the final CBWAM-influenced output feature map, F_4.

BWAM is tasked with the macro-level allocation of spatial attention, while SAM refines this distribution at the pixel level, ensuring a judicious assignment of spatial attention resources.

Equation (10) represents the result of global average pooling along the channel direction, denoted as A_3, while Equation (11) represents the result of global maximum pooling along the same direction, denoted as M_3.

$$A_3 = H_{AP}(F_3) = \frac{1}{C}\sum_{i=1}^{C} x(i) \tag{10}$$

$$M_3 = H_{MP}(F_3) = \max x(i) \tag{11}$$

where C is the number of channels, $i \in [1, C]$ and $i \in \mathbb{N}^{\circ}$.

Equation (12) represents the output feature map, denoted as F_4, obtained after the application of CBWAM.

$$F_4 = F_3 \otimes \delta(Conv(S(M_3) \oplus S(A_3))) \tag{12}$$

The synergetic application of BWAM and SAM ensures a comprehensive and hierarchical distribution of spatial attention, from broader regions down to individual pixels. This collaborative mechanism ensures that the spatial attention within the final output feature map F_4 is both rational and efficient. Such a stratified approach to spatial attention allocation substantially amplifies the model's capability to discern lanes in intricate driving scenarios.

2.2.4. *Skip connection.* Accurate local feature detection is paramount, often dictating the efficacy of vehicular decision-making systems. Skip connection, a prevalent architectural feature within neural networks, forges direct informational conduits between disparate network layers, thereby facilitating the retention of critical local feature information. This mechanism enhances the model's proficiency in identifying salient lane attributes. FIGURE 10 delineates the structural schema of skip connection.

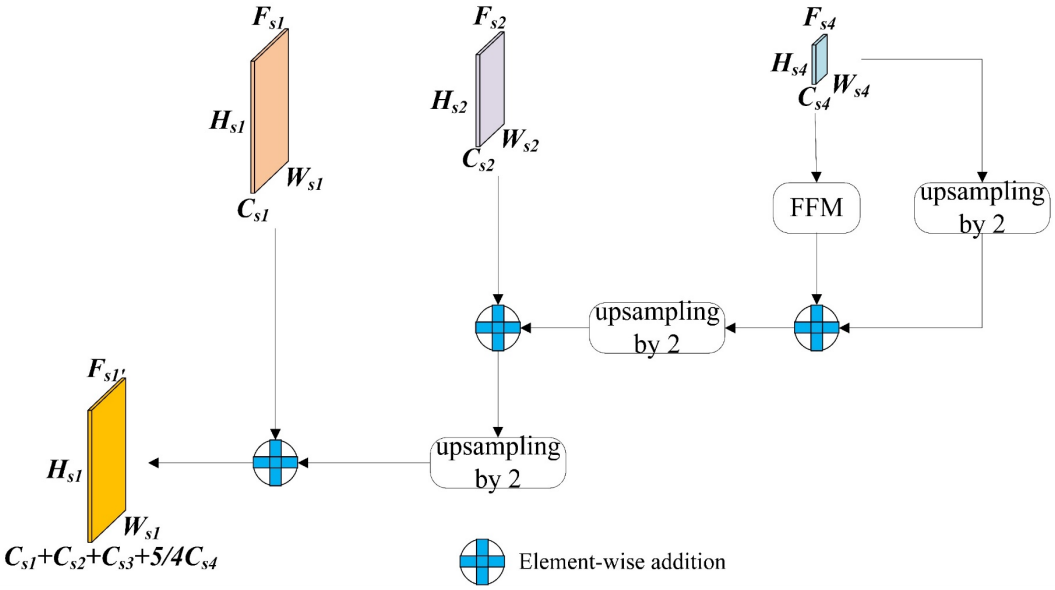

Figure 10. Schematic diagram of skip connection

In the context of the DeeplabV3+ architecture, the integration of deep and shallow feature information across the channel dimension demands a fourfold upsampling of the deep feature map. However, such an elevated upsampling factor risks obfuscating pixel-level detail. Furthermore, a rudimentary structural design interlinking the encoding and decoding layers can precipitate substantial loss of feature information during transference. Informed by these insights, the present study revisits the conceptual underpinnings of Fully Convolutional Networks (FCNs) to refine the DeeplabV3+ framework.

At the juncture between the encoding and decoding layers, skip connection are strategically deployed to incrementally amalgamate shallow features emanating from various convolutional strata of the primary network. This exploitation of shallow features, which encapsulate local details such as texture and hue associated with lane markings, serves to markedly elevate segmentation precision, particularly for diminutive targets such as peripheral and remote lanes.

The process is outlined as follows: initially, the encoded output bifurcates—one pathway feeds into the Feature Fusion Module (FFM) for super-resolution upsampling, while the other pathway is subjected to a two-stage upsampling. The outputs of both pathways are then fused and subsequently undergo another round of twofold upsampling. This fused map is then adjoined, in terms of the channel dimension, to the feature map derived from the octuple downsampling of the backbone network. The process of merging and upsampling is reiterated, doubling the scale of the result until the dimensionality of the output feature map is halved relative to the input image. The final feature map, a product of the skip connection, is channeled into the decoding layer, culminating in the pixel-wise segmentation of the image.

3. **Experimental results and analysis.**

3.1. **Data set selection and pre-processing.**

3.1.1. *Overview of the Baidu's unmanned vehicle dataset.* For the purposes of this research, the dataset utilized during the training phase was sourced from the Baidu's unmanned vehicle dataset, specifically from the semi-final stage dataset. This dataset, which was amassed from road segments within the metropolises of Shanghai and Beijing, is optimally conducive for the development of autonomous driving algorithms that are customized for the unique driving environment prevalent in mainland China. Beyond the inclusion of images depicting straight roadways under optimal lighting conditions, this dataset comprises a spectrum of complex driving scenarios, such as those involving glare, low light conditions, and inconsistent illumination patterns. FIGURE 11 presents a selection of images from the dataset, each paired with its respective annotation.

Additionally, the experimental protocol of this study extended beyond the use of the Baidu's unmanned vehicle dataset, incorporating the Tusimple dataset—a benchmark dataset in the field of lane detection. Collected from various segments of highways, the Tusimple dataset includes a variety of traffic scenarios encountered during different times of the day, under moderate weather conditions, and ranging from two-lane to multi-lane configurations. Each image within this dataset is annotated with precise lane markings, offering a comprehensive depiction of diverse traffic situations and roadway scenes.

The images displayed in FIGURE 11 illustrate a variety of challenging conditions for lane detection: (a) illustrates lanes in low light conditions; (b) shows lanes affected by glare; (c) captures lanes with significant curvature; (d) displays interference from road markings; (e) depicts lanes obscured by vehicles ahead; (f) includes erroneous lane marking disruptions; (g) features interference from pedestrian crossings; (h) shows lanes with signs of wear; and (i) details further disruption from road markings.

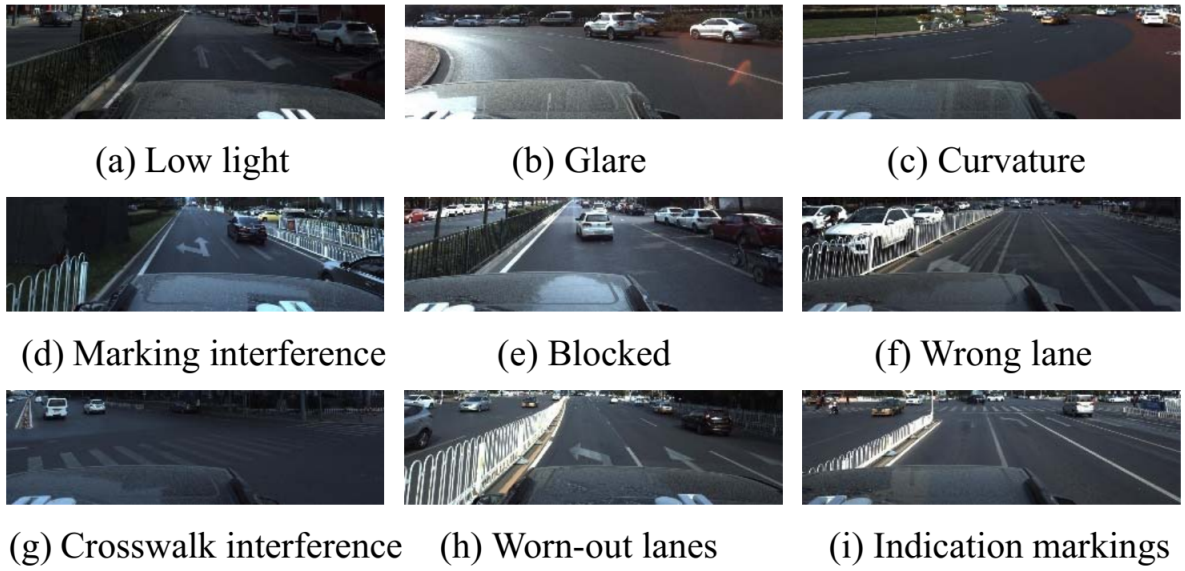

Figure 11. Lane images in various environments in Baidu's unmanned vehicle dataset

3.1.2. *Image pre-processing.* In the upper half of the image, about 1/3 of the area contains a large number of objects unrelated to lane detection, such as the sky, trees, etc., and to speed up the training process and take into account the sufficient field of view, the lower half of the image, about 2/3 of the area with a size of 3384×1020 pixels, is retained as the region of interest using a cropping method, and then it is scaled down to 1128×340 pixels to prevent memory overflow during training.

As one of the most widely used regularization methods, data augmentation can not only reduce the dependence of the model on feature points appearing at the same location with high frequency and avoid the interference of noise points by rotating the images at different angles, compressing them, and artificially adding noise but also expand the size of the dataset to prevent model overfitting.

In this study, before training, the number of lanes with large curvature in the dataset is widened by using horizontal flip to balance the sample types; in the training process, the images in each batch of the dataset are firstly cropped randomly, and the brightness, contrast, and saturation are randomly increased to maintain the training speed while simulating the light brightness changes during the driving process. A total of 11,608 images were included in the dataset before the training, and the lanes in different environments were counted, and the statistical results before and after the enhancement are shown in FIGURE 12.

3.2. **Experiment and Analysis.**

3.2.1. *Experimental environment configuration.* The model described in this paper utilizes the following hardware and software configuration: Ubuntu 20.04 operating system, Python 3.7 programming language, PyTorch 1.7.0 deep learning framework, AMD EPYC 7302 CPU, Nvidia RTX 3090 GPU with 24GB video memory, 63GB RAM, and 50GB of available hard disk space.

3.2.2. *Experimental procedure.* In this study, the weights were initialized using `torch.manual_seed(340`
The step decay strategy, in conjunction with the Adam optimizer, was used to determine the learning rate. The expanded dataset was divided into training, validation, and test

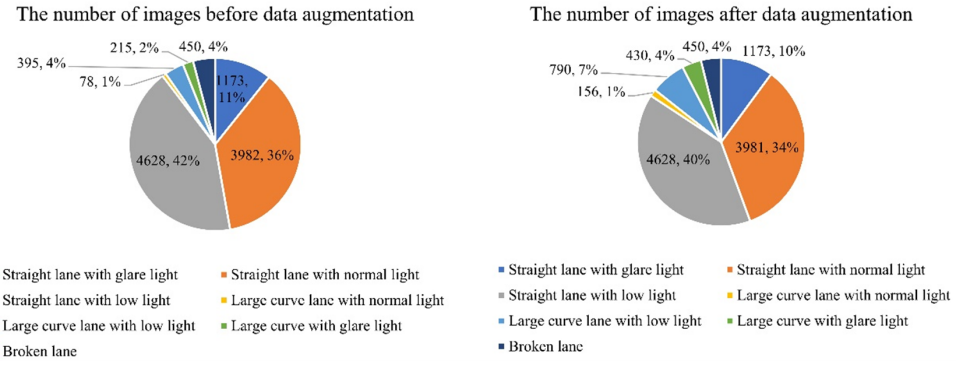

Figure 12. Statistical results of lane images in different environments before and after data augmentation

sets in an 8.5:1:0.5 ratio. The initial learning rate was set to 0.0005, with a minimum of 5e-6, and a batch size of 16.

Three different loss functions were employed in the experiments: the two-category cross-entropy loss function, the weighted two-category cross-entropy loss function, and the combination of the weighted two-category cross-entropy loss function and Dice Loss. The experiments were conducted over 100 rounds, with the training loss and validation loss values recorded during each round. Furthermore, model evaluation was performed every 5 rounds, saving the weights associated with the optimal evaluation metric $mIoU$. Finally, the loss curves and $mIoU$ curves were plotted as a function of the number of training rounds.

The experiments in this paper were conducted as follows: (i) Four different loss functions were individually applied to the modified DeeplabV3+ model to determine the optimal one for this dataset. (ii) Ablation experiments were conducted to assess the necessity of each designed module. (iii) A comparison was made between different deep learning models to evaluate the segmentation performance of the improved DeeplabV3+ model, considering real-time capabilities and lane detection accuracy.

3.2.3. *Evaluation indicators.* The main performance evaluation metrics for the experiments are the Mean Intersection over Union ($mIoU$), *Accuracy*, and mean pixel accuracy (mPA). One of the crucial metrics for evaluating the performance of the semantic segmentation model is $mIoU$. It is computed by determining the IoU of the predicted and labeled values for all image categories and subsequently calculating the average of these ratios.

In the binary classification problem studied in this paper, the average intersection over union ($mIoU$) is computed as the mean of the intersection over union ratios for the lane IoU_{lane} and the background $IoU_{background}$. The values of $mIoU$ are determined using Equations (13) and (14), respectively.

$$mIoU = \frac{1}{M}\sum_{i=1}^{M}\frac{TP}{TP+FN+FP} \tag{13}$$

$$IoU_{lane} = \frac{TP_{lane}}{TP_{lane}+FN_{lane}+FP_{lane}} \tag{14}$$

Equation (15) presents the formula employed to quantitatively measure the accuracy of the model's predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

Equation (16) presents the formulas that calculate the ratio of pixels with positive predictions in each category to the total number of pixels in that category. Additionally, they compute the average for each category.

$$mPA = \frac{1}{M} \sum_{i=1}^{M} \frac{TP}{FP + TP} \tag{16}$$

Where M is the number of categories, TP is the sum of the number of pixels predicted to be positive but actually positive, TN is the sum of the number of pixels predicted to be negative but actually negative, FP is the sum of the number of pixels predicted to be negative but actually positive, and FN is the sum of the number of pixels predicted to be positive but actually negative.

3.2.4. *Different loss functions.* In an effort to rectify the prevalent class imbalance between positive and negative samples in lane imagery, the present study introduces the application of Dice Loss. This loss function stands out for its enhanced efficacy in imbalanced sample contexts, prioritizing the extraction of feature information from the foreground during the training phase, which in turn significantly bolsters the model's proficiency in identifying lane features.

Within the scope of this work, four distinct loss functions were integrated into the refined DeeplabV3+ architecture: (1) Binary Cross Entropy (BCE) Loss; (2) a combination of BCE Loss and Dice Loss; (3) Weighted BCE (WBCE) Loss; and (4) a fusion of WBCE Loss and Dice Loss.

The *BCE Loss* measures the discrepancy between the true and predicted probability distributions and is calculated as shown in Equation (17):

$$BCE\ Loss = \frac{1}{N} \sum_{i} -(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \tag{17}$$

Here, y represents the label, and $\hat{y}$ represents the model prediction.

The *WBCE Loss* assigns different weights to each class based on Equation (17). This weighting scheme strengthens its contribution to the loss for classes with a small number of samples and reduces its contribution to the loss for classes with a large number of samples. The calculation formula is given in Equation (18).

$$WBCE\ Loss = \frac{1}{N} \sum_{i} -(\omega \cdot y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \tag{18}$$

The ω is the weight.

The *Dice Loss* is commonly employed as an ensemble similarity measure function to quantitatively assess the degree of similarity between two samples. Its calculation formula is depicted in Equation (19).

$$Dice\ Loss = 1 - \frac{2 \sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{19}$$

Here, the variables y_i and $\hat{y}_i$ represent the labeled and predicted values of pixel i, respectively. N denotes the total number of pixel points.

A series of experiments were conducted employing these varied loss functions, as delineated in TABLE 1.

Table 1. Experiments with different loss functions

Exp No.	**Loss**	$\omega = \omega_{lane} : \omega_{background}$	**mIoU (%)**	**Accuracy (%)**	**mPA (%)**
No.1	*BCE Loss*	\	86.95	99.41	91.40
No.2	*BCE Loss+Dice Loss*	\	87.47	99.41	93.86
No.3	*WBCE Loss*	2:1	87.86	99.44	93.19
No.4	*WBCE Loss+Dice Loss*	2:1	88.22	99.48	91.64

The data in TABLE 1 reveal that exclusive utilization of *BCE Loss* (Experiment No.1) yielded a *mIoU* of 86.95%. Incorporation of *Dice Loss* alongside *BCE Loss* (Experiment No.2) resulted in an augmented *mIoU* of 87.47%, representing an increment of 0.52 percentage points relative to Experiment No.1. Experiment No.3 further advanced the *mIoU* to 87.86% by employing *WBCE Loss* with a lane-to-background weight ratio of 2:1, thereby validating the efficacy of the weighting approach in mitigating class imbalance. The culmination of these efforts in Experiment No.4, which combined *WBCE Loss* with *Dice Loss*, led to an *mIoU* of 88.22%, surpassing Experiment No.1 by 1.27 percentage points, with *Accuracy* improving to 99.48%, and *mPA* attaining 91.64%. Such marked improvements are ascribed to the synergistic effect of *WBCE Loss* and *Dice Loss*: the former equilibrates the class disparity, while the latter ensures the model's predictions closely align with the actual labels. The confluence of these loss functions thus culminates in superior performance, conclusively demonstrating their collective superiority in countering the challenge of class imbalance. The findings accentuate the pronounced advantage of this methodology in refining the model's discrimination of lanes, especially within intricate traffic environments, and thereby enhancing the accuracy of lane detection.

FIGURE 13 illustrates the visual detection maps derived from models trained with the respective loss functions, demonstrating the effectiveness of each approach.

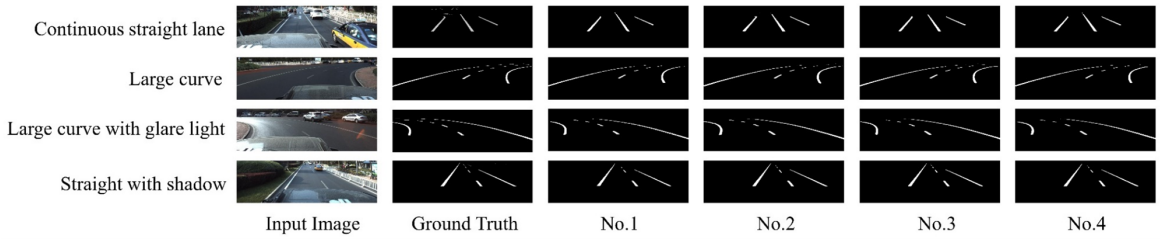

Figure 13. Visual Detection Maps for Different Loss Functions

Based on the findings from FIGURE 13, employing *WBCE Loss+Dice Loss* as the loss function enhances the model's capability to detect lanes, resulting in more comprehensive and detailed markings. It exhibits the ability to detect distal lanes more comprehensively in scenarios such as Continuous straight lanes and Straight with shadow, while also capturing smoother and fuller lanes in situations like Large curves and Large curves with glare light. In summary, the identified lanes exhibit greater completeness and richer details. Conversely, models trained using alternative loss functions exhibit certain shortcomings. This indicates that adopting *WBCE Loss+Dice Loss* as the loss function aids the model in addressing the issue of category imbalance and emphasizes the similarity between the predicted results and actual labels, thereby enhancing the performance of the lane detection model.

3.2.5. *Ablation experiments.* In the ablation study section, a quantitative analysis was undertaken to ascertain the individual contributions of the Multi-scale Feature Extraction Enhancement Module, the FFM, Skip connection, and the CBWAM to the overall performance of the model. The findings of this analysis are detailed in **TABLE 2**.

Table 2. Ablation experiments of different modules in the network structure

Experiment serial number	**Different modules in the network structure**					**mIoU(%)**
	Backbone	**Multi-scale feature extraction module**	**FFM**	**Skip Connection**	**CBWAM**	
No.1	✓					84.85
No.2	✓	✓				86.05
No.3	✓	✓	✓			85.71
No.4	✓	✓	✓	✓		87.10
No.5	✓	✓	✓	✓	✓	87.41
No.6 (Ours)	✓	✓	✓	✓	✓	88.22

TABLE 2 delineates the results from six distinct ablation experiments. The baseline model (Experiment No.1) achieved a $mIoU$ of 84.85%. With the integration of solely the Multi-scale Feature Extraction Enhancement Module (Experiment No.2), the $mIoU$ saw an increment to 86.05%, marking an improvement of 1.20%. The addition of only the CBWAM (Experiment No.3) realized an $mIoU$ of 85.71%, an uplift of 0.86%. The employment of Skip connection alone (Experiment No.4) culminated in an $mIoU$ of 87.10%, translating to an augmentation of 2.25%. These experiments distinctly demonstrate the enhancement provided by the Multi-scale Feature Extraction Enhancement Module, the CBWAM, and the Skip connection to the field of lane detection. Moreover, upon the addition of the Feature Fusion Module to the foundation established by Experiment No.4 (Experiment No.5), the $mIoU$ further rose to 87.41%, a supplementary increase of 0.31%, thereby evidencing the Feature Fusion Module's enhanced capability in capturing and refining the features of lanes, particularly their intricate details. In conclusion, the proposed model (Experiment No.6), which synthesizes all aforementioned modules, attained an $mIoU$ of 88.22%, translating to a significant enhancement of 3.37% compared to the baseline model. This improvement underscores the synergistic impact of the integrated modules. The comprehensive ablation study not only substantiates the effectiveness of each module but also furnishes critical insights for the architectural design of future lane detection models.

The visualization of the ablation experiment is detected in the graph shown in FIGURE 14.

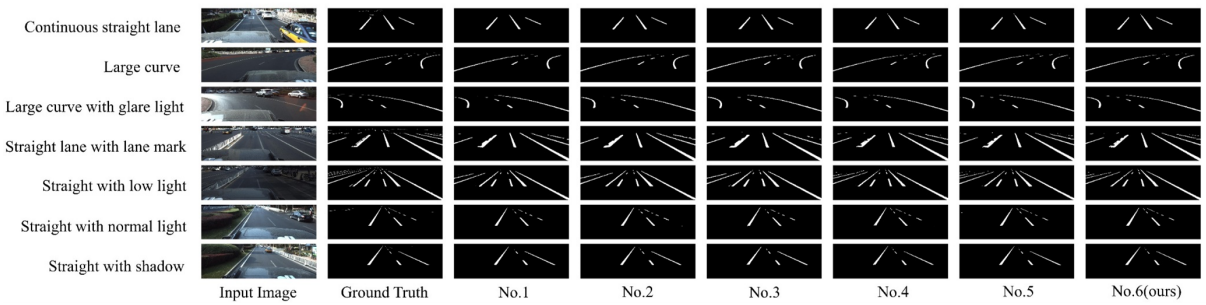

Figure 14. Visualization and detection map of the ablation experiment

By analyzing the visual detection graph and considering the previous data, it becomes evident that the model augmented solely with the multi-scale feature extraction module

(No.2) outperforms the baseline model (No.1) in detecting small targets like distal lanes in Large curves and Straight with low light scenes, thereby affirming the effectiveness of the multi-scale feature extraction module. Furthermore, the model equipped solely with CBWAM (No.3) exhibits improved discrimination of boundaries for similar objects such as white railings in Straight lanes with lane marks, highlighting CBWAM's effective allocation of attention resources and enhanced feature extraction capabilities. The inclusion of skip connection alone (No.4) results in fuller lanes compared to the baseline model, demonstrating the effectiveness of combining skip connection with shallow features for lane feature extraction. Additionally, the model combining skip connection and FFM (No.5) achieves superior overall detection performance across all scenarios when compared to No.4. Finally, our proposed model (No.6) outperforms models No.1-No.5, providing enhanced overall detection performance and more detailed information.

3.2.6. *Comparative experiments on the Baidu's unmanned vehicle dataset.* This paper presents comparison experiments, as shown in TABLE 3, to validate the superior segmentation capability and real-time performance of the proposed enhanced DeeplabV3+ model for lane segmentation. UNet, PSPNet, and the previous enhanced version of DeeplabV3+ are employed for comparative analysis. To ensure experimental fairness, all models are evaluated under identical conditions, utilizing the same input image size of 1128×340 pixels.

Table 3. Evaluation results of various image segmentation models on the Baidu's unmanned vehicle dataset.

Exp No.	**Model**	**mIoU(%)**	**Accuracy (%)**	**mPA(%)**	**Params(MB)**	**Single Image prediction time (ms)**
No.1	VGG16+UNet	89.56	99.54	93.51	95.0	47.30
No.2	MobilenetV2+PSPNet	72.26	98.47	81.03	9.3	20.41
No.3	MobilenetV2+DeeplabV3+	84.85	99.30	91.12	22.3	26.29
No.4	Ours	88.22	99.48	91.64	26.7	35.12

In the comparative evaluations on the Baidu's unmanned vehicle dataset, our proposed architecture (Experiment No.4) attained an $mIoU$ of 88.22%, an *Accuracy* of 99.48%, and an mPA of 91.64%, with the model footprint being a mere 26.7MB and the inference latency for a single image at 35.12 ms. In juxtaposition with the VGG16+UNet configuration (Experiment No.1), the latter realized a marginally superior $mIoU$ of 89.56%, yet was encumbered by a significantly larger size of 95.0MB and a longer prediction duration of 47.30 ms. Conversely, our model markedly diminishes both the complexity and the inference time while preserving a performance level that is on par. Relative to the MobilenetV2+PSPNet framework (Experiment No.2), our model exhibited a notable $mIoU$ enhancement of 15.96%, underscoring the capability of our design to retain a lightweight structure while elevating accuracy. These findings not only affirm the precision advantage of our model but also underscore its efficiency gains, which are particularly pivotal for applications necessitating real-time lane detection.

The visualized detection plots of different image segmentation models are shown in FIGURE 15.

From FIGURE 15, it can be observed that our proposed model (No.4) achieves a detection effect similar to the Ground Truth and a prediction effect comparable to the UNet model (No.1). The overall lane detection effect is good with clear edges. In comparison to the PSPNet model (No.2), our proposed model demonstrates superior prediction accuracy at the edges, exhibiting a significantly higher number of correctly detected lane

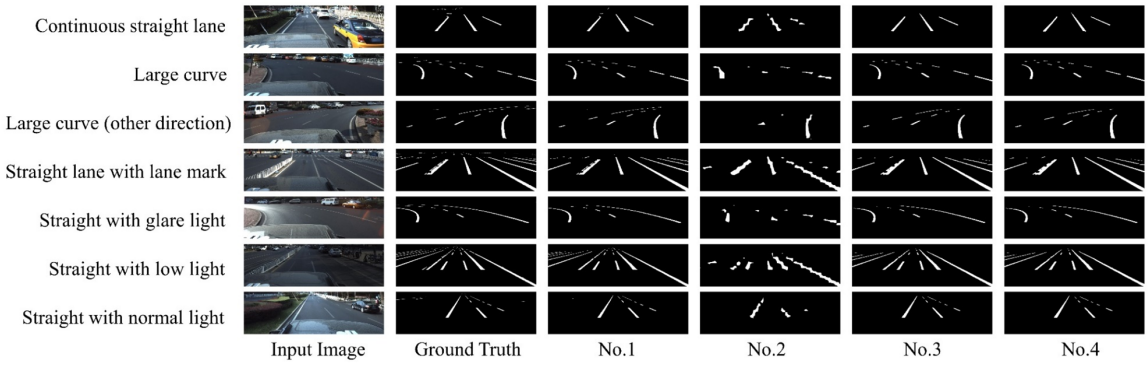

Figure 15. Visualization and detection plots of different segmentation models

pixels. Furthermore, when compared to the baseline model (No.3), our proposed model outperforms in both overall and detailed lane detection results.

3.2.7. *Evaluation of Classical Image Segmentation Algorithms on the TuSimple Dataset.* In this investigation, we conducted a comparative analysis of several semantic segmentation models on the TuSimple dataset, which included the well-established UNet, PSPNet, DeeplabV3+, as well as our novel proposed model. The experimental outcomes, delineated in TABLE 4, encapsulate the performance benchmarks of each model on the TuSimple dataset, which comprise metrics such as $mIoU$, mPA, *Accuracy*, and the F_1-score.

Table 4. Comparative Performance Metrics of Classical Image Segmentation Algorithms on the TuSimple Dataset

Exp No.	**Model**	$mIoU$ (%)	mPA (%)	**Accuracy** (%)	F_1**-score** (%)
No.1	VGG16+UNet	77.59	85.14	98.52	85.9
No.2	MobilenetV2+PSPNet	57.41	60.06	97.35	64.6
No.3	MobilenetV2+DeeplabV3+	75.0	81.64	98.37	83.8
No.4	Ours	**78.24**	**89.85**	98.41	**86.5**

The data unequivocally demonstrate that the proposed model outshines the other three experimental groups in aggregate, registering the most elevated scores in $mIoU$, mPA, and F_1-score. Moreover, the proposed model consistently excels over the VGG16+UNet model across all evaluative indicators. To elucidate, the proposed model (denoted as 'Ours') exhibits an increment of 0.65 percentage points in $mIoU$ and 4.71 percentage points in mPA over the VGG16+UNet model, and it is marginally lower in *Accuracy* by a negligible 0.11 percentage points. The F_1-score is particularly noteworthy, where the proposed model achieves 86.5%, outperforming the 85.9% secured by VGG16+UNet, which further corroborates the effectiveness of our approach.

In essence, the proposed model has demonstrated superlative performance on the TuSimple dataset, with marked improvements in critical metrics such as $mIoU$ and mPA, indicating a substantial enhancement in the segmentation of intricate structures relative to other models. These findings substantiate the sound design and robustness of the proposed model.

3.2.8. *Comparative Study of Classical Lane Detection Algorithms on the TuSimple Dataset.* Within the scope of this experiment, we evaluated a spectrum of state-of-the-art lane detection techniques and introduced our MobilenetV2-based model (referred to as 'Ours').

These methodologies incorporate a variety of technical stratagems and foundational networks, including but not limited to Eiglanes, SCNN, ENet-SAD, RESA, LaneAF, LSTR, FOLOLane, CLRNet, LaneATT, and CondLane. Eiglanes, CLRNet, and LaneATT are predicated on anchor-based methods; SCNN, ENet-SAD, RESA, and LaneAF leverage semantic segmentation; LSTR utilizes model-based approaches; FOLOLane relies on keypoint estimation; and CondLane employs a row-based detection strategy.

Among the methodologies under scrutiny, the proposed model manifests a conspicuous performance ascendancy, attaining a lofty *Accuracy* of 98.41% while sustaining a processing velocity of 28.4 frames per second. In comparison to other methodologies underpinned by ResNet and DLA architectures, such as CLRNet, which peaks at an *Accuracy* of 96.87%, and LaneATT, with an *Accuracy* of 96.83%, our proposed model secures an *Accuracy* augmentation of 1.54% and 1.57%, respectively. Against the SCNN method, another semantic segmentation paradigm, our proposed model records an *Accuracy* improvement of 1.88%. This underlines that the proposed model not only confers superior *Accuracy* but also significantly bolsters lane detection efficacy whilst upholding real-time processing capabilities. These outcomes accentuate the viability and applicability of our methodology in lane detection endeavors, especially in contexts that necessitate high *Accuracy* coupled with immediate responsiveness.

Table 5. Comparative Performance Metrics of Classical Lane Detection Algorithms on the TuSimple Dataset

Method	**Backbone**	**Accuracy(%)**	**FPS**
Eigenlanes [22]	-	95.62	-
SCNN [9]	VGG16	96.53	7.5
ENet-SAD [23]	-	96.64	75.0
RESA [24]	ResNet34	96.82	-
LaneAF [25]	DLA-34	95.62	-
LSTR [26]	ResNet18	96.18	420
FOLOLane [27]	ERFNet	96.92	-
CLRNet [28]	ResNet18	96.84	-
CLRNet [28]	ResNet34	96.87	-
CLRNet [28]	ResNet101	96.83	-
LaneATT [10]	ResNet122	96.10	26
LaneATT [10]	ResNet18	96.84	-
LaneATT [10]	ResNet34	96.87	-
CondLane [29]	ResNet18	95.48	220
CondLane [29]	ResNet34	95.37	154
CondLane [29]	ResNet101	96.54	58
Ours	MobileNetV2	**98.41**	28.4

4. **Conclusion.** The real-time lane detection methodology introduced in this study, predicated on an enhanced DeeplabV3+ framework, adeptly surmounts the segmentation intricacies associated with diminutive targets, including edge and distant lanes within vehicular imagery. This approach satisfies the exigencies of precision and immediacy required for lane segmentation in real-world applications. Initially, the MobilenetV2 architecture serves as the backbone network, distilling lane features across a hierarchy of levels. This is followed by the deployment of a Multi-scale Feature Extraction Enhancement Module, which seizes the contextual nuances of lanes across varied scales. Subsequently, FFM —

underpinned by a Residual Module and Hybrid Dilated Convolution — performs super-resolution upscaling on the deeply-layered features derived post a 16-fold downsampling, engendering feature maps replete with semantic richness and detailed granularity.

In an effort to capitalize on the a priori distributional knowledge of lanes, the CBWAM meticulously distributes attentional resources across both channel and spatial domains, thus bolstering the model's lane discernment capabilities and mitigating the attrition of feature information. The incorporation of Skip connection seamlessly amalgamates disparate levels of features, engendering a synergistic interplay between profound and superficial semantic information, which markedly amplifies the model's segmentation prowess. Empirical evidence corroborates that the model, while maintaining real-time efficacy, significantly augments the segmentation of diminutive targets, thereby contributing a novel research vista to the sphere of lane detection.

Notwithstanding the considerable strides made by the proposed model across various dimensions, it is not devoid of limitations. For example, the model's robustness in the face of severe meteorological conditions is slated for further amelioration. Ensuing research endeavors will be channeled towards enhancing the model's generalizability, thereby rendering it more adaptable to a diverse array of driving milieus. Moreover, the pursuit of more streamlined network structures, with the aim of bolstering the real-time aspect of lane detection, will constitute a principal objective of our future undertakings. It is our ambition to perpetuate the refinement and innovation of our techniques, thereby furnishing increasingly efficacious and precise lane detection solutions for the tangible implementation of autonomous vehicular technologies.

Acknowledgment. This work is supported by Zhejiang Sci-Tech University 2021 National University Students Innovation and Entrepreneurship Training Program, China (11120032662125). This work is also supported by the Key R&D Program of Zhejiang Province (2022C03136).

REFERENCES

[1] Y. U. Yim and S. Y. Oh, "Three-feature based automatic lane detection algorithm (TFALDA) for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 4, pp. 219-225, 2022.

[2] Z.-Q. Ying, G. Li, and G.-Z. Tan, "An illumination-robust approach for feature-based road detection," *IEEE International Symposium on Multimedia (ISM)*, pp. 278-281, 2015.

[3] H.-Y. Zhou and X. Song, "Lane Detection Algorithm Based on Haar Feature Based Coupled Cascade Classifier," *IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pp. 286-291, 2021.

[4] Q. Chen and H. Wang, "A real-time lane detection algorithm based on a hyperbola-pair model," in *IEEE Intelligent Vehicles Symposium*, IEEE, 2006, pp. 510-515.

[5] H.-R. Xu, X.-D. Wang, H.-W. Huang, K.-S. Wu, and Q. Fang, "A fast and stable lane detection method based on B-spline curve," *IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design*, pp. 1036-1040, 2009.

[6] W.-H. Li, F. Qu, Y. Wang, L. Wang, and Y.-H. Chen, "A robust lane detection method based on hyperbolic model," *Soft Computing*, vol. 23, pp. 9161-9174, 2019.

[7] Z. W. Kim, "Robust lane detection and tracking in challenging scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 16-26, 2008.

[8] J. Kim, J. Kim, G. J. Jang, and M. Lee, "Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection," *Neural Networks*, vol. 87, pp. 109-121, 2017.

[9] T.-M. Deng and Y.-J. Wu, "Simultaneous vehicle and lane detection via MobileNetV3 in car following scene," *PLoS One*, vol. 17, no. 3, pp. e0264551, 2022.

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 294-302, 2021.

[11] L.-C. Chen, Y.-K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801-818, 2018.

[12] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast LIDAR-based road detection using fully convolutional neural networks," *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1019-1024, 2017.

[13] D.-H. Lee and J.-L. Liu, "End-to-end deep learning of lane detection and path prediction for real-time autonomous driving," *Signal, Image and Video Processing*, vol. 17, no. 1, pp. 199-205, 2023.

[14] L.-C. Chen, X.-Z. Xu, L.-H. Pan, J.-F. Cao, and X.-M. Li, "Real-time lane detection model based on non bottleneck skip residual connections and attention pyramids," *PLoS One*, vol. 16, no. 10, pp. e0252755, 2021.

[15] X.-G. Pan, J.-P. Shi, P. Luo, X.-G. Wang, and X.-O. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[16] R.-Z. Ding, Z.-Y. Liu, T.-W. Chin, D. Marculescu, and R. D. Blanton, "FlightNNS: Lightweight quantized deep neural networks for fast and accurate inference," in *Proceedings of the 56th Annual Design Automation Conference*, pp. 1-6, 2019.

[17] X.-R. Jiang, N.-N. Wang, J.-W. Xin, X.-B. Xia, X. Yang, and X.-B. Gao, "Learning lightweight super-resolution networks with weight pruning," *Neural Networks*, vol. 144, pp. 21-32, 2021.

[18] S. Swaminathan, D. Garg, R. Kannan, and F. Andres, "Sparse low rank factorization for deep neural network compression," *Neurocomputing*, vol. 398, pp. 185-196, 2020.

[19] D. Yoon, J. Park, and D. Cho, "Lightweight deep CNN for natural image matting via similarity-preserving knowledge distillation," *IEEE Signal Processing Letters*, vol. 27, pp. 2139-2143, 2020.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.

[21] X.-Y. Zhang, X.-Y. Zhou, M.-X. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848-6856, 2018.

[22] M.-X. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, pp. 6105-6114, 2019.

[23] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self-attention distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1013-1021, 2019.

[24] D.-H. Lee and J.-L. Liu, "End-to-end deep learning of lane detection and path prediction for real-time autonomous driving," *Signal, Image and Video Processing*, vol. 17, no. 1, pp. 199-205, 2023.

[25] Z.-T. Yao and X.-Y. Chen, "Efficient lane detection technique based on lightweight attention deep neural network," *Journal of Advanced Transportation*, 2022.

[26] S. Song, W. Chen, Q.-J. Liu, H.-S. Hu, T.-C. Huang, and Q.-Y. Zhu, "A novel deep learning network for accurate lane detection in low-light environments," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 236, no. 2-3, pp. 424-438, 2022.

[27] H. Chang, D.-Y. Yeung, and Y.-M. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I-624, 2004.

[28] P.-Q. Wang, P.-F. Chen, Y. Yuan, D. Liu, Z.-H. Huang, X.-D. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451-1460, 2018.

[29] L. Liu, X. Chen, S. Zhu, and P. Tan, "Condlanenet: a top-to-down lane detection framework based on conditional convolution," In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3773-3782, 2021.