# THE LANE DETECTION ALGORITHM BASED ON MULTISCALE AGGREGATED ATTENTION FUSION NETWORK

Hong Wang[1], Yin Ang[1], Yilin Kang[1], Shasha Tian[1], Lu Zheng[1], Aifei Wang[1]
[1] College of Computer Science, South-Central Minzu University, Wuhan 430074, China, Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises, Wuhan 430074, China

NOTE: Corresponding author: Yilin Kang, ylkang@mail.scuec.edu.cn

***Abstract*** *– High accuracy and high stability are key elements of the lane detection algorithm in an autonomous driving system. Traditional algorithms are having difficulty extracting detailed features due to the complex geometric structure and background interference of lanes in real scenarios. Therefore, this paper proposes a Multiscale Aggregated Attention Fusion (MAAF) network, which integrates attention mechanisms to improve the accuracy and robustness of lane detection. Firstly, the Recurrent Feature-Shift Aggregator for Lane Detection (RESA) is improved to increase the effective sensory field and improve the efficiency of feature aggregation. Then, the ECANet attention module is used to extract features across channels, enhancing the model's focus on lane details. Finally, a spatial attention mechanism is incorporated to make the network more attentive to lane features, acquire more semantic information, and reduce the influence of background interference and clutter. Experimental results show that this method achieves 96.84% and 76.5% metrics on the TuSimple and CuLane datasets, respectively, surpassing the baseline network. Furthermore, it demonstrates good generalization and robustness, enabling accurate lane detection in complex road environments.*

**Keywords** – Attention mechanism, deep learning, feature aggregation, lane detection, mixed dense hole convolution.

## 1. INTRODUCTION

Lane detection is one of the key technologies in autonomous driving [1], attracting extensive attention from scholars at home and abroad, and its real-time performance and accuracy are crucial to the stability of autonomous driving systems. The development of lane detection is expected to drive the further development of autonomous driving technology. Advanced assisted driving systems typically utilize perceptual layer data such as lidar, high-accuracy positioning, and mapping to provide additional information, but lane detection is still an integral component.

Lane detection algorithms offer the advantages of real-time, high accuracy, and independence compared to mapping applications [2]. Lane detection algorithms can perform lane detection in real-time video streams, provide instant lane information, and can still accurately detect lane positions and shapes by analyzing images even when there is no network connection or map data available. In contrast, lane information for mapping applications is based on pre-stored map data and relies on the availability of external data, which may be limited by data accuracy and update frequency. However, lane detection algorithms can also be affected by the environment and the shape of the lane, and it is crucial to choose a suitable technical solution considering the specific application scenario and requirements.

Traditional lane line detection mainly relies on manual feature extraction and heuristic methods [3], which can be roughly divided into two methods based on road features and road models, respectively, to extract color features, texture features, or multifeature fusion of lane lines as a straight-line model and curve model [4]. However, its extracted features are limited and cannot be applied to complex scenarios, which is gradually replaced by deep learning-based methods. Therefore, they are gradually replaced by deep learning methods, which are usually based on pixel-by-pixel prediction methods and consider lane detection as a semantic segmentation problem [5-7]. These methods use an encoder-decoder framework, which first applies a Convolutional Neural Network (CNN) as an encoder to extract semantic information into the feature maps, and then uses an upsampling decoder to restore the feature maps to their original sizes and perform

pixel-by-pixel prediction. However, due to the elongated nature of lane lines and the insufficient number of labeled lane line pixels, these methods have difficulty in extracting subtle lane features and may ignore the shape ahead or there may be a high correlation between lanes, resulting in poor detection performance. For complex driving scenarios, the low-quality features extracted by conventional CNNs tend to lose subtle lane features. To solve this problem, some methods try to pass spatial information in the feature map but the operation of passing sequential information leads to a significant increase in the number of parameters and computation of the algorithm.

The lane detection task needs to consider the diversity in size and shape of lane markings. Additionally, lane markings are a small proportion of the foreground in the image, and there are various challenges such as lane markings being worn out, obscured, or affected by rain and snow. All these factors present significant challenges for achieving high accuracy in lane detection.

To extract more structural and spatial features of lane markings, this paper proposes a Multiscale Aggregated Attention Fusion network (MAAF) that incorporates the attention mechanism, which improves on RESA by expanding the feature fusion process by using a hybrid continuous null convolution to enlarge the receptive field, which moves cyclically in multiple directions of the feature map to achieve feature aggregation in both horizontal and vertical directions. Meanwhile, a Joint Attention (JATT) module is proposed to be connected in parallel with the improved RESA module to realize cross-channel information interaction while improving the algorithm's attention to lanes, and to reduce information loss while improving the fusion of global features. The F1-measure metric on the public dataset CuLane is 76.4%, which can accomplish the lane line detection task in complex environments, and reaches 30+fps on the RTX3090 graphics card, which meets the real-time requirements.

## 2. RELATED WORK

The methods for lane detection can be divided into traditional methods and deep learning-based methods, with the development of intelligent transport, lane detection methods based on radar and other high-precision devices also being proposed. Traditional methods mainly include algorithms such as edge detection [8-9], color segmentation [10-11], Hough transform [12], Kalman filtering [13], and curve fitting [14]. Li et al. [15] proposed an adaptive lane detection method using the Canny operator for edge detection, combined with the Hough transform and Otsu algorithm, to solve the problem of poor results of traditional algorithms at night. M Aly et al. [16] proposed an efficient method for lane detection, which uses selective directional Gaussian filtering, RANSAC, and other methods. This method can detect all lanes in still images under various conditions and achieve a running speed of 50Hz. Traditional methods are faster but due to the diversity of lanes and the complexity of the environment, their accuracy does not meet the requirements, and they are difficult to adapt to challenging real-world scenarios and cannot be deployed for use on vehicles.

Deep learning-based methods have better robustness and can adapt to more complex scenarios. Yu et al. [17] combined traditional methods with deep learning methods to decompose the inverse perspective mapping process into multiple successive microscopically-single-strain transform layers and refine the interpolated feature maps by subsequent convolutional layers, thus reducing artifacts and improving accuracy. Hou et al. [18] proposed a lightweight lane detection network (SAD) based on a self-attentive distillation module but the accuracy is low in complex scenarios. Pan et al. [19] proposed SCNN, which achieves better lanes' spatial structure feature extraction by slicing features and performing feature propagation across rows and columns. Zheng et al. [20] continued the idea of SCNN, optimizing the iterative manner of propagation across rows and columns, and proposed a new feature aggregation module RESA, which has a lower computation time and better performance. Qin et al. [21] proposed the UFSD algorithm, viewing the lane detection task as a classification problem on rows and using a larger receptive field, enabling it to handle complex scenarios and achieve a speed of 300+ fps. Feng et al. [22] proposed a lane detection method based on curve parameters, using a parameterized Bezier curve model for end-to-end detection and proposing a feature inversion fusion method based on deformable convolution, resulting in a model with small parameters and high detection accuracy.

PEAK et al. [23] proposed a lidar lane detection algorithm based on the K-Lane dataset, which showed good performance even under various lighting conditions and severe occlusions. BAI et al.

[24] proposed a deep neural network model combining lidar and camera sensors for direct output of lanes in 3D space to solve the problem of existing methods where the image is in the three-dimensional space with low accuracy. Lidar-based lane detection techniques perform lane detection by reflectance thresholding or by clustering, which is too dependent on fixed parameters and is only suitable for fixed scenarios. And it is limited by the small dataset, and the development of the technology is slow.

## 3. LANE DETECTION NETWORK FRAMEWORK

The lane detection task is a task highly dependent on the surrounding environment clues. Even if a lane is obscured, it is still possible to infer using other lanes or the positions of other vehicles. To improve the ability of the lane detection network to extract effective information in complex scenes, this paper proposes a multiscale feature information aggregation network with a fusion attention mechanism by focusing on the feature information of different receptive fields. The network structure is shown in Fig. 1. The network consists of an encoding network, a feature information aggregation network, and a decoding network. The encoding network uses the ResNet [25] network to perform preliminary feature extraction on the input image. The feature information aggregation network (MAAF) not only aggregates feature information from different receptive fields in four directions, but also enables feature information interaction across channels, focusing more on the position of the lanes and reducing the interference from the background area during the detection process. In the decoding network, the feature map is restored to the original size through upsampling, and then the lanes are predicted using convolution operations.

This paper adds a global average pooling branch before decoding in the network to fully utilize the high-level semantic information of lane existence. This branch takes the output of MAAF as input to determine whether the lanes exist. The global average pooling layer has a smaller computational complexity and parameter quantity than the fully connected layer. By averaging the points on the feature map, it can capture useful details and eliminate useless noise, making the network focus more on the positions where the lanes exist and less on the positions where the lanes don't exist. By introducing the high-level semantic information of lane existence, the lanes can be better located, and the accuracy and robustness of lane detection can be improved. In addition, the global average pooling layer can also reduce the risk of overfitting and the number of parameters and complexity of the model, thereby improving the optimization and generalization ability of the model.

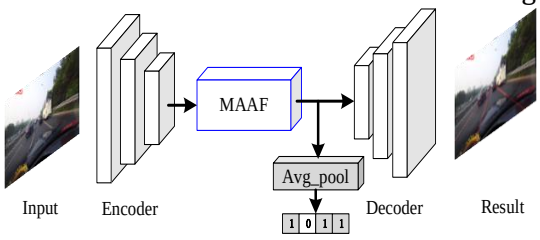



**Fig. 1** – Overall structure of the network

### 3.1 Multiscale feature information aggregation network with attention mechanism fusion

Traditional convolutional semantic segmentation networks based on traditional standards are no longer applicable to long and thin targets like lanes. Currently, some methods use information propagation or self-attention mechanisms to enhance the transmission of feature information between lanes, thereby improving the accuracy of lane detection. Inspired by the REcurrent Feature-Shift Aggregator (RESA) and Global Attention Mechanism (GAM) [26] attention mechanisms, this paper designs a multiscale feature information aggregation network with a fusion attention mechanism, effectively improving the extraction of lanes' detail features and reducing background interference. The structure is shown in Fig. 2. The backbone network extracts the feature map $X \in R^{C\times H\times W}$ of the original image at a size of 1/8. First, feature map X is simultaneously inputted into the channel attention module (ECANet) [27] and the sliced feature fusion module (IMRESA), enhancing the features in the channels and spatial dimensions, respectively. The channel attention ECANet captures lane information between different channels, while IMRESA aggregates lane features in the spatial dimensions. Finally, spatial attention highlights the key positions of the lanes, making the model pay more attention to the target regions in the image that play a decisive role in segmentation while ignoring the interference from background regions.

### 3.2 Multiscale slice feature aggregation module

The traditional convolution method performs

convolution calculations in the channel direction and accumulates the results. However, this approach has greatly limited receptive fields. IMRESA, on the other hand, uses mixed dilated convolutions in the width and height dimensions. This module uses one 3x3 regular convolution (with a dilation rate of 1) and three dilated convolutions with dilation rates of 2, 5, and 9. Compared to traditional convolutions, dilated convolutions add zeros to the convolution kernels to increase the distance between them, thus enlarging the receptive field of each kernel. After K iterations, the feature maps not only receive feature information from each location but also aggregate information from different receptive fields.

The feature map $X \in R^{C\times H\times W}$ represents the features extracted by the encoding network, where C, H, and W represent the number of channels, number of rows, and number of columns of the feature map X, respectively. The IMRESA branch first slices the feature map horizontally and vertically into H and W slices, and $X^k_{c,i,j}$ represents the feature map at the kth iteration, with indices c, i, and j indicating the channel, row, and column indexes, respectively. F is the continuous dilated convolution, k is the iteration count, ReLU is the non-linear activation function, α is the weight for feature information aggregation, and $Z^k_{c,i,j}$ represents the feature map element after the mixed continuous dilated convolution. $X^{k'}_{c,i,j}$ represents the updated feature map element. The calculation formula for the entire process is as follows:

$$Z^k_{c,i,j} = \sum_{m,n} F_{m,c,n} \cdot X^k_{m,(i+s_k)\ mod\ H, j+n-1} \quad (1)$$

$$Z^k_{c,i,j} = \sum_{m,n} F_{m,c,n} \cdot X^k_{m,i+n+1,(j+s_k)\ mod\ W} \quad (2)$$

$$X^{k'}_{c,i,j} = X^k_{c,i,j} + \alpha ReLU\left(Z^k_{c,i,j}\right) \quad (3)$$

$$s_k = 2^k, \quad k = 0,1,\dots,K-1 \quad (4)$$

K represents the total number of iterations. When iterating horizontally, $K = \lfloor log_2 W \rfloor$, and when iterating vertically, $K = \lfloor log_2 H \rfloor$. $s_k$ represents the step size of the k-th iteration. "D", "U", "L", and "R" represent the four directions: down, up, left, and right, respectively. After K iterations of feature aggregation in four different directions and scales, each pixel of the original feature map X will receive feature information from different positions and receptive fields of other pixels.

The actual process of feature aggregation includes four directions: down, up, left, and right. Fig. 3 takes the example of aggregating from left to right (IMRESA_R) to explain how each slice can obtain information from the entire feature map after K iterations. First, the feature map is divided into W slices vertically, denoted as $\{X_0, X_1, X_2, \dots, X_{w-2}, X_{w-1}\}$. The iteration count is 4. In the first iteration, with k = 0, the step size $s_1$ is set to 1. $X_1$ receives feature information from $X_0$, $X_{w-i}$ receives feature information from$X_{w-i-1}$, and $X_0$ receives feature information from $X_{w-1}$. In the second iteration, with k = 1, the step size $s_2$ is set to 2. $X_2$ receives feature information from $X_0$, $X_{w-i}$ receives feature information from $X_{w-i-2}$, and similarly, $X_0$ receives feature information from $X_{w-2}$. After two iterations, $X_2$ has received information from $X_1$ and $X_0$, while $X_0$ has already received information from $X_{w-1}$ in the first iteration. Therefore, at this point, $X_2$ has already received feature information from $X_1$, $X_0$, and $X_{w-1}$. Finally, after K iterations, each $X_i$ will receive feature information from the entire feature map, as shown in Table 1.

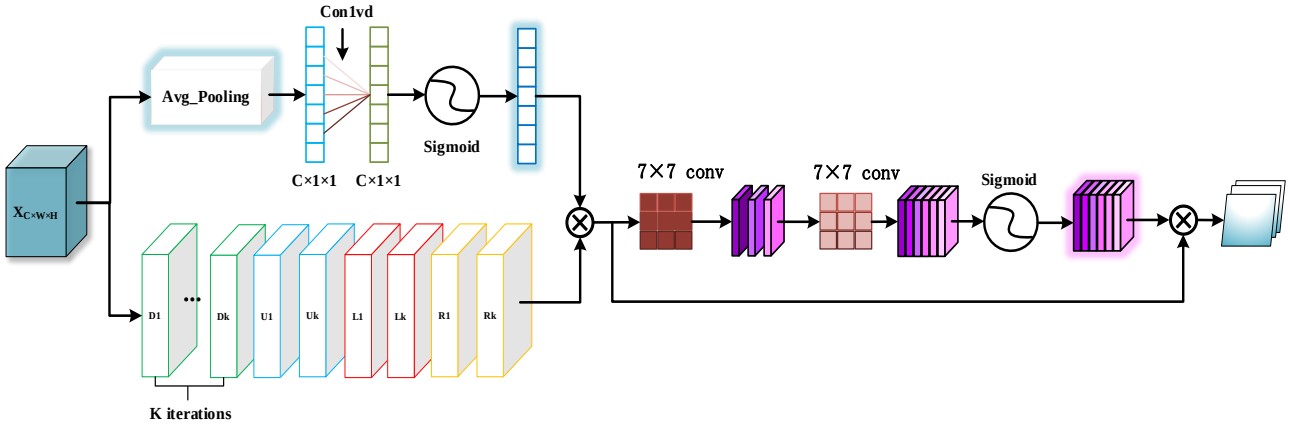



**Fig. 2** – Feature aggregation network structure

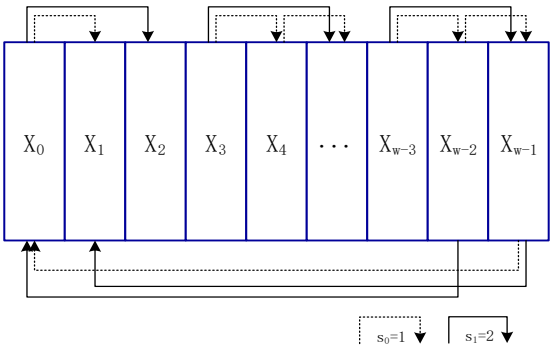



**Fig. 3** – IMRESA_R iteration process

**Table 1** – Feature information obtained by $X_7$ after k iterations

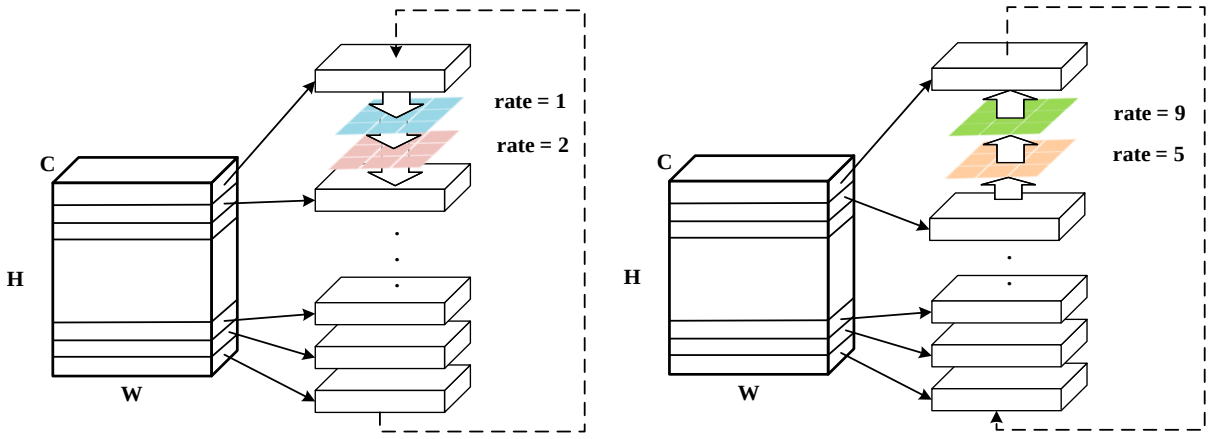
| $k$ | $s_k$ | Features received by $X_7$ |
|---|---|---|
| 0 | $s_0$=1 | $X_6$ |
| 1 | $s_1$=2 | $X_4, X_5, X_6$ |
| 2 | $s_2$=4 | $X_0, X_1, X_2, X_3, X_4, X_5, X_6$ |

## 3.3 Multidirectional mixed continuous cavity convolution

Compared with traditional convolution, dilated convolution fills in zeros in the convolution kernel of the original convolution to increase the distance between convolution kernels. This enlarges the receptive field of each convolution kernel, improves the network's receptive field, allows the network to learn more global information, better distinguishes the relationships between each pixel and feature, and improves model accuracy. However, dilated convolution has the drawback of kernel discontinuity, which means not all points in the feature map will participate in the convolution calculation, resulting in the risk of losing feature information, especially for objects with thin and continuous structural features like lanes. Hybrid Dilated Convolution (HDC) [28] can effectively solve this problem.

In this study, multiple experiments were conducted on the convolution method and different dilation rates. Hybrid dilated convolution with dilation rates of 1 and 2 was used for top-down aggregation, while dilation rates of 5 and 9 were used for bottom-up aggregation. Similarly, hybrid dilated convolution with dilation rates of 1 and 2 was used for left-right aggregation, and dilation rates of 5 and 9 were used for right-left aggregation. Performing hybrid dilated convolution in all four directions not only provides a larger receptive field but also ensures that the features at each position can participate in the calculation during the iteration process, without losing the structural feature information of lanes.

In each iteration, IMRESA divides the feature maps into H and W slices in both horizontal and vertical directions, and then performs feature aggregation in the "D", "U", "L", and "R" directions. Taking feature aggregation in the "IMRESA_D" and "IMRESA_U" directions as an example, as shown in Fig. 4, the feature maps are divided into H slices represented as $\{H_0, H_1, H_2, \ldots, H_{h-2}, H_{h-1}\}$. First, aggregation is performed from top to bottom, where feature slice $H_{h-1}$ undergoes continuous dilated convolutions with dilation rates of 1 and 2, and then the result is feature fused with feature slice $H_{h-2}$. Then, aggregation is performed from bottom to top, where feature slice $H_{h-3}$ undergoes continuous dilated convolutions with dilation rates of 5 and 9, and then the result is feature fused with feature slice $H_{h-2}$. Similarly, the same mixed continuous dilated convolutions are used for feature fusion in the two vertical iterative directions.




**(a) IMRESA_U** **(b) IMRESA_D**

**Fig. 4** – Multidirectional mixed continuous hole convolution

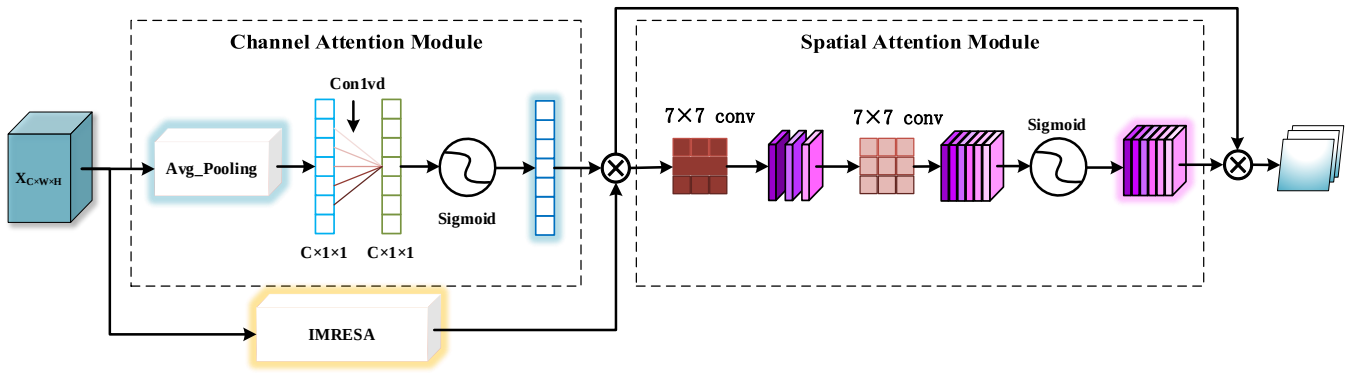



**Fig. 5** – Joint attention module

### 3.4 Joint Attention Module (JATT)

In the image, the lane is the foreground, with a small proportion of pixel points, so the background has a significant impact on the recognition of lanes. To improve the accuracy of network recognition, this paper improves the GAM attention module and proposes a new joint attention module JATT, which consists of channel attention and spatial attention in series. JATT first uses channel attention to interact with channel-wise information in the feature map and then uses spatial attention to enhance the focus on the lanes area.

The JATT module proposed in this paper is composed of ECANet channel attention and spatial attention modules in series, as shown in Fig. 5. In the ECANet channel attention module, the input feature map $X \in R^{C\times H\times W}$ is first average-pooled, the feature map is converted from a $C \times H \times W$ matrix to a $C \times 1 \times 1$ vector. Then, an adaptive one-dimensional convolution kernel size is calculated.

ECANet channel attention achieves information interaction between feature maps across channels by using one-dimensional convolution. The model has a relatively small complexity. The input feature map is represented as $X \in R^{C\times H\times W}$. First, the computed convolution kernel size is applied to the one-dimensional convolution. This allows layers with larger channel numbers to have more thorough cross-channel interaction, thereby obtaining channel weights for each feature map.

Finally, the normalized weights are multiplied channel-wise with the feature map obtained through IMRESA aggregation, resulting in a weighted feature map. The calculation formula for this process is shown in equation (5):

$$M_c(X) = \sigma(F_{1,k}(avg(X))) \qquad (5)$$

Where $M_c(X)$ represents the feature map generated by ECANet attention, $F_{1,k}$ represents a one-dimensional convolution, and $avg$ represents global average pooling. The calculation formula for the one-dimensional convolutional kernel of size k is as shown in equation (6):

$$k = \left|\frac{log_2 C}{2} + 1\right| \qquad (6)$$

To make the model pay more attention to lane features, the feature map $X_1$ generated from channel attention is used as input for spatial attention, which mainly focuses on the positional information of lanes within the image. It selectively enhances each spatial feature through weighted selection. $X_1$ is the input feature. Using a 7×7 convolution kernel, the number of channels in the feature map is reduced, and then it is restored using another 7×7 convolution kernel. Finally, the resulting feature map is normalized to obtain $M_s(X_1)$. $X_1$ and $M_s(X_1)$ are then multiplied together. This process can be represented by equation (7):

$$M_s(X_1) = \sigma\{F_{7,7}[F_{7,7}(X_1)]\} \qquad (7)$$

In the equation, $\sigma$ represents the $Sigmoid$ function, and $F_{7,7}$ corresponds to a $7 \times 7$ convolution operation.

By combining the attention mechanism with the slice feature aggregation module, the model can effectively enhance the fusion of global features and increase its focus on lane details, while reducing interference from background and other information.

## 4. EXPERIMENT

To validate the performance of the algorithm proposed in this paper, experiments were conducted on the TuSimple and CuLane public datasets. Detailed information on both datasets is provided in Table 2.

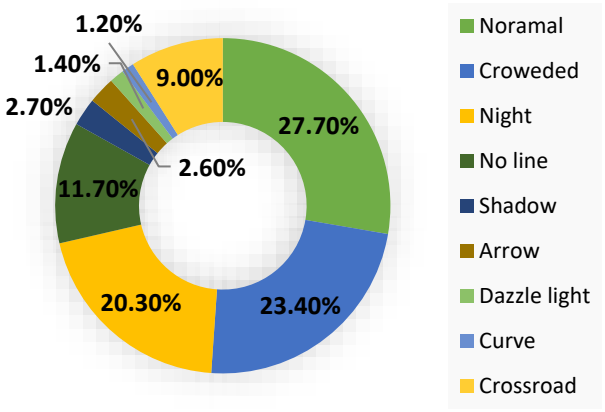



**Fig. 6** – Proportions of different classes in the CuLane dataset

## 4.1 The TuSimple dataset

The TuSimple dataset is an autonomous driving dataset for real highway scenarios, with a total size of approximately 20G. The dataset includes lanes from highways in good weather or moderate weather conditions, with relatively simple scenes. There are 3626 images used for training and 2782 images used for testing. The evaluation metrics include False Positive (FP), False Negative (FN), and accuracy. The calculation formula for accuracy is given by equation (8):

$$accurary = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}} \quad (8)$$

The $S_{clip}$ represents the total number of lane points, and the $C_{clip}$ represents the number of correctly predicted lane points.

## 4.2 The CuLane dataset

The CuLane dataset was proposed by the SCNN paper and captured using six different vehicles in different lane environments in Beijing. The total duration of the dataset exceeds 55 hours, with a size of approximately 40 GB. It includes various types of lanes such as urban, rural, and highways, and the lane markings are divided into nine categories: normal, crowded, night, no line, shadow, arrow, dazzle light, curve, and crossroad. The ratios for each category are shown in Fig. 6. The dataset consists of 88,880 images for training, 9,675 images for validation, and 34,680 images for testing.

In the CuLane dataset, each lane is considered as a line with a width of 30 pixels. The accuracy of predictions is assessed using the Intersection over Union (IoU) metric, with a threshold of 0.5. The main evaluation metrics in the experiment are F1-measure, precision, and recall. Precision is calculated using equation (9):

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

The formula for calculating recall rate is equation (10):

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

The calculation formula for F1-measure is expressed as equation (11):

$$F1 = \frac{2\times precision\times recall}{precision+recall} \quad (11)$$

TP stands for true positive, FP stands for false positive, and FN stands for false negative.

## 4.3 Experimental setup

This experiment uses the Ubuntu 18.04 operating system. The GPU used for training is an RTX 3090 and the deep learning framework used is PyTorch 1.8.1. Before experimenting, relevant hyperparameters need to be set. The SGD optimizer is used with a momentum of 0.9 and a weight decay coefficient of 1e-4. For the TuSimple dataset, the learning rate is set to 0.02 with a batch size of 4. For the CuLane dataset, the learning rate is also set to 0.02 but with a batch size of 8. Before training, image preprocessing is performed, and the image size is adjusted to 288 pixels × 800 pixels for the TuSimple dataset and 368 pixels × 640 pixels for the CuLane dataset.

## 4.4 Experimental result analysis

This article conducts experiments on two public lane datasets and compares them with other popular lane detection methods such as SCNN, SAD, LaneNet [27], Enet [28], UFSA, RESA18, RESA-34, and RESA-50.

**Table 2** – Description of TuSimple and CuLane datasets

| Dataset | Resolution | Total | Training | Testing | Complexity | Lines |
|---|---|---|---|---|---|---|
| TuSimple | 1280×720 | 6408 | 3626 | 2782 | simple | ≤5 |
| CuLane | 1649×590 | 133235 | 98555 | 34680 | complex | ≤4 |

The use of three backbone networks, ResNet18, ResNet34, and ResNet50, with the identifiers -18, -34, and -50 respectively, are considered. Since the TuSimple dataset has simpler scenarios, the models chosen are also simpler, whereas on the CuLane dataset some models are added that work better for complex scenarios. The experimental results on the TuSimple dataset are shown in Table 3. When the ResNet34 backbone network is selected, the accuracy of this method reaches 96.84%, surpassing the highest RESA-34. In addition, the values of FP and FN are compared on the TuSimple dataset, and the FP value of this method is significantly reduced. This indicates that this method can avoid misjudgment of non-lane as lanes, thus achieving higher accuracy in lane detection tasks. The experimental results on the CuLane dataset are shown in Table 4, where FP is used as the evaluation metric in the crossroad scenario, and the F1-measure is used in other scenarios. This method obtains the best performance in all scenarios. Compared with RESA-50, the improvement in the shadow scenario reaches 4.4%, the overall accuracy improves by 1.0%, and MAAF-50 can achieve 35fps on the RTX3090 graphics card. These results demonstrate the effectiveness and real-time performance of this method.

To demonstrate the effectiveness of our method more intuitively, Fig. 7 shows the qualitative experimental results of MAAF-50 on the CuLane dataset. The first column shows the lane detection results obtained by RESA-50, the second column shows the lane detection results obtained by MAAF-50, and the third column shows the ground truth annotation images for the lanes. Note that the crossroad's images were not originally labeled with ground-truth. From the images, the proposed multiscale feature information aggregation network with the fused attention mechanism in this paper can correctly detect lanes in different scenarios and lane shapes, such as crowded, night, shadow, and no line, which proves that the algorithm proposed in this paper has good robustness.

## 4.5 Ablation experiment

To verify the reasonableness and effectiveness of the proposed improvements, this paper uses ablation experiments on the CuLane dataset using the RESA-50 model as the baseline. First, each module proposed in this paper was gradually added, and the effectiveness of each part was verified. The detailed results of the experiments are listed in Table 5, where JATT represents the joint attention mechanism, Conv represents the dilation rate of 1,2, 5, and 9 for the dilation convolution, and Avg_pool represents the global average pooling. The experimental results show that the addition of each

**Table 3** – Comparison results of different algorithms on the TuSimple dataset

| Network | Accuracy (%) | FP | FN |
|---|---|---|---|
| **SCNN** | 96.53 | 0.0617 | **0.0180** |
| **LaneNet** | 93.38 | 0.0780 | 0.0224 |
| **ENet** | 93.02 | 0.0886 | 0.0734 |
| **RESA-18** | 96.70 | 0.0395 | 0.0283 |
| **RESA-34** | 96.82 | 0.0363 | 0.0248 |
| **MAAF-18** | 96.79 | 0.0270 | 0.0291 |
| **MAAF-34** | **96.84** | **0.0253** | 0.0265 |

**Table 4** – Comparison results of various algorithms on the CuLane dataset

| Category | SCNN | SAD | ENet | UFSA | RESA-34 | MAAF-34 | RESA-50 | MAAF-50 |
|---|---|---|---|---|---|---|---|---|
| Normal(%) | 90.6 | 89.9 | 88.4 | 90.7 | 91.9 | 92.7 | 92.1 | **92.8** |
| Crowded(%) | 69.7 | 68.4 | 67.0 | 70.2 | 72.4 | 73.6 | 73.1 | **74.4** |
| Night(%) | 66.1 | 64.3 | 61.4 | 66.7 | 69.8 | 70.9 | 69.9 | **70.8** |
| No line(%) | 43.4 | 42.1 | 42.9 | 44.4 | 46.3 | 48.7 | 47.7 | **48.8** |
| Shadow(%) | 66.9 | 67.8 | 63.4 | 69.3 | 72.0 | 74.0 | 72.8 | **77.2** |
| Arrow(%) | 84.1 | 83.1 | 81.9 | 85.7 | 88.1 | 88.8 | 88.3 | **88.9** |
| Dazzle light(%) | 58.5 | 59.7 | 57.4 | 59.5 | 66.5 | 67.8 | 69.2 | **70.6** |
| Curve(%) | 64.4 | 66.2 | 62.6 | 69.5 | 68.6 | 70.3 | 70.3 | **72.3** |
| Crossroad(FP) | 1990 | 1982 | 2768 | 2037 | 1896 | 1610 | 1503 | **1364** |
| Total(%) | 71.6 | 70.6 | 68.8 | 72.3 | 74.5 | 76.0 | 75.3 | **76.3** |

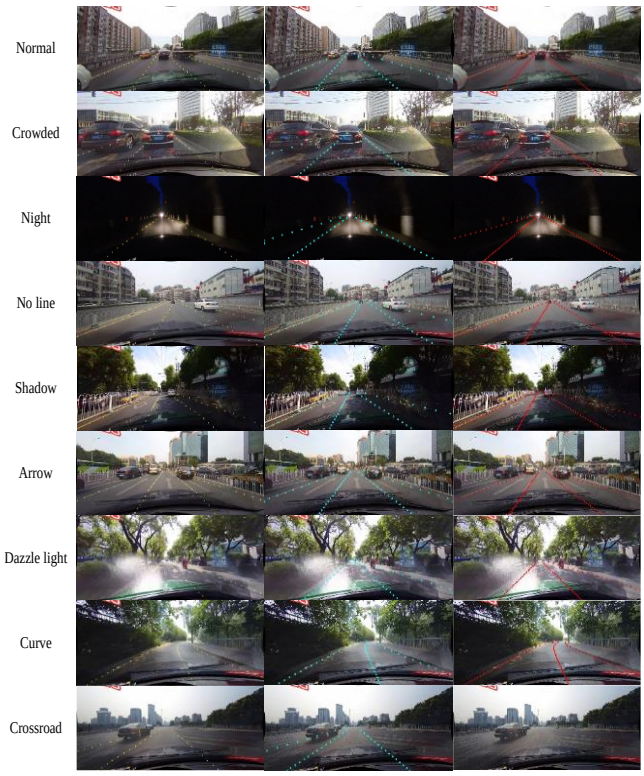



**Fig.** 7 – Detected results of the algorithm in the CuLane dataset in this article

part in RESA improves the F1-measure, especially for the scenarios of no line, shadow, and curve, where the improvement is particularly significant. This indicates that the proposed method in this paper can better extract lane features and effectively utilize the thin and continuous structural features of lines.

To verify the improved rationality proposed in this article, the CBAM attention module, the GAM attention module, and the improved joint attention module were selected for experimental comparison, shown in Table 6. The experiments were all based on the RESA-50 baseline framework. RESA-GAM represents the combination of the GAM attention module with RESA-50, RESA-CBAM represents the combination of the CBAM attention module with RESA-50, and RESA-JATT represents the use of the proposed joint spatial attention module in this article.

This study also conducted experiments using different groups of dilation rates to verify the effectiveness of mixed dilated convolutions with dilation rates of 1, 2, 5, and 9. MAAF-Con1 represents the use of dilated convolutions with rates of 1, 2, 5, and 9 in each iteration. MAAF-Con2 represents the use of dilated convolutions with rates of 1, 2, 3, and 4 in each iteration. MAAF-Con3 represents the use of dilated convolutions with rates of 1, 2, 4, and 8 in each in each iteration. The experimental results are shown in Table 7.

The above experimental results indicate that compared to CBAM and GAM attention mechanisms, the proposed joint attention mechanism in this paper demonstrates a stronger ability to extract channel and spatial features. In addition, using dilated convolutions with dilation rates of 1, 2, 5, and 9 in the iterative process can effectively increase the receptive field and ensure that each slice obtains feature information from other slices in each iteration.

## 5. CONCLUSION

To enhance the effective receptive field of the network while preserving feature continuity and detail, this paper proposes a multiscale feature information aggregation network with a fusion mechanism. First, the RESA module is optimized through an iterative process, and combined with ECANet channel attention and spatial attention to enhance the network's focus on the lane while increasing the network's ability to extract lane features.

**Table 5** – Effectiveness experiment of improvement points

| RESA-50 | Conv | JATT | Avg_pool | No line(%) | Shadow(%) | Curve(%) | F1-measure(%) |
|---|---|---|---|---|---|---|---|
| √ | | | | 47.7 | 72.8 | 70.3 | 75.3 |
| √ | √ | | | 48.0 | 73.3 | 70.6 | 75.5（+0.2） |
| √ | √ | √ | | 48.4 | 75.1 | 71.6 | 75.9（+0.6） |
| √ | √ | √ | √ | 48.8 | 77.2 | 72.3 | 76.3（+1.0） |

**Table 6** – Attention mechanism experimental results

| Network | RESA-50 | RESA-GAM | RESA-CBAM | RESA-JATT |
|---|---|---|---|---|
| F1-measure(%) | 75.3 | 75.7 | 75.2 | **76.3** |

**Table 7** – Experimental results of void convolution rate

| Network | MAAF-Con1 | MAAF-Con2 | MAAF-Con3 |
|---|---|---|---|
| F1-measure(%) | **76.3** | 75.4 | 76.0 |

The algorithm in this paper achieves an accuracy of 96.84% on the TuSimple dataset and a comprehensive F1 score of 76.3% on the CuLane dataset, reaching 77.2% in shadow scenes. Compared to other efficient lane networks, the algorithm in this paper effectively improves the accuracy of lane detection tasks in complex scenes. To further reduce the number of parameters and calculations and improve the inference speed of the model, future research will optimize the iterative steps of feature fusion according to the computing power and characteristics of embedded platforms and implement lightweight processing of the network.

## ACKNOWLEDGEMENT

This work is supported and assisted by the National Ethnic Affairs Commission of the People's Republic of China (Training Program for Young and Middle-aged Talents, MZR20007), the Special Project on Regional Collaborative Innovation in Xinjiang Uygur Autonomous Region (Science and Technology Aid Program)(2022E02035), the Hubei Provincial Administration of Traditional Chinese Medicine Research Project on Traditional Chinese Medicine (ZY2023M064), the Fundamental Research Funds for the Central Universities of South-Central Minzu University (Grant Number: CZO23040), and the General Project of University Industry-University Research Innovation Fund, Science and Technology Development Center, Ministry of Education(20200T08), the Fundamental Research Funds for the Central Universities, South-Central Minzu University(CZY23020,CZY23007).

## REFERENCES

[1] Li X, Li J, Hu X, et al. Line-CNN: End-to-End Traffic Line Detection With Line Proposal Unit [J]. IEEE Transactions on Intelligent Transportation Systems, 2019:1-11.DOI:10.1109/TITS.2019.2890870.

[2] Narote, Sandipann, P. et al. A review of recent advances in lane detection and departure warning system [J]. Pattern Recognition the Journal of the Pattern Recognition Society, 2018.

[3] Mammeri A, Boukerche A, Lu G. Lane detection and tracking system based on the MSER algorithm, Hough transform and Kalman filter [J]. ACM, 2014. DOI:10.1145/2641798.2641807.

[4] Wang J, Ma Y, Huang S, et al. A Keypoint-based Global Association Network for Lane Detection [J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2204.07335.

[5] HOU Y, MA Z, LIU C, et al. Learning Lightweight Lane Detection CNNs by Self Attention Distillation[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South). 2019. http://dx.doi.org/10.1109/iccv.2019.00110.DOI: 10.1109/iccv.2019.00110.

[6] ROMERA E, ALVAREZ J M, BERGASA L M, et al. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation [J/OL]. IEEE Transactions on Intelligent Transportation Systems, 2018: 263-272.http://dx.doi.org/10.1109/tits.2017.2750080. DOI:10.1109/tits.2017.2750080.

[7] Chen Z, Chen Z. RBNet: A Deep Neural Network for Unified Road and Road Boundary Detection [J]. 2017. DOI:10.1007/978-3-319-70087-8_70.

[8] Zhang H, Liang J, Jiang H, et al. Lane line recognition based on improved 2D-gamma function and variable threshold Canny algorithm under complex environment: [J]. Measurement and Control, 2020, 53(9-10):1694-1708. DOI:10.1177/0020294020952477.

[9] K Dinakaran, et al. Advanced lane detection technique for structural highway based on computer vision algorithm [J]. 2020. DOI:10.1016/j.matpr.2020.09.605.

[10] Hu J, Xiong S, Zha J, et al. Lane Detection and Trajectory Tracking Control of Autonomous Vehicle Based on Model Predictive Control [J]. International Journal of Automotive Technology, 2020, 21(2):285-295. DOI:10.1007/s12239-020-0027-6.

[11] Du X, Tan K K. Comprehensive and Practical Vision System for Self-Driving Vehicle Lane-Level Localization [J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2016, 25(5):2075-2088. DOI:10.1109/TIP.2016.2539683.

[12] Luo S, Zhang X, Hu J, et al. Multiple Lane Detection via Combining Complementary Structural Constraints [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, PP(99):1-10. DOI:10.1109/TITS.2020.3005396.

[13] Borkar A, Hayes M, Smith M T. Robust lane detection and tracking with ransac and Kalman filter[C]//Image Processing (ICIP 2009), 2009.

[14] Tang S L P. A review of lane detection methods based on deep learning [J]. Pattern Recognition: The Journal of the Pattern Recognition Society, 2021, 111(1).

[15] Li Y, Chen L, Huang H, et al. Nighttime lane markings recognition based on Canny detection and Hough transform[C]//IEEE International Conference on Real-time Computing & Robotics. IEEE, 2016. DOI:10.1109/RCAR.2016.7784064.

[16] Aly M. Real time Detection of Lane Markers in Urban Streets [J]. IEEE, 2014. DOI:10.1109/IVS.2008.4621152.

[17] YU Z, REN X, HUANG Y, et al. Detecting Lane and Road Markings at A Distance with Perspective Transformer Layers[C/OL]//2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece. 2020. http://dx.doi.org/10.1109/itsc45102.2020.9294383. DOI:10.1109/itsc45102.2020.9294383.

[18] HOU Y, MA Z, LIU C, et al. Learning Lightweight Lane Detection CNNs by Self Attention Distillation[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South). 2019. http://dx.doi.org/10.1109/iccv.2019.00110. DOI:10.1109/iccv.2019.00110.

[19] PAN X, SHI J, LUO P, et al. Spatial as deep: Spatial CNN for traffic scene understanding [J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022. http://dx.doi.org/10.1609/aaai.v32i1.12301. DOI:10.1609/aaai.v32i1.12301.

[20] ZHENG T, FANG H, ZHANG Y, et al. RESA: Recurrent Feature-Shift Aggregator for Lane Detection [J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 3547-3554. http://dx.doi.org/10.1609/aaai.v35i4.16469. DOI:10.1609/aaai.v35i4.16469.

[21] QIN Z, WANG H, LI X. Ultra Fast Structure-Aware Deep Lane Detection[M/OL]//Computer Vision – ECCV 2020,Lecture Notes in Computer Science. 2020: 276-291. http://dx.doi.org/10.1007/978-3-030-58586-0_17. DOI:10.1007/978-3-030-58586-0_17.

[22] FENG Z, GUO S, TAN X, et al. Rethinking Efficient Lane Detection via Curve Modeling [J].

[23] PAEK D H, KONG S H, WIJAYA K T. K-Lane: Lidar Lane Dataset and Benchmark for Urban Roads and Highways[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA. 2022. http://dx.doi.org/10.1109/cvprw56347.2022.00491. DOI:10.1109/cvprw56347.2022.00491.

[24] BAI M, MATTYUS G, HOMAYOUNFAR N, et al. Deep Multi-Sensor Lane Detection[C/OL]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid. 2018. http://dx.doi.org/10.1109/iros.2018.8594388. DOI:10.1109/iros.2018.8594388.

[25] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. IEEE, 2016.DOI:10.1109/CVPR.2016.90.

[26] Liu Y, Shao Z, Hoffmann N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions [J]. 2021. DOI:10.48550/arXiv.2112.05561.

[27] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA. 2020. http://dx.doi.org/10.1109/cvpr42600.2020.01155. DOI:10.1109/cvpr42600.2020.01155.

[28] Wang P, Chen P, Yuan Y, et al. Understanding Convolution for Semantic Segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV).IEEE, 2018.DOI:10.1109/WACV.2018.00163.

[29] Neven D, De Brabandere B, Georgoulis S, et al. Towards End-to-End Lane Detection: an Instance Segmentation Approach [J]. IEEE, 2018.DOI:10.1109/IVS.2018.8500547.

[30] HOU Y. Agnostic Lane Detection. [J]. arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition, 2019.

## AUTHORS

**Hong Wang** received a master's degree in Intelligent Science and Technology from the School of Computer Science, South-central Minzu University in July 2001. She is currently an associate professor with the School of Computer Science, South-Central Minzu University(SCMU), Wuhan, China. Her current research interests include autonomous driving, computer vision, and machine learning.

**Yin Ang** received a B.S. degree in Intelligent Science and Technology from the School of Computer Science, South-central Minzu University in July 2021, and is currently pursuing the M.S. degree in Computer Technology from the School of Computer Science, South-Central Minzu University. His research interests include autonomous driving, computer vision, and machine learning.

**Yilin Kang** received a PhD degree in computer science from Nanyang Technological University, Singapore, in 2015. She is currently an assistant professor with the School of Computer Science, South-Central Minzu University (SCMU), Wuhan, China. Her current research interests include knowledge discovery, cognitive and neural systems, and brain-inspired computing. Prior to joining SCMU, she was a research fellow with the Nanyang Technological University-University of British Columbia Joint Research Centre of Excellence in Active Living for the Elderly (LILY). She serves as the PC member of AAAI 2019–2022, IJCAI 2021, 2016, and OC member of IEEE WI/IAT2015. She serves as the reviewer in several major journals, such as IEEE Trans. On NNLS, IEEE Trans. On SMC, and JAAMAS.

**Shasha Tian** obtained a Ph.D. from the School of Computer Science, Wuhan University in June 2021, and went to the School of Computer Science, South-Central Minzu University in July 2003, with research interests in autonomous machine exploration and autonomous driving.

**Lu Zheng**, doctoral candidate, South-Central Minzu University(SCMU). His research interests include artificial intelligence and autonomous driving.

**Aifei Wang** received a B.S. degree in Intelligent Science and Technology from the School of Computer Science, South-central Minzu University in July 2022, and she is currently pursuing her M.S. degree in Computer Technology from the same institution. Her research interests include computer vision, machine learning, and embedded development.