Computer Science and Information Systems 21(4):1699–1723 https://doi.org/10.2298/CSIS240314054Z

Multi-frame Network Feature Fusion Model and Self-attention Mechanism for Vehicle Lane Line Detection

Guang Zhu[1], Yajuan Liu[2], and Jiyue Wang[1,3,⋆]

[1] School of Vehicle and Traffic Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450064 China
zhuguangzz@163.com
[2] School of Foreign Languages, Zhengzhou University of Science and Technology
Zhengzhou 450064 China
xdwangxd@163.com

Abstract. The traditional lane detection networks mainly use independent single frame images to extract features first and then detect them, which cannot deal with the scene with complex background well. Therefore, this paper proposes a lane parallel detection network based on multi-frame network feature fusion model and self-attention mechanism according to the scene characteristics that vehicles can obtain continuous images during normal driving. Firstly, a parallel feature extraction structure is designed. On the one hand, a single frame network with high precision is used to extract the features of the current frame. On the other hand, a lightweight multi-frame network is designed to extract features of low-resolution multi-frame temporal images. And the recurrent neural network module is used to fuse the extracted temporal features and obtain multi-frame features. Self-attention mechanism can effectively capture the relevant information of internal features. Then the fusion module of single frame feature, multi-frame feature and self-attention feature is designed. The feature map of lane line is output by up-sampling network. The experimental results show that the network in this paper has significant improvement in both objective detection accuracy and subjective effect compared with other methods.

Keywords: vehicle lane detection, multi-frame network feature fusion, self-attention mechanism.

1. Introduction

With the continuous breakthrough of artificial intelligence theory and technology represented by deep neural network, autonomous driving and advanced assisted driving have entered a period of rapid development. In practical application, it can not only effectively reduce driving accidents caused by human factors, protect people's life safety, but also greatly improve traffic efficiency, which is the inevitable development trend of intelligent transportation. Lane detection, as an important part of fully automatic driving and advanced assisted driving, is of great significance to the vehicle's environmental awareness ability and lane keeping system [1-3]. At present, although there are some commercial lane detection applications, their application scenarios are limited and cannot be applied

⋆ Corresponding Author

to complex scenarios such as short-time lane occlusion and road surface light and dark changes. Therefore, lane detection in complex scenarios is an important research direction in the field of automatic driving [4].

Due to the excellent performance of neural networks in image segmentation, scholars have proposed various lane detection networks. At present, most lane detection networks are mainly developed from semantic segmentation networks based on single frame images. Li et al. [5] proposed the classical semantic segmentation network (SegNet), which was a symmetric network structure. It recorded the position index of the pooled value in the feature map through a specially designed pooling index, and directly assigned the value to the corresponding position during up-sampling, so as to recover the image segmentation information. Liu et al. [6] proposed U-shaped network (U-NET), which was combined with deconvolution and lateral connection, to make the up-sampling operation learnable. Meanwhile, the up-sampling network can restore details by integrating high-level features to make the prediction result smoother.

In addition to traditional semantic segmentation networks, existing networks are also designed according to lane features. Ko et al. [7] proposed the end-to-end Lane Detection Network. LaneNet took advantage of the feature that lane lines had multi-lane instances, considered lane line detection as an instance segmentation problem, added an additional instance segmentation network branch, and divided all lane line pixel regions into different lane line instances. Lee et al. [8] proposed that based on the vanishing point principle and the end-to-end Vanishing Point Guided Network (VPGNet), lane lines and road markers could be processed at the same time, and the global information provided by vanishing points could be used to improve the detection accuracy in complex weather environments. Mei et al. [9] proposed Spatial Convolutional Neural Network for lane structure (SCNN), modified the original convolution model into a slice convolution suitable for predicting the strip structure, so that the information flow between row and column pixels could be transmitted in the convolution network, thus increasing the weight of lane structure in the network. Garnett et al. [10] proposed the end-to-end 3D Multiple Lane Detection Network (3D-LANenet). The front view and the transformed top view provided by the vehicle multi-camera system were used for lane line prediction. The forward stream processed and retained image information. The head-view flow provided translation invariance characteristics and output the final 3D lane line information. Although the above network can effectively detect the lane lines, it only uses the current moment image to detect the lane lines, ignoring the lane line features in the past time series, under some sudden lane occlusion, Angle of view transformation, and the change of light and shade on the ground. Due to the lack of time domain context information, the detection accuracy is not high.

At present, there are few researches on using video stream as network input to realize lane detection. Zhang et al. [11] proposed the method of combining CNN and Recurrent Neural Network (RNN) for lane line detection, which had better performance than that using only CNN. Although the network introduced the RNN to improve the detection and segmentation effect. However, when the same network structure is used for feature extraction for each frame in the temporal sequence, the network will weaken the most critical feature information of the current frame, and the computing resources will be consumed in the multi-frame features with a large amount of redundancy. In addition,

once the network structure is modified, the global features will be greatly affected, which is not conducive to the optimization of network structure.

In order to make up for the lack of time domain information in lane detection network based on single ton image, the detection speed and accuracy are taken into account. This paper proposes a parallel lane detection network based on image sequence. On the basis of the single-ton image segmentation network, a multi-frame network based on RNN is added, which takes the multi-frame images in the past, including the current frame, as the input of the network. The Convolutional long-short Term Memory (ConvLSTM) gate mechanism [12] is used to fuse and extract the target feature information and environment context information in multi-frame images according to the time series.

In order to reduce the new computational burden caused by multi-frame network, the low-resolution input strategy is adopted, and the lightweight network is selected as the skeleton network of multi-frame network, which can effectively reduce the number of parameters and computational complexity of multi-frame network. At the same time, self-attention mechanism is adopted to extract context features. In order to effectively fuse the multi-frame time-domain features extracted by multi-frame network and the global semantic features extracted by single frame network, a feature fusion module with channel connection is designed. So that the single frame feature can effectively fuse all the multi-frame time domain information, so that the fusion feature has the ability to represent both the spatial information and the time domain information. Finally, the up-sampling network is used to output the lane line feature map.

The paper is organized as follows. Section 2 introduces the related works for this paper. In section 3, we introduce the proposed lane parallel detection network. Data enhancement and model training strategy are shown in section 4. Experiments are displayed in section 5. There is a conclusion in section 6.

2. Related Works

From the perspective of how to define lane detection tasks, lane detection methods based on deep learning can be divided into four categories:

1. Category-based methods [13], which usually combine prior information to determine lane location. For example, DeepLane2 overcomes the shortcoming that the position of lane lines cannot be directly obtained by image classification by adding some priori knowledge related to location into the network [14], and directly estimates the position of lane lines by using deep neural networks.
2. The lane detection method based on classification has a simple network structure, but limited application scenarios, and high computational complexity of post-processing. The method based on object detection uses coordinate regression to detect the boundary box position of the lane line. For example, VPGNet1 uses an improved vanishing point location method to propose an end-to-end multi-task network [15], which can identify lane lines in rainy and low-lighting weather conditions, but requires complex post-processing operations such as point sampling, clustering and lane line regression. Reference [16] proposed a lane detection method based on spatiotemporal deep learning, which used time and space constraints to narrow the scope of the search

area and could accurately detect the lane boundary. However, its data flow and network junction architecture were complex, and the preprocessing could not provide effective results when the initial assumptions were not satisfied.

3. Model-based approach. It uses mathematical model to model the lane lines, and fits the lane lines by solving the model parameters. In reference [17], segmented straight line model was used to fit curved lane lines, heuristic search algorithm was used to search lane boundaries in ROI, and the detected lane lines were fitted as continuous smooth curves, which could well approximate curves. However, this method was only applicable to lane lines with small curvature, and the detection effect of lane lines with large curvature was not accurate enough. In reference [18], a hyperbola model was used to fit the lane lines, so that the two hyperbola converged to the same point, which effectively overcomed the problem of discontinuity of the linear and parabolic models at the junction between the straight and the curve. However, its detection effect was poor in bad weather or when there was interference from other vehicles. Each mathematical model has its own advantages and disadvantages, and it is usually not possible to perfectly fit the lane lines through only one model.
4. Method based on image segmentation. It marks the lane lines and background pixels into different classes, and the detection results are in the form of pixel-level classification. Compared with traditional methods, the detection effect is better and the robustness is better through end-to-end learning. Reference [19] proposed an improved segmentation network (SegNet) algorithm, which used convolution and pooling to extract lane features, and used connected domain constraints to achieve accurate segmentation of lane lines in binary images, but its structure was complex and its real-time performance was poor.

LaneNet4 is a lane detection method based on instance segmentation, with two decoders, one for segmentation and one for embedding,LaneNet can cluster each lane in the training phase rather than the post-processing phase. CNN-LSTM [20] is a hybrid depth network structure that combines CNN and RNN. Two layers of LSTM are added between encoder and decoder, and time domain information is integrated. This method can also achieve better performance under vehicle occlusion, shadow and bad weather conditions. However, the combination of CNN and LSTM requires a large amount of computation. The relevant MMA-Net algorithm in this paper belongs to the method based on image segmentation, which uses 4-layer CNN as the encoder to extract the local and global memory features of the lane lines. The Local and GlobalMemory Aggregation Module (LGMA) is used to aggregate the local and global memory features of multiple layers to enhance the features of the target frame. Finally, the results of lane segmentation are predicted by the decoder. Aiming at the problems of low accuracy and high omission rate of MMA-Net algorithm, this paper proposes a lane parallel detection network based on multi-frame network feature fusion model and self-attention mechanism.

3. Proposed Lane Parallel Detection Network

The structure of lane parallel detection network based on image sequence in this paper is shown in Figure 1, which is mainly composed of two network modules: multi-frame network and single-frame network. Multi-frame network is used to extract time-domain fea-

tures from multi-frame temporal images. The single frame network is based on the coding-decoding model, which is used to extract the global semantic features of the current moment image, and output the final result through the fusion module and the up-sampling module. Because of the parallel structure design, single-frame network and multi-frame network are endowed with different information flows, so that multi-frame network can learn multi-frame time domain characteristics. Single frame network can learn the spatial semantic features of a single frame image, connect these features, and express the lane information more comprehensively. In addition, the relatively independent parallel network structure is also convenient for structural tuning.

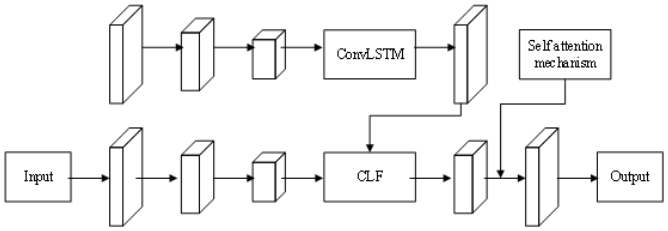

Fig. 1. Proposed network

The multi-frame network in this paper uses a lightweight skeleton network, which takes the multi-frame images in the past, including the current frame, as the network input. Since the introduction of multi-frame network will bring great computational complexity, the low resolution input strategy is adopted. By reducing the input resolution of the multi-frame image, the network computation load is obviously reduced and the network computation time is shortened. After extracting the features of each image output by skeleton network in different time domain, how to effectively fuse the features extracted by each image becomes the main problem. In this paper, the ConvLSTM network module is used to process these features, and the Long short-term memory network (LSTM) [21,22] is used to process the key target feature information in the multi-frame feature sequence while forgetting the unimportant feature information, so as to obtain the final multi-frame time-domain feature.

The single-frame network in this paper is more complex than the multi-frame network. This is because in reality, compared with the image frames at other moments in the past, the image frames at the current moment contain more accurate semantic information, and high-complexity network structures often have stronger generalization learning ability and representation ability for semantic information [23]. Therefore, single-frame network adopts more complex network structure and deeper network level to learn higher level abstract features, and integrates feature information of different sizes through pyramid module structure to obtain single-frame features.

A ConvLSTM fusion module (CLF) is designed to fuse the feature information of single-frame network and multi-frame network. The single frame feature can be fused with the complete multi-frame feature in time domain by means of channel connection. Thus, it makes up for the lack of single-frame features in time domain information, and single-frame features occupy more feature dimensions in the fused features. Ensure that

the single frame feature can be used as the dominant feature in the fusion feature, and the fusion feature can output the final fusion feature map through the up-sampling network.

3.1. Multiple Frame Network

In order to reduce the computational burden of multi-frame network due to the processing of continuous multi-frame images, this paper adopts multi-resolution strategy, which takes low-resolution continuous images as the input of multi-frame network. The high resolution single frame image is used as the input of the single frame network. The following describes the influence of the image input resolution on the network computation in neural networks when the resolution increases gradually. The classic Residual Network with 50 Layers (ResNet50) [24] was used as the test object, and three conventional resolution images were selected as the input. The image input resolution is linearly increased by multiples based on 320x180. The influence of image input resolution on network computation amount and frame number is shown in Table 1.

Table 1. Influence of image input resolution on network computation amount and frame number

Resolution	Network computation amount	Frame/s
320×180	392.3	151.1
480×270	758.6	123.6
640×360	1336.7	91.9

As can be seen from Table 1, the increase of image input resolution is not equivalent to the linear growth of network computation amount, but increases in accordance with the exponential trend. At the same time, it is found that the more the number of channels in the network layer, the greater the increase of the computation. Therefore, by reducing the resolution of the image input, the computation amount of the network can be effectively reduced, and the saved computation amount can be exchanged for the application of more complex and deeper network structure, so as to make up for the loss of image details caused by the resolution. When focusing on semantic segmentation problems such as lane lines, reducing the resolution can also dilute some unimportant background details and highlight the difference between lane lines and background. Therefore, it is reasonable and desirable to reduce the resolution to some extent.

In this paper, a multi-frame network is parallel to a single-frame network. In order to reduce the computational burden caused by the introduction of multi-frame network and achieve better network performance, the second edition of efficient spatial pyramid is selected after analyzing the existing lightweight neural networks network (ESPNetV2) serves as a skeleton network for multi-frame networks. Compared with Image Cascade Network (ICNet) [25], Efficient Residual Factorized Convolutional Network (ERFNet) [26] and the Second Edition of Mobile Vision Convolutional Network [27], the number of floating point operations per second of ESPNetV2 decreases to $[1/9-1/12]$, and the accuracy decreased by only 2%-4%. ESPNetV2 network further optimizes the convolution mode of ESPNet, reduces the number of training parameters of the network by point-group convolution and void depth-separable convolution , while maintaining the original

network structure similar to spatial pyramid [28], and proposes the highly efficient spatial pyramid module (EESP).

The EESP module is mainly based on the block convolution principle and spatial pyramid theory. Firstly, the dimension of the input features is reduced by Group Convolution (GConV), and then the Depthwise Separable Convolution (DSConV) operation is performed by using Convolution of different scales to verify low-dimensional features. Finally, the features are concatenated by element summation. The specific structure of the module is shown in Figure 2.

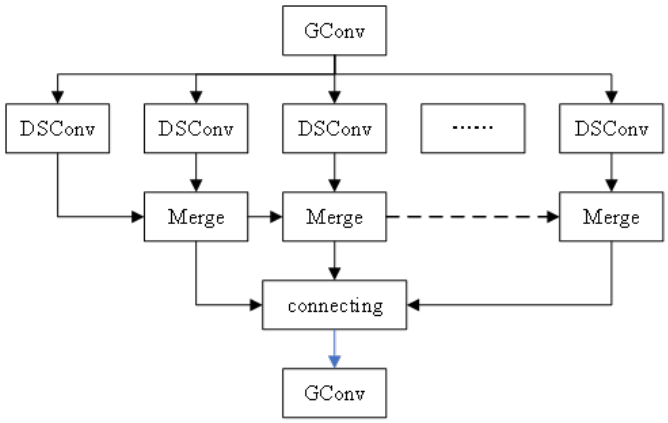

Fig. 2. Structure of EESP module

The EESP module replaces conventional convolution by grouping convolution and depth-separable convolution respectively, which can effectively reduce the computational cost of convolution operation. The features are convolved by expanding convolution kernels with different scales in Hierarchical Feature Fusion (HFF) module. The local feature information and global feature information under different receptive fields are fused by summing element by element. The information receiving field of the whole network is expanded, so as to effectively improve the detection effect. At the same time, the overall module network structure of EESP is similar to the spatial pyramid structure, so that the network has stronger semantic information and enriches the receptive field.

The ESPNetV2 skeleton network is constructed based on the EESP network module. After initial convolution, the skeleton network structure is divided into four spatial levels. Each space level employs one or more EESP network modules. After analyzing the original network structure, this paper finds that the original skeleton structure applies multiple EESP modules to the same scale in Part 2 and Part 3, and a large number of repeated convolution operations will repeatedly extract feature information, resulting in certain information redundancy. Therefore, in this paper, the number of EESP modules in part 2 and part 3 of the skeleton network is further compressed, and finally two and four EESP modules are adopted, respectively, to further reduce the computational burden of multi-frame network. The skeleton network structure of the final multi-frame network is

shown in Table 2. Each temporal image will pass through the skeleton network to extract features and obtain multi-frame feature sequences.

Table 2. Backbone network structure of multi-frame network

Layer	Number	Output size
Convolution	1	112×112
Strided EESP	1	56×56
Strided EESP	1	28×28
EESP	2	28×28
Strided EESP	1	14×14
EESP	4	14×14
Strided EESP	1	7×7
EESP	3	7×7

When feature sequences are obtained through multi-frame network, how to extract the target feature information that changes in time domain from feature sequences. To effectively ignore the background features that do not change much, that is, how to abstract features selectively is very important for segmentation target. Therefore, this paper studies the characteristics of RNN and introduces LSTM network module to obtain target feature information selectively. As a special form of RNN, LSTM uses three different gate functions to extract long-term features, which are the input gate that controls the addition of new information, the forgetting gate that controls the passage of information, and the output gate that determines the output of information.

In this paper, the gate mechanism of LSTM is used to process multi-frame feature sequence, and the target feature information in multi-frame feature sequence is extracted according to the time domain. At the same time, the unimportant feature information is forgotten. However, LSTM is a fully connected network model, which will lead to extra time consuming. In this paper, the ConvLSTM structure is adopted, and the convolution operation is used to replace the matrix multiplication operation in each gate function, so as to capture the potential spatio-temporal features in the temporal features. Convolution is also more effective and easy to understand the image feature extraction, effectively improve the image feature processing speed and adaptability to feature maps. A regular ConvLSTM memory unit is computed at time t as follows.

$$i_t = \sigma(W_{xi} \times X_t + W_{hi} \times H_{t-1} + W_{co} \times C_{t-1} + b_i). \tag{1}$$

$$f_t = \sigma(W_{xf} \times X_t + W_{hf} \times H_{t-1} + W_{cf} \times C_{t-1} + b_f). \tag{2}$$

$$C_t = f_t \times C_{t-1} + i_t \tanh(W_{xc} \times X_t + W_{hc} \times H_{t-1} + b_c). \tag{3}$$

$$o_t = \sigma(W_{xo} \times X_t + W_{ho} \times H_{t-1} + W_{co} \times C_{t-1} + b_o). \tag{4}$$

$$H_t = o_t \times \tanh(C_t). \tag{5}$$

Where, i_t represents the exit entry at time t. f_t represents the forgetting gate at time t. C_t represents the ConvLSTM cell state at time t. o_t represents the output gate at time t. represents the state of the hidden layer at time t. σ stands for sigmoid function. $tanh$ is the hyperbolic tangent function. $\times$ denotes the convolution operation. X_t represents the input feature at time t. W_{xi} represents the weight matrix of input X_t under the input gate. b_i represents the offset value under the input gate. In general, C_t changes slowly. H_t varies a lot from node to node.

3.2. Single Frame Network

Since the single frame feature at the current moment needs to be the dominant feature, the skeleton network of the single frame network needs to use a more complex network structure to improve the generalization learning ability and representation ability of features. In this paper, classical neural networks such as ResNet, VGG16-BN and GoogleNet are selected as the skeleton network of single-frame network, and the fully connected layer in the network structure is removed.

The number of channels in the output feature map of different skeleton networks is different. In order to keep the output feature map of the single-frame network and the multi-frame network have the same scale and equal ratio of channels, and strengthen the feature information of the single frame feature at different scales, the multi-scale feature enhancement structure [29] is added after the single-frame skeleton network, and the feature information is enhanced at three different scales. After using the enhanced structure, each spatial location can view the local environment in different scale space, further expand the information receiving domain of the whole network, and improve the effect of lane detection.

The ConvLSTM fusion module (CLF) is designed to effectively fuse the single frame features and multi-frame features extracted through parallel network. As shown in Figure 3, CLF first up-samples the multi-frame feature map to restore the feature map size to the same size as the single-frame feature. Then 1x1 convolution kernel is used for single-frame features and multi-frame features to smooth features. Then, multi-frame features and single-frame features are fused by channel connection, in which single-frame features occupy more feature dimensions in the fused features and serve as the dominant features. Finally, the fused features were activated at the site to the reasonable Satisfaction Linear Unit (ReLU) to reduce the interdependence between parameters.

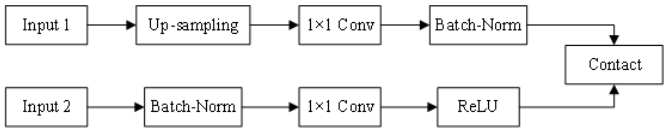

Fig. 3. CLF structure

3.3. Self Attention Mechanism

Self-attention mechanism can effectively capture the relevant information of internal features. The self-attention mechanism is used to automatically learn the context information of the users in the group and the influence of other users, which can solve the problem of dynamic complexity of the members in the group. The structure of the self-attention mechanism is shown in Figure 4. For the analysis of one module, c_1 represents the context information vector representation of the first user u_1. $u_2, \cdots, u_n$ stands for other user vector representation. According to the context information of the user and the embedding vector of the user into a matrix $U_1 = [c_1, u_2, u_3, \cdots, u_n]$, $U_1 \in R_d$. d represents the embedding vector dimension of the user, and other modules learn in the same way. The input to the self-attention mechanism used in this article consists of query, key, and value. The input of query, key and value is composed of the matrix U_1 formed by embedding vectors of user and context information, and the three are equal, and the output is the final weighted $A_{u,i}$. The specific calculation steps are as follows:

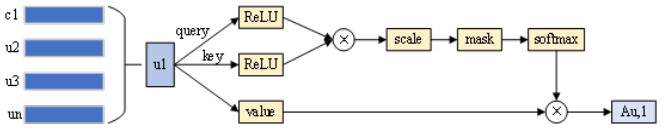

Fig. 4. Structure of self attention mechanism

Project the input query and key into the same space and introduce nonlinear functions to learn Q and K.

$$Q = \sigma(U_1 \cdot W_a). \tag{6}$$

$$K = \sigma(U_1 \cdot W_k). \tag{7}$$

Where σ is the Relu activation function. $W_a = W_k$ is the weight matrix of Query and key respectively, $W_a = W_k \in R_d \times d$.

This paper uses the compressed inner product attention function. The results in (1) are multiplied to get the corresponding weight matrix B_1, and then normalized.

$$B_1 = \sigma(\frac{Q \cdot K^T}{\sqrt{d}}). \tag{8}$$

Where σ is the softmax function. The scale operation uses $\sqrt{d}$ to expand and shrink attention to desaturate the function value to avoid too low gradient. Secondly, through mask operation, the diagonal elements of the matrix are hidden to avoid high matching scores between the same vectors of Query and key.

The weight matrix B_1 can be obtained from Equation (8), and the group matrix represents $A_{u,1}$ by multiplying the weight moment B_1 and value obtained. In this method, the user context information and user information are concatenated into a matrix U_1 and

the final weighted output is learned by the self-attention mechanism to obtain the group representation $A_{u,1}$.

$$A_{u,1} = B_1 \cdot U_1. \tag{9}$$

Where U_1 is the matrix spliced by the context information vector of u_1 and other member user vectors in the group. B_1 is the weight matrix obtained by compressing the inner product attention function. The above is the first module that integrates contextual information to obtain group representation $A_{u,1}$ through self-attention mechanism learning. Other modules learn in the same way, that is, n groups can be obtained, representing $A_{u,2}, \cdots, A_{u,n}$.

3.4. Model Optimization

Bayesian personalized ranking (BPR) establishes a triplet model for users and items. Assuming the user views the positive items, the user must prefer the items he sees to the negative items he does not see. Therefore, the sorted items of positive items are better than the negative items. The BPR optimization objective is to perform the optimal personalized ranking based on the maximum a posteriori estimator, and its formula is as follows:

$$L = argmin \sum_{(g,j,k) \in D_S} - \ln \sigma(X_{u,a}(\theta) - X_{u,b}(\theta) + \lambda ||\theta||^2). \tag{10}$$

Where (g, j, k) belongs to the triplet of set D_S and contains all pairs of positive and negative items of each user. $X_{u,a}$ represents the predicted score of user u for Project a. $X_{u,b}$ represents user's prediction score for project b. σ is the sigmoid function. λ is the regularization parameter. is the model parameter.

This paper takes BPR as the basis of model learning, treats a group as a user, uses BPR algorithm to make recommendations, and then uses matrix factorization method. Therefore, the objective function of this paper is:

$$L = arg \min_{\theta} \sum_{(g,j,k) \in D_S} - \ln \sigma(g_1 \cdot v_a - g_1 \cdot v_b + \lambda ||U||^2 + \lambda ||C||^2). \tag{11}$$

Where g_1 represents the group representation of the mean embedding. U and C are parameters which represent the set of user latent vector and user context vector, respectively. $g_1 \cdot v_a$ represents the group's predicted score on a. $g_1 \cdot v_b$ represents the group's predicted score on b. (g, j, k) belongs to the triplet of set D_S, including all pairs of positive and negative items of each group.

3.5. Up-sampling Network Structure

As a decoding network, the up-sampling network uses the underlying features to recover the target information by up-sampling. The up-sampling network in this paper adopts the same decoding structure as LaneNet. LaneNet consists of four convolutional layers and one up-sampling layer. The up-sampling layer smooths the sampled feature map through the up-sampling algorithm of bilinear interpolation. Finally, the fused feature map is enlarged to the same size as the input feature map, and this feature map was output as the result feature map.

4. Data Enhancement and Model Training Strategy

4.1. Data Enhancement

In this paper, network training is based on TuSimple road data set [30], and the data set is enhanced accordingly. The TuSimple road dataset contains a total of 6570 groups of lane line image sequences. Each sequence contains 20 consecutive frames of images acquired within 1s. All images are based on the foreground view of the car. The original resolution of each frame is 1280×720. For each set of image sequences, only the last 20 frame image marks the true lane line label. In TuSimple dataset, 3626 sets of image sequences are used as the training set and 2944 sets of image sequences are used as the test set, including different time periods during the day, different number of lanes and various traffic conditions. To a certain extent, it is in line with the realistic road situation.

In order to avoid over-fitting and improve the performance of the dataset, data enhancement was performed on the dataset. According to the network structure in this paper, for each group of image sequence, 4 images are sampled from the 20 consecutive frames at different intervals, and a group of training data is composed by combining the label images. The road images collected by the camera at different speeds can be simulated by acquisition at different intervals. In this paper, 1, 2, 3, 4 and 5 frames are respectively used as sampling benchmarks, and this standard is applied to each group of image sequences. The sampling methods of continuous ton images are shown in Table 3.

Table 3. Sampling of continuous frame image

Sampling interval	The sampling frame
1	17,18,19,20
2	14,16,18,20
3	11,14,17,20
4	8,12,16,20
5	5,10,15,20

After the above processing, 5 times of the original data set image sequence data is obtained. In order to further improve the diversity of training data, image horizontal flip is used for data enhancement. Aiming at the ratio of 0.55:0.45 in the training set and test set on the original data set, the more commonly used ratio of 0.8:0.2 is used to construct the training set and test set. Finally, a training set including 51260 groups of image sequences and a test set containing 12820 groups of image sequences were constructed.

In order to test whether the proposed network can effectively cope with complex scenes such as lane occlusion and light shadow, 540 groups of complex image sequences were further selected from the test set as the complex test set for subsequent experiments, including 435 groups of lane occlusion image sequences and 105 groups of light shadow image sequences. In addition, in order to improve the training speed, 640×360 is finally selected as the basic resolution.

4.2. Training Strategy

After the parallel network structure is established in this paper, the design of reasonable training strategy is also a key part of neural network training. In order to effectively extract the features of the lane model through continuous update and learning of network parameters, gradually reduce the error value with the real label, and finally achieve the effect of separating the lane line from the background, this paper formulated the training strategy as follows.

1. Consider the initialization of network parameters. The methods of network parameter initialization mainly include network pre-training parameter initialization, random parameter initialization and fixed value parameter initialization. Because the network in this paper adopts parallel network structure design, the initialization parameter strategy will be different. For the skeleton network of a single frame network, the pre-trained parameters of the network model on the ImageNet dataset are used for parameter initialization. For the convolutional layer in the multi-frame network, the Kaiming normal distribution initialization algorithm is applied to initialize the parameters of the convolutional layer. For the Batch Normalization (BN) layer in the multi-frame network structure, according to the initialization method in reference [31], the weight value and bias value are filled with fixed 1 and fixed 0, respectively
2. In this paper, the network adopts Stochastic Gradient Descent (SGD) as the optimization algorithm, with the initial learning rate=0.03, weight $decay = 5e^{-5}$, and momentum parameter=0.9. The learning rate strategy selects Poly as the way to adjust the learning rate. The logarithmic function makes the learning rate decay according to the number of cycles, which avoids the problem that the learning rate is too large to converge in the later stage of neural network training.
3. Weighted cross-entropy is used as the loss function of the network in this paper, which has considerable effect on image classification and segmentation. According to the idea of classification, the cross drop loss function first makes class prediction for each pixel of the segmented image, then calculates the loss value of a single pixel by comparing the prediction results with the label pixels, and finally calculates the average value of the total loss value of all pixels. Therefore, each pixel can be learned by the network and continuously improve the output result, but such loss function can not be directly used to deal with the lane segmentation problem. This is because in the actual image, the background pixels are far more than the lane pixels, which leads to the category imbalance in the learning process, making the final loss value dominated by the background category, and the lane features are difficult to be learned by the network. With different categories in order to solve this problem, can be obviously improved by means of weighted study effect, loss of background value multiplied on a smaller weight value, can significantly reduce background for the effects of loss value, otherwise, the target class on a large weight value, increase the probability of the target class are studied, the loss function are defined as follows:

$$Loss = \sum_{x \in \Omega} w(x) \ln(p_{l(x)}(x)). \tag{12}$$

Where x represents the target category of segmentation. $w(x)$ represents the weighted value of the current category. $p_{l(x)}$ represents the probability that the network fits the

predicted value l to the true value x which is calculated by KL divergence. If the predicted value is close to the true value, the loss value is small. Otherwise, the loss is greater. In this paper, the W-weighted value of the background class is set as 0.02, and the W-weighted value of the lane line is set as 1.02.

5. Experiments and Result Analysis

The network in this paper is implemented based on Pytorch.1 and runs on 64-bit Ubuntu 18.04 system. In the computer hardware configuration: the processor is Intel 9700K CPU4.50 GHz; GPU is GeForceGTX-1080TI; The video memory is 11GB.

5.1. Evaluation Indexes

In order to comprehensively evaluate the network performance in this paper, the evaluation indexes used in the experiment are as follows.

Accuracy. Measurement of the proportion of the number of correctly classified pixels in the total number of pixels:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

Where TP represents the number of pixels correctly predicted as road labels and it is defined as the true positive sample. TN denotes the number of pixels correctly predicted as road labels defined as the true negative sample. FN denotes the number of pixels that incorrectly predict a road label as a background label defined as a false negative sample. FP represents the number of pixels that incorrectly predict the background label to be a road label defined as a false positive sample.

The accuracy rate cannot accurately describe the lane detection accuracy in the actual situation. This is because in ordinary road environment, most pixels are background pixels rather than lane pixels. The extreme imbalance of this category leads to a high accuracy in the image with fewer pixels to predict lane lines, even if all of them are predicted as background labels. Therefore, in this paper, F1 is used to measure the comprehensive performance of the network model for the single category of lane. As a model accuracy index, the F1-measure evaluates the comprehensive performance of the network through the weighted average of Precision and Recall. The detailed calculation is as follows:

$$F1 = \frac{2P \cdot R}{P + R}. \tag{14}$$

$$P = \frac{TP}{TP + FP}. \tag{15}$$

$$R = \frac{TP}{TP + FN}. \tag{16}$$

5.2. Skeleton Network Selection

In order to further verify that the network in this paper still plays a role when other single-frame skeleton networks are used, VGG16-BN, GoogleNet, ResNet32 and ResNet50 are selected as the skeleton networks of the single-frame network, and comparative experiments are carried out. The experimental results are shown in Table 4. It can be seen from the table that different skeleton networks have a certain degree of influence on the detection effect. The index values of ResNet50 are the highest, followed by VGG16 and ResNet32, and GoogleNet is the lowest.

Table 4. Experimental result comparison of different backbone networks

Backbone network	Accuracy	P	R	F1	Frame/s
GoogleNet	0.970	0.891	0.913	0.901	33.2
ResNet32	0.973	0.895	0.928	0.910	28.3
VGG16	0.975	0.905	0.929	0.918	22.6
ResNet50	0.984	0.914	0.948	0.928	19.9

Figure 5 shows the subjective detection results of different skeleton networks. As can be seen from the figure, the subjective detection effect of ResNet50 and VGG16-BN is better, and the edge of ResNet50 is finer. GoogleNet detection effect is poor, the edge is rough, and the lane line is broken.

5.3. Network Structure Test

In this paper, the TuSimple dataset is enhanced by randomly dividing the dataset according to the ratio of 0.8:0.2. But there is a definite chance of randomization. In order to eliminate the contingency, the experiment uses the way of 5 fold cross validation to carry out multiple tests. The original data set was divided into 5 groups, each group of subset data was used as a test set, and the remaining 4 groups of subset data were used as training set. Finally, five model results were obtained. The results of these five models on their respective test sets are calculated and the average value is calculated, which is used as the performance index of the network in this paper. In cross-validation, 640×360 is selected as the input image resolution of single-frame network, and 320×180 is selected as the input image resolution of multi-frame network.

The cross-validation results of the network in this paper are shown in Table 5. The experimental data are relatively stable, and the floating range of all parameters is within 0.005, indicating the stability of the network in this paper.

In order to test the impact of the proposed multi-frame network on the performance of the proposed single-frame network, the ablation experiment was further used to comprehensively test the network structure, using only the single-ton network, using only the multi-frame network, and using the proposed network, respectively. The test results are shown in Table 6. As can be seen from Table 6, compared with multi-frame network and single-frame network, the proposed network achieves better results in all parameters,

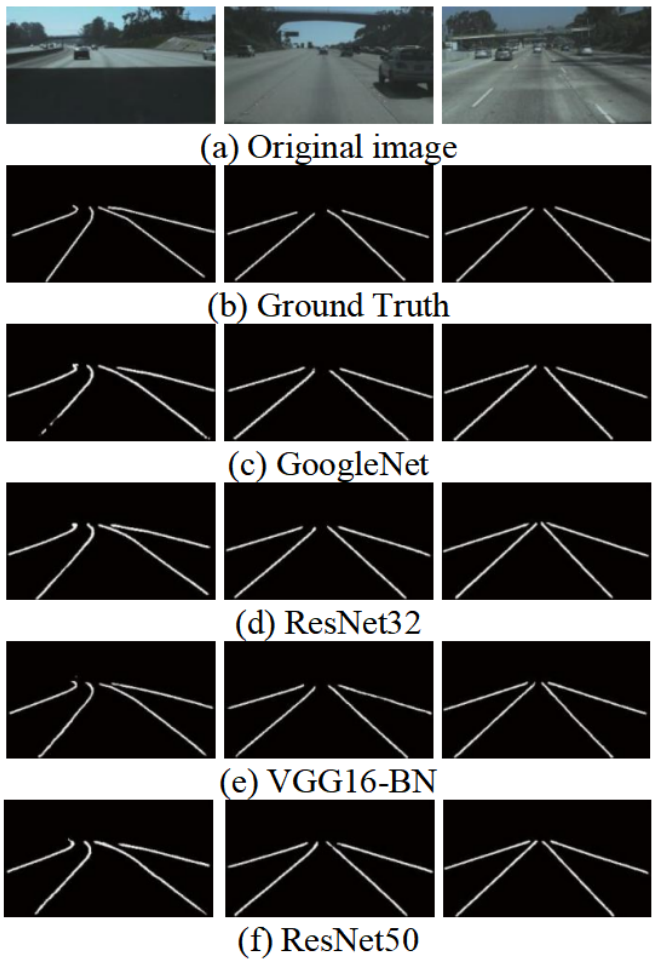

Fig. 5. Subjective results of different backbone networks

Table 5. Cross validation results of the proposed network

Cross validation	Accuracy	P	R	F1
1	0.985	0.919	0.942	0.931
2	0.985	0.921	0.941	0.931
3	0.985	0.918	0.937	0.928
4	0.984	0.913	0.943	0.928
5	0.985	0.916	0.936	0.927
Average	0.985	0.917	0.940	0.929

which objectively verifies the effectiveness of the proposed network structure and indicates that single-ton network can effectively improve network performance after adding multi-ton network.

Table 6. Results of different network structures on test set

Network	Accuracy	P	R	F1	Frame/s
Single frame	0.981	0.885	0.916	0.893	29.1
Multi-frame network	0.976	0.828	0.892	0.857	44.5
Proposed	0.985	0.917	0.942	0.939	22.4

Figure 6 shows the subjective detection results of the three networks under three different scenarios: no lane line is blocked, there is lane line is blocked, and light shadow. As shown in figure 6, in the image with no lane line blocked on the left, the three network structures can effectively detect lane lines. But the details at the edges of this network are smoother. In the image with blocked lane lines in the middle, the multi-frame network and the proposed network can effectively detect the blocked lane lines due to the advantage of image sequence, while the single-frame network has lane line breaks at the occlusion. In the image of light shadow on the right, the lane line is discontinuous in both multi-frame network and single-frame network, and the proposed network can still detect the lane line effectively.

According to the above experimental data, because the multi-frame network adopts lightweight skeleton network, the ability to learn lane features is not strong, resulting in lane discontinuity and poor edge quality. Although the parameters of the single-frame network are better than those of the multi-frame network, it lacks the time domain perception ability of lane features in complex scenes such as lane occlusion and light shadow, and the lane line discontinuity occurs at the occlusion. In this paper, the multi-frame network module is added on the basis of the single-frame network, which not only strengthens the generalization learning ability and representation ability of the single-frame network for complex scenes, but also integrates the time-domain context features given by the multi-frame network, and obtains the blocked lane features according to the RNN features in the case of short-term occlusion.

From the above experimental results, it can be seen that compared with the independent single-frame network and multi-frame network, the proposed network achieves better detection effect in both subjective and objective evaluation.

5.4. Multiple Resolution Selection

A low resolution image input strategy is used in multi-frame networks to reduce network computation. In order to test the influence of different input image resolutions on the performance of the overall network, three different resolutions are selected as the input image resolutions of the multi-frame network while keeping the resolution of the input image of the single-frame network as 640×360. The specific results are shown in Table 7.

As can be seen from Table 7, reducing the input resolution of the multi-frame network will slightly reduce the network detection effect, but effectively reduce the running time

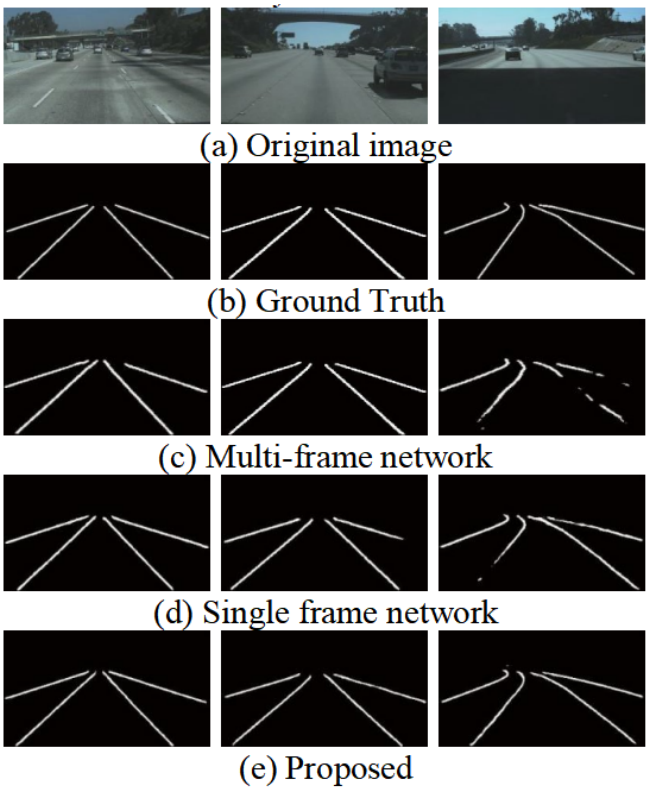

Fig. 6. Subjective results of different network structures

Table 7. Experimental result comparison with different input resolutions

Resolution	Accuracy	P	R	F1	Frame/s
640×360	0.982	0.918	0.941	0.928	21.1
480×270	0.982	0.919	0.938	0.927	22.0
320×180	0.982	0.914	0.937	0.925	22.7

of the multi-frame network and improve the number of network calculations. Reducing the input resolution will make the edge sampling less refined, and the edges of the longer lane lines will be rougher, resulting in a slightly lower overall accuracy.

Figure 7 shows the subjective detection results under different input resolutions. As can be seen from the comparison of the detected images in the figure, the subjective visual gap is not obvious. In view of subjective and objective evaluation, 320×180 is finally selected as the input resolution of multi-frame network.. The experiments in this section also further show that the performance of multi-frame network has a positive influence on the overall parallel network to a certain extent, and optimizing the performance of multi-frame network can also improve the performance of parallel network in this paper.

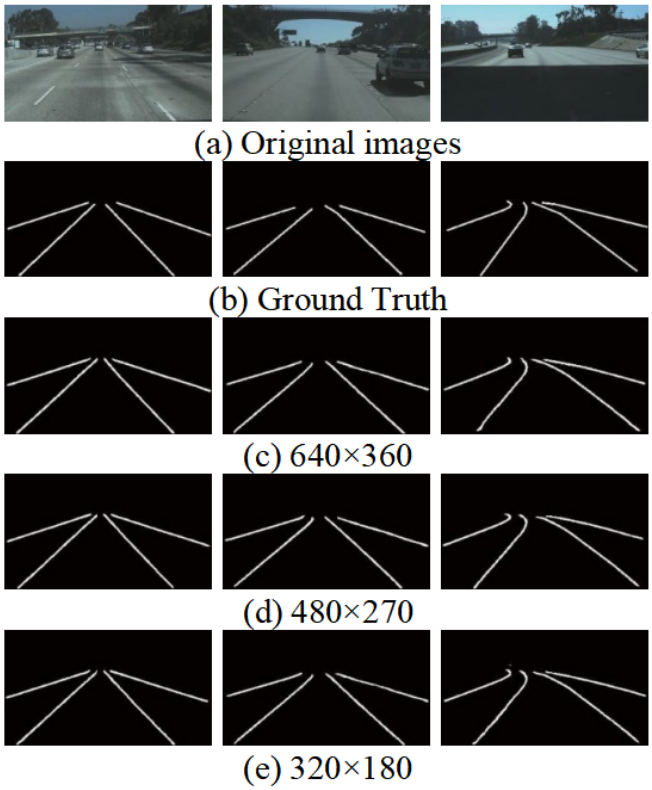

Fig. 7. Subjective results with different input resolutions

5.5. Comparison with Advanced Networks

In order to further evaluate the proposed network, several existing neural networks that also use VGGNet as the skeleton network are selected for performance comparison, including SegNet, UNet, LaneNet, SegNET-CL and UNet-CL. In order to better evaluate the

performance of the proposed network and other neural networks based on single frame detection and multi-frame detection, the performance factors of skeleton network are excluded. The above networks are trained with the same training set, and tested with the same overall test set and complex test set with lane line occlusion respectively. The experimental results are shown in Table 8 and Table 9, respectively.

Table 8. Performance comparison of different networks on TuSimple test set

Network	Accuracy	P	R	F1	Frame/s
LaneNet	0.981	0.875	0.916	0.893	30.1
UNet	0.982	0.888	0.910	0.899	20.0
SegNet	0.982	0.887	0.918	0.901	26.0
SegNet-CL	0.983	0.908	0.916	0.909	25.1
UNet-CL 0.983	0.899	0.927	0.912	19.1	
Proposed	0.985	0.917	0.940	0.928	23.6

Table 9. Performance comparison of different networks on complex test set

Network	Accuracy	P	R	F1
UNet	0.978	0.831	0.853	0.841
UNet-CL	0.980	0.853	0.889	0.865
Proposed	0.982	0.878	0.901	0.887

It can be seen from Table 8 that, compared with the existing neural networks for multiple single-frame detection, the proposed network can improve the accuracy, recall and F1 metric by about 2% when the time complexity is slightly increased.

As can be seen from Table 9, when the difficulty of the dataset increases, the performance index of each network decreases to a certain extent. But the accuracy rate of this network is still high. This further illustrates the advantages of the network in this paper. The reason is that the lane information in the image sequence is relatively continuous and changes little, but the overall background changes are more dynamic and complex. In the face of occlusion, it is difficult to accurately detect lane features in single frame detection. Through multi-frame network, continuous and less variable features such as lane lines have a higher probability to be detected.

Figure. 8 shows the comparison of subjective detection results of each network. Based on the analysis of the actual detection effect, it can be concluded that the network in this paper can still effectively remove the influence of occlusion when the lane line is suddenly blocked in columns 2 to 6 in Figure 8. For complex scenes with varying light and dark scenes such as column 4 and column 6 in Figure 8, the multi-frame features enhance the semantic information of single frame features, so the detection effect is also better.

SegNetConvLSTM and UNet-ConvLSTM, which use the same network structure to extract features for each image, weaken the most critical feature information of the current frame and produce more redundant features. Therefore, excessive redundant features need

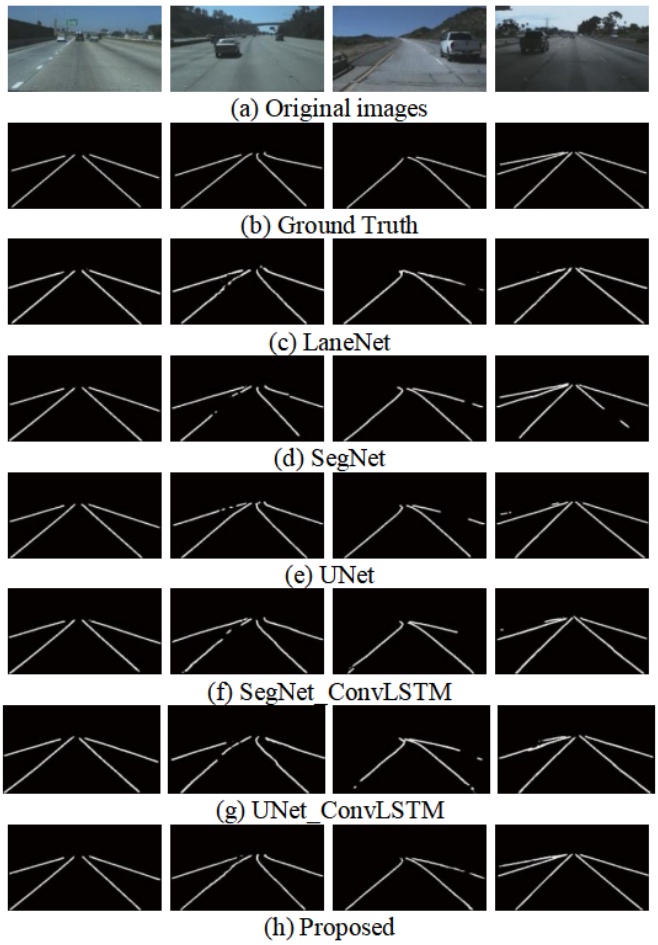

Fig. 8. Subjective result comparison of different networks

to be processed when using RNN to fuse features. This leads to a decrease in accuracy. In this paper, by giving different information flow to single-frame network and multi-frame network, and adopting skeleton network with different complexity, multi-frame network can learn multi-frame time-domain semantic features. Single-frame network learns spatial semantic features of a single frame image, which not only effectively reduces redundant features, but also strengthens the interpretation of the network and extracts lane features more effectively. According to the results in Tables 8 and 9, compared with the lane detection network based on video streaming, the accuracy rate, recall rate and F1 measure of the proposed network on the overall test set are all improved by about 1%. On the complex test set containing only lane line occlusion, the performance improvement of the proposed network is more obvious.

6. Conclusion

This paper proposes a parallel lane detection network based on image sequence. On the basis of the single frame based on the original network parallel rise frame extraction network, using RNN module containing the current frame and the frame more sequences of time-domain characteristics, environmental context feature extraction and fusion, not only effectively make up for the single frame network of temporal context information is missing, also assist improve the lane line semantic information. Since the current frame image is the most important temporal frame, compared with the multi-frame network, the single-frame network adopts the skeleton network with higher complexity to improve the generalization learning performance and representation ability of the current frame features. Experiments show that the introduction of multi-frame network module in the network of this paper has better performance index than using only single-frame network. Due to the low resolution strategy, lightweight network structure and reuse of the feature map of the past frame, the time of adding multi-frame feature module is only increased by about 25%. Although the network in this paper can make up for the missing temporal feature information of the single-frame network to a certain extent, it is still not suitable for the situation that the lane line is blocked for a long time and the road light has many shadows. In the subsequent research, the feature extraction and fusion of multi-frame network and single-frame network and the optimization of long-term occlusion and other complex actual scenes will be further studied, so as to make the spatial feature information of single frame target more complementary to the temporal information of image sequence.

Acknowledgments. This work was supported by Research project of Zhengzhou University of Science and Technology: Vibration Analysis and Control Based on Flexible Joints and Loads of Industrial Robots (2022XJKY05); Key Teaching reform project of Zhengzhou University of Science and Technology, "Research on Cultivating Interdisciplinary Innovative Practice Teaching Based on Robot Technology" (2022JGZD03); 2023 Teacher Development Research Project of Zhengzhou University of Science and Technology, Research on Career Development Dilemma and Implementation Path of Young Teachers in Application-oriented Universities in the New Era (JSFZ-ZXKT2023004); Key research projects of universities in Henan Province, "Development and Research of Intelligent Gauge Detection System of Rail Transit Based on Dynamic Distance Measurement" (24B460027), "Application Research of Rhodes Correlation Failure Model Based on Big Data Acquisition in high-speed Cutting Analysis" (24B460028); "Research on Structure Optimization and vibration Damping Performance of high-frequency needling Equipment for Nonwoven

Fabrics" (23B460011), "Research on Key Technologies of Supercapacitor Module Equalization and Safety Monitoring and early Warning based on Evidence Theory" (23B480003).

The author is grateful for the anonymous review by the review experts.

References

1. A. Dominic Savio, C. Balaji, D. kodandapani, K. Sathyasekar, R. Naryanmoorthi, C. Bharatiraja, and Bhekisipho Twala. DC Microgrid Integrated Electric Vehicle Charging Station Scheduling Optimization [J]. Journal of Applied Science and Engineering. 26(2), 253-260, 2022. https://doi.org/10.6180/jase.202302_26(2).0011
2. Liu S, Liu L, Tang J, et al. Edge computing for autonomous driving: Opportunities and challenges[J]. Proceedings of the IEEE, 2019, 107(8): 1697-1716.
3. Liu L, Lu S, Zhong R, et al. Computing systems for autonomous driving: State of the art and challenges[J]. IEEE Internet of Things Journal, 2020, 8(8): 6469-6486.
4. Tang J, Li S, Liu P. A review of lane detection methods based on deep learning[J]. Pattern Recognition, 2021, 111: 107623.
5. Li Y, Shi T, Zhang Y, et al. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 175: 20-33.
6. Liu L, Cheng J, Quan Q, et al. A survey on U-shaped networks in medical image segmentations[J]. Neurocomputing, 2020, 409: 244-258.
7. Ko Y, Lee Y, Azam S, et al. Key points estimation and point instance segmentation approach for lane detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 8949-8958.
8. Lee S, Kim J, Shin Yoon J, et al. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1947-1955.
9. Mei S, Jiang R, Li X, et al. Spatial and spectral joint super-resolution using convolutional neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(7): 4590-4603.
10. Garnett N, Cohen R, Pe'er T, et al. 3d-lanenet: end-to-end 3d multiple lane detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2921-2930.
11. Zhang W, Liu H, Wu X, et al. Lane marking detection and classification with combined deep neural network for driver assistance[J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2019, 233(5): 1259-1268.
12. Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional long short-term memory[C]//2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2017: 1-6.
13. Takahashi M, Iino K, Watanabe H, et al. Category-based memory bank design for traffic surveillance in context R-CNN[C]//International Workshop on Advanced Imaging Technology (IWAIT) 2023. SPIE, 2023, 12592: 84-87.
14. Battarra I, Accorsi R, Manzini R, et al. Hybrid model for the design of a deep-lane multisatellite AVS/RS[J]. The International Journal of Advanced Manufacturing Technology, 2022, 121(1-2): 1191-1217.
15. Garg T, Bachu S, Panda D, et al. Multi-Stage Pyramid Parsing Network For Lane Marking Detection[C]//2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, 2022: 1-6.
16. Dong Y, Patil S, van Arem B, et al. A hybrid spatial-temporal deep learning architecture for lane detection[J]. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(1): 67-86.

17. Modi S, Bhattacharya J, Basak P. Multistep traffic speed prediction: A deep learning based approach using latent space mapping considering spatio-temporal dependencies[J]. Expert Systems with Applications, 2022, 189: 116140.
18. Luo S, Yao J, Hu J, et al. Using deep learning-based defect detection and 3D quantitative assessment for steel deck pavement maintenance[J]. IEEE Transactions on Intelligent Transportation Systems, 2022. doi: 10.1109/TITS.2022.3169164.
19. Ghandorh H, Boulila W, Masood S, et al. Semantic segmentation and edge detection: Approach to road detection in very high resolution satellite images[J]. Remote Sensing, 2022, 14(3): 613.
20. Kortli Y, Gabsi S, Voon L F C L Y, et al. Deep embedded hybrid CNN-LSTM network for lane detection on NVIDIA Jetson Xavier NX[J]. Knowledge-based systems, 2022, 240: 107941.
21. Li P, Laghari A A, Rashid M, et al. A deep multimodal adversarial cycle-consistent network for smart enterprise system[J]. IEEE Transactions on Industrial Informatics, 2022, 19(1): 693-702.
22. Wang L, Yin S, Alyami H, et al. A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images[J]. Geoscience Data Journal. 2022. https://doi.org/10.1002/gdj3.162.
23. Tsironi E, Barros P, Weber C, et al. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition[J]. Neurocomputing, 2017, 268: 76-86.
24. Theckedath D, Sedamkar R R. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks[J]. SN Computer Science, 2020, 1(2): 1-7.
25. Cui Z, Chang H, Shan S, et al. Deep network cascade for image super-resolution[C]//European Conference on Computer Vision. Springer, Cham, 2014: 49-64.
26. Romera E, Alvarez J M, Bergasa L M, et al. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(1): 263-272.
27. Lee J, Tang H, Park J. Energy efficient canny edge detector for advanced mobile vision applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(4): 1037-1046.
28. Shoulin Yin, Hang Li, Asif Ali Laghari, et al. A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection[J]. EAI Endorsed Transactions on Scalable Information Systems. 21(33), e8, 2021. http://dx.doi.org/10.4108/eai.6-10-2021.171247
29. Liang H, Yang J, Shao M. FE-RetinaNet: Small Target Detection with Parallel Multi-Scale Feature Enhancement[J]. Symmetry, 2021, 13(6): 950.
30. Cao J, Song C, Song S, et al. Lane detection algorithm for intelligent vehicles in complex road conditions and dynamic environments[J]. Sensors, 2019, 19(14): 3166.
31. Gao S H, Han Q, Li D, et al. Representative batch normalization with feature calibration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8669-8679.

Guang Zhu is with the School of Vehicle and Traffic Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064 China. Research direction: Image processing, data analysis, artificial intelligence.

Yajuan Liu is with the School of Foreign Languages, Zhengzhou University of Science and Technology, Zhengzhou 450064 China. Research direction: English data processing, big data, pattern recognition.

Jiyue Wang is with the School of Vehicle and Traffic Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064 China. Research direction: Image processing, data analysis, artificial intelligence.

Received: March 14, 2024; Accepted: June 29, 2024.