Neurocomputing 465 (2021) 15–25

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps

Jingyu Li [a,b], Fengling Jiang [a,b,d,1], Jing Yang [a,c], Bin Kong [a,c,*], Mandar Gogate [e], Kia Dashtipour [e], Amir Hussain [e]

[a] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[b] University of Science and Technology of China, Hefei 230026, China
[c] Anhui Engineering Laboratory for Intelligent Driving Technology and Application, Hefei 230088, China
[d] Hefei Normal University, Hefei 230061, China
[e] Edinburgh Napier University, School of Computing, Merchiston Campus, Edinburgh, Scotland EH10 5DT, UK

ARTICLE INFO

Article history:
Received 22 June 2021
Revised 14 August 2021
Accepted 24 August 2021
Available online 27 August 2021
Communicated by Zidong Wang

Keywords:
Lane detection
Semantic segmentation
High-definition maps
Attention mechanism

ABSTRACT

Accurate high-definition maps with lane markings are often used as the navigation back-end for commercial autonomous vehicles. Currently, most high-definition maps are manually constructed by human labelling. Therefore, it is urgently required to propose a multi-class lane detection method that can automatically mark the road lanes to assist in generating high-precision maps for autonomous driving. We propose a lane segmentation detection method, named Lane-DeepLab, which is based on semantic segmentation for detecting multi-class lane lines in unmanned driving scenarios. The proposed method is based on the DeepLabv3+ network as the baseline, and we have redesigned the encoder-decoder structure to generate more accurate lane line detection results. More specifically, we restructure the atrous convolution at multi-scale by applying attention mechanism. Subsequently, we employ the Semantic Embedding Branch (SEB) to combine the high-level and low-level semantic information to obtain more abundant features, and use the Single Stage Headless (SSH) context module to obtain multi-scale information. Finally, we fuse the results to generate automatic high-precision mapping results. Our method has improved performance compared with other methods in the ApolloScape part of the dataset. Besides, in the database of Cityscapes, our approach has also achieved good results in semantic segmentation. Experimental results demonstrate that our proposed Lane-DeepLab can provide excellent performance in real traffic scenarios.

1. Introduction

Lane detection, a technique that helps to detect the road marked lines automatically to ensure the vehicles drive in the assigned lane rather than crash into other lanes, has played an important role in intelligent driving fields (e.g. lane departure warning system, lane keeping assist system, automatic parking assist system, driving assist system). There are many lane detection methods, and generally, they can be divided into two categories by purposes. One is for real-time detection which contains algorithms require a fast detection speed but low detection accuracy. The other is for the accurate high-definition maps restructure; Since it is applied for the navigation back-end in commercial autonomous vehicles, it requires high detection accuracy but low detection speed. Our proposed method belongs to the latter. The challenge in the high-definition maps is that most maps are manually constructed by human labelling. Since the lanes are numerous,long and thin, not only the line pixels but also the lane categories (i.e. dash line, solid line, and the signs (left arrow sign, straight arrow sign and et al.) should be marked accurately. Besides, the real lane surroundings are sophisticated,because the lane itself may damage or deteriorate caused by shadows, light, and occlusion, which bring obstruction and negative-effects on lane detection accuracy.

Therefore, in the past two decades, many researchers have devoted themselves to the field of lane detection, emerging many state-of-arts [1–5]. Due to the development of deep learning, many methods are proposed to improve the computer vision tasks performance compared to the traditional methods. The fully

* Corresponding author.
E-mail address: bkong@iim.ac.cn (B. Kong).
[1] This author contributed equally to this work and should be considered as co-first author.

https://doi.org/10.1016/j.neucom.2021.08.105
0925-2312/

convolutional network (FCN) [6] belonging to the deep learning method is for the task of semantic segmentation, which has rapidly used in a number of methods [7,8], as well as for the lane detection methods [9,10]. Another network model structure, an encoder-decoder structure [11] as well as an end-to-end architecture, is widely used in many computer vision tasks [12,13]. In this work, we propose an efficient approach based on DeepLabv3+ [14] to generate more accurate lane line detection results (see Fig. 1), named Lane-DeepLab. DeepLabv3+ is an encoder-decoder FCN-based module, the backbone of it is ResNet101 [15] for feature extracting, the atrous spatial pyramid pooling (ASPP) is employed in the encoder part. However, attention mechanism is essential in computer vision tasks. In our proposed method, we redesign the ASPP module, replacing the 1×1 convolutional layer and image pooling layer with our designed attention module. Our goal is to detect the lane for the accurate high-definition maps.

To summarize, the contributions of this study are: **(1)** We add attention mechanism to the encoder part, redesign the ASPP module, and name it attention atrous spatial pyramid pooling (ARM-ASPP). **(2)** We use feature fusion method in the decoder module to combine the high-level and low-level semantic information to obtain more abundant features. **(3)** We employ the SSH context module to get multi-scale information to obtain more robust lane detection results. Our Lane-DeepLab model, which can serve as road lane marked for accurate high-definition maps restructure, can achieve state-of-the-art lane detection performance on ApolloScape [16].

The rest of the paper is organized as follows: Section 2 presents related work on state-of-the-art semantic segmentation approaches for lane detection. Section 3 presents our proposed method. Section 4 discusses comparative experimental results and ablation studies. Finally, Section 5 concludes this work with limitations of our current approach and outlines future research directions.

2. Related work

There are two ways (lidar and camera) for detecting the road lines. Although the price of lidar is becoming cheaper and cheaper, a lot of studies [17,18] have been conducted for lane detection using lidar, but we focus on the high-definition cameras, which are mainly used in application. The road line detection environment has two categories, structured roads and unstructured roads. Structured roads generally refer to highways, urban arterial roads and other well-structured roads. These roads have clear road markings with relatively simple background environment and more obvious geometric characteristics. Therefore, the road detection problem can be simplified to the detection of lane lines or road boundaries. On the other hand, unstructured roads generally refer to less structured roads such as urban non-arterial roads and rural streets. Such roads have no lane lines or any clear road boundaries. Fig. 2 shows a better illustration between structured roads and unstructured roads.

In this paper, we focus on the structured road for lane detection, and we take lane detection as a type of image semantic segmentation task, which is led to determine whether each pixel-point on the image is a lane line pixel or not. Specifically, we discuss the issue from two aspects briefly, traditional methods and deep learning-based methods.

Traditional methods Traditional image semantic segmentation methods mainly use the visual characteristics of the image itself, such as gray value, colour, texture and other features. Niu et al. [19] proposed a lane detection method that regarded lane boundary as collection to small line segments, then used a modified Hough Transform to detect it, and clustered the all small line to detect the continuous lines. In the literature [20], the authors proposed a lane detection method for the highway, using GraphSLAM algorithm to accumulate and fuse the results to obtain the multi-lane. Song et al. [21] presented a real-time and robust lane detection method by using stereo cameras; obstacles image is obtained, then by fusing the original image to generate a low-noise top view. Later,the lane is detected through Hough Transform and vanishing points. In [22], the authors divided the road area into three regions and only used the related regions, then by using the prior information of lane colour and the width of the lane, using the Canny algorithm to detect the lane. The traditional methods for lane detection are often fast and simple, which can satisfy the real-time, but the road surroundings are fluxing due to the weather, light, vehicles; the segmentation results are not qualified with low accuracy.

Deep learning-based methods The convolutional neural network(CNN) is developed for the tasks of image classification. It can extract the features of the input images. However, the output of the image is one-dimensional information, only can predicate the images belong to what kinds of objects. Besides, the pooling layers lost many low-level features. Nevertheless, the FCN network can overcome these problems and detect more accurate two-dimensional semantic information. There are many semantic segmentation networks, FCN [6], U-net [23], ENet [24], SegNet [11], and DeepLab serials [25–28]. Based on these, many methods are proposed for lane detection. In [29], the authors proposed a fast lane detection model called LaneNet, the architecture of it is based on ENet, which has two branches, the segmentation branch produced a binary lane mask, and the embedding branch generated an embedded feature map, then these two maps are fused to classify the pixel into different lanes. Sun et al. [30] proposed a lane detection method based on atrous convolution and spatial pyramid pooling, the architecture of the proposed method has one encoder and two decoders, the encoder module is similar to DeepLabv3+, and the two decoders are similar to LaneNet, which obtained superior performance. Method [31] adopted the encoder-decoder structure model for lane detection. The encoder part is based on VGG-16. The decoder part has two branches; one can obtain semantic

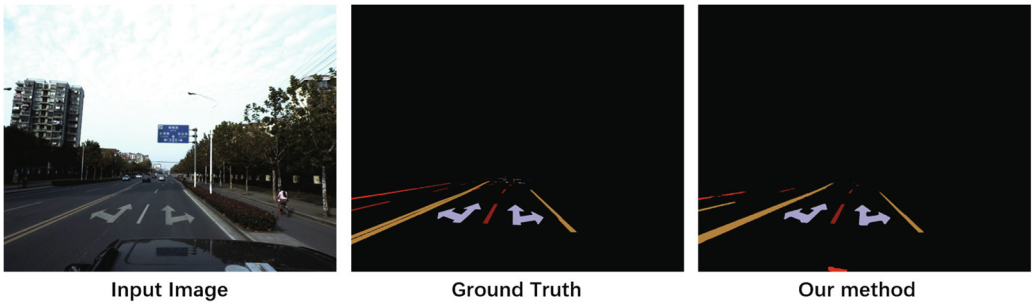

Fig. 1. Lane detection for accurate high-definition maps restructure.

Fig. 2. Structured road (left) has clear road markings, simple background environment and obvious geometric characteristics; unstructured road environment (right) lacks obvious road borders or parallel line.

segmentation map, and the other can produce instance segmentation, then cluster the two maps together to generate the final results. Hou et al. [32] proposed a light-weight lane detection model based on ENet by using self-attention distillation to train the model to achieve high accuracy of mAP in some datasets. In addition, evolutionary algorithms combined with deep learning have also yielded great successes in various fields [33–35]. Based on the evolutionary algorithm, a framework named CurveLane-NAS can capture both long-ranged coherent and accurate short-range curve information for curve lane detection [36].

From the aforementioned methods, these methods do not classify lane lines and neglect the attention mechanism,which is useful in automatic driving scenarios. Recently, attention mechanism has provided significant improvements in natural language processing and various computer vision fields. For lane detection task, since lanes are long and thin, the number of annotated lane pixels is far fewer than the background pixels, which are hard to learn for a network. In addition, attention mechanism can emphasize important spatial information of feature maps. Specifically, attention mechanism can increase the weighted information of lane line targets while reducing irrelevant information. Moreover, in our method, we adopt the attention mechanism based on DeepLapv3 + and partly select the original ApolloScape datasets with 35 categories into seven categories when training, achieving state-of-the-art results.

3. Proposed method

Our proposed method is based on DeepLabv3+, which is the best structure for semantic segmentation in the DeepLab series. The pipeline of our proposed method is shown in Fig. 3. The 'X' sign indicates element-wise multiplication. From the pipeline, the backbone of our structure is ResNet101 with 101 layers, a filter size of 3×3, a stride of 2 and the shortcut connections, which are for the feature information extracting. The output results from the ResNet101 are then fed into our proposed ARM-ASPP module, which is an attention mechanism module with three 3×3 convolutional layers owning atrous rates of 6, 12 and 18. The number of filters is 256, including the batch normalization layer and global average pooling. Subsequently, all feature maps are fused and cascaded together through a 1×1 convolutional layer to obtain high-level feature maps. The high-level feature map is up-sampled twice by bilinear interpolation to obtain an enlarged feature map. After another two rounds of upsampling, the feature map is restored to the same size as the low-level feature map extracted by the ResNet101 structure. The high-level and low-level feature maps are connected and then 3×3 convolutional layer is used to fine-tune the features. Finally, the bilinear interpolation method is used to perform upsampling by 4 to obtain the final prediction lane segmentation map. The cross-entropy loss function [37] is used, which is shown as Eq. (1):

$$L = -\sum_{i=1}^{N} y^{(i)} log\bar{y}^{(i)} + (1 - y^{(i)}) log(1 - \bar{y}^{(i)}) \tag{1}$$

3.1. Backbone

The ResNet, a convolutional neural network that is a short name for a residual network, ResNet101 belongs to the ResNet series, which has 101 layers deep, is the backbone used in our model. In the traditional convolutional neural network with the network depth increasing, a degradation problem is exposed: accuracy gets saturated and then degrades rapidly. In order to solve this problem, He et al. [15] proposed the ResNet network with short connects to achieve better performance. Our proposed network uses 3×3 Conv layers, the downsampling with the stride 2, global average pooling layer and a 1000-way fully-connected layers with softmax in the end. The ResNet has two types of short connects. One is in the satiation when the input and output are of the same dimensions, which can express as,

$$y = F(x, W_i) + x. \tag{2}$$

The other appears when the dimensions change, A) The shortcut still performs identical mapping, with extra zero entries padded with the increased dimension. B) The projection shortcut is used to match the dimension (done by 1×1 Conv) using the following formula.

$$y = F(x, W_i) + W_s x. \tag{3}$$

From the DeepLabv3+, we choose to use ResNet101 for our backbone in the model since it can achieve better results.

3.2. Encoder structure optimization

The encoder-decoder structures are gradually applied to semantic segmentation task in deep neural networks. Typically, an encoder module can capture higher semantic information and a decoder module can recover the spatial information. In the encoder

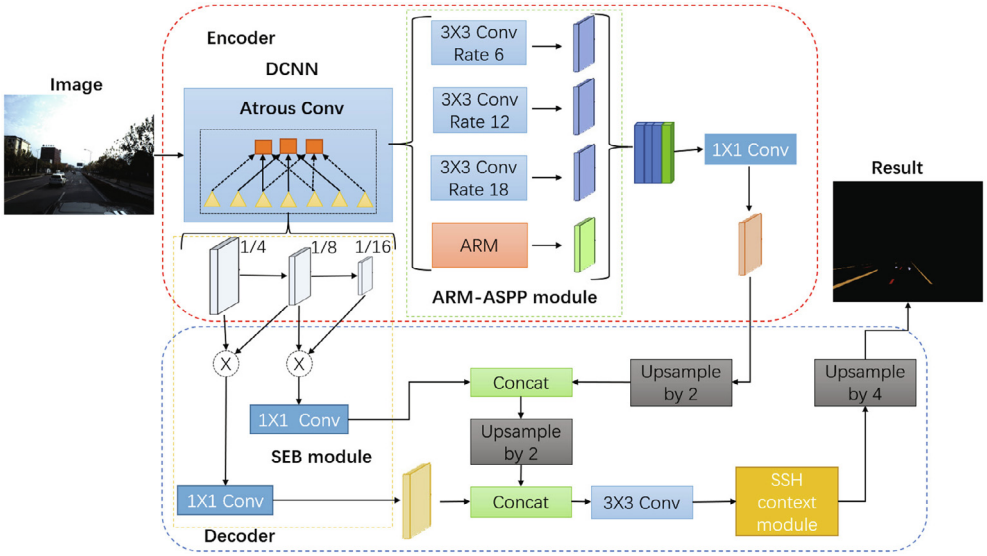

Fig. 3. The pipeline of our Lane-DeepLab.

part, atrous convolution, also known as dilated convolution, can control the resolution of feature responses. Considering two-dimensional signals, the process is denoted as Eq. (4). Here, i stands for each location on the output feature map y, and w represents a convolution filter. Atrous convolution is used in the input Resnet101 backbone feature map x. The rate parameter r stands for the stride. To change the rate value, we can adaptively modify the filter's field-of-view, because it can employ the different dilation rates in convolutional layers to capture diverse context information without increasing the number of parameters or computation. DeepLabv3+ uses the dilated convolution to preserve the spatial size of the feature map.

$$y[i] = \sum_{k} x[i + r * k]w[k] \tag{4}$$

However, the context information of the image is also crucial to predicting the high-quality semantic segmentation output. Atrous spatial pyramid pooling (ASPP) employs multiple parallel atrous convolutional layers with different rates to capture objects and context at multiple scales. DeepLabv3+ uses global average pooling to capture the global context of the image. Inspired by this, we modify the original ASPP module of DeepLabv3+. We removed the original 1×1 Conv layer, and add it to the attention refinement module (ARM). Then, we combine original ASPP module with ARM, and name it attention atrous spatial pyramid pooling (ARM-ASPP). ARM-ASPP not only uses global average pooling to capture global context but also computes an attention vector to guide the feature learning. Therefore, ARM-ASPP can refine the output features better than the original ASPP. The ARM module is addressed as follow.

The New Attention Refinement Module We implement the attention refinement model by adding a 1×1 Conv to the other sub-path. So the modified ARM has a global pool layer, a 1×1 Conv layer, a batch norm layer and a sigmoid activate function. The two sub-path results are multiplied together to obtain the attention features. See Fig. 4 for details.

3.3. Decoder structure optimization

In the decoder part, object details and spatial dimension are gradually recovered. In order to make full use of semantic information and context information, we made two optimizations for the decoder structure. One is Semantic Fusion Module (SFM), which can get the high-level and the low-level semantic information; and the other is the Single Stage Headless context module (SSH context module) [38], which can achieve multi-scale results. Some details are described below.

A. Semantic Fusion Module

In general, a common feature fusion method in the U-net segmentation framework is to express it as a residual form:

$$y_l = Upsample(y_{l+1}) + F(x_l) \tag{5}$$

Among them, x_l represents the l-th feature generated by the encoder, y_l is the fusion feature of the l-th layer. As far as we know, due to differences in semantic level and spatial resolution, the fusion of simple low-level features and high-level features may be less effective. The low-level features contain too less semantic information to restore the semantic resolution. Inspired by [39], the insight is to introduce more semantic information from high-level features into low-level features. We rewrite the fusion process described as Eq. (6).

$$y_l = Upsample(y_{l+1}) + F(x_l, x_{l+1}) \tag{6}$$

The Semantic Embedding Branch (SEB) module from the literature [39], which fuses the low-level features and high-level features to obtain the rich feature information. In our proposed SEB module (see Fig. 5), we have two branches, one is to fuse the 1/4 and 1/8 feature maps to output a 1/4 feature map by SEB module, the other is to fuse the 1/8 and 1/16 feature maps to output a 1/8 feature map by SEB module. From Fig. 3, we can see the output of the encoder part is a 1/16 feature map, then upsample by 2 is a 1/8 feature map, and concatenate with the SEB module result and upsample by 2 is a 1/4 feature map. The other branch is a 1/4 fea-

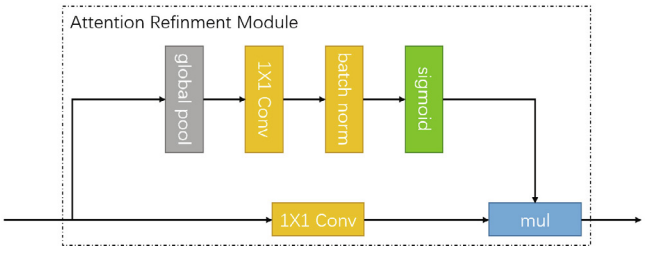

Fig. 4. The attention refinement module.

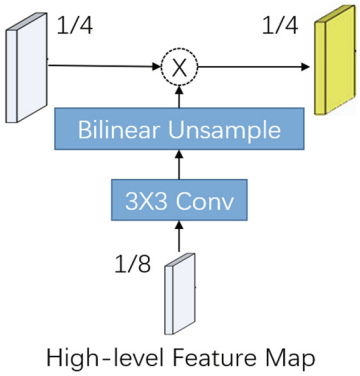

Fig. 5. The SEB module.

ture map obtained by the SEB module, then to concatenate with another 1/4 feature map, to obtain a 1/4 feature map, then through the 3×3 Conv to feed into the SSH context module, which is introduced next.

B. SSH context module

However, in the U-shape structure, some lost context information cannot be easily recovered. In general, we can enlarge the window around proposals to model context in semantic segmentation task. SSH uses simple convolutional layers to achieve the same larger window effect, leading to more efficient context modeling. It is used for face detection, through which can detect multi-scale faces. Inspired by this, we use this module to our proposed method, to detect multi-scale lines on the road to improve the robustness of our method. In this way, our proposed model provides more context information to generate a high-quality result. The structure of the SSH context module is shown as Fig. 6. To reduce the number of parameters, we deploy sequential 3×3 filters instead of larger convolutional filters. The channel of the module is set to 256.

Finally, we give the vertical structure of our proposed method, from which we can see the detailed pipeline of the proposed method. In Fig. 7, we give the encoder and decoder structure with the attention module, SFM module and SSH module.

4. Experiment results

We evaluate our network on the ApolloScape dataset [16] and Cityscapes dataset [40] respectively for the multi-category lane line semantic segmentation and general semantic segmentation tasks.

4.1. ApolloScape benchmark

1) Dataset. In order to evaluate the performance of the proposed Lane-DeepLab algorithm, experiments on the ApolloScape dataset performed. As we know, ApolloScape lane dataset is very challenging, which provides high-quality pixel-level annotations of 110 000 + frames and lane attributes, including 6 dividing markings, 4 guiding markings, 2 stopping lines, 12 turning markings and so on. In

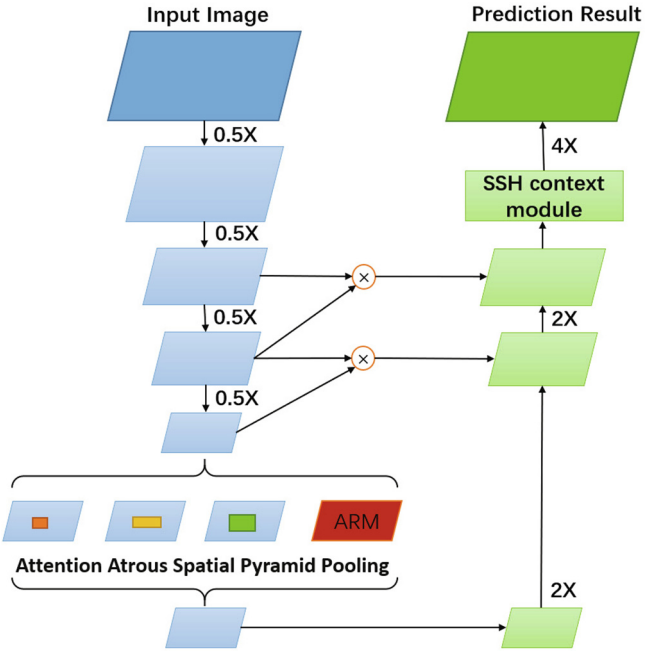

Fig. 7. The vertical structure of our proposed method.

our experiment, we use multi-class for training. ApolloScape has three separate datasets and we have chosen one of them for lane detection task. This is a relatively large data set which has around 19040 images (12400 for the training, 3320 for validation, and 3320 as test sets respectively). There are much semantic segmentation information on the road, e.g. stopping line, zebra line, single solid line, single dash line, double solid lines and so on. We select 7

Table 1
Details of lane mark labels in our dataset.

Class	category	colour
1	Dividing	
2	Guiding	
3	Stopping	
4	Zebra	
5	Turn	
6	Reduction	
7	Others	

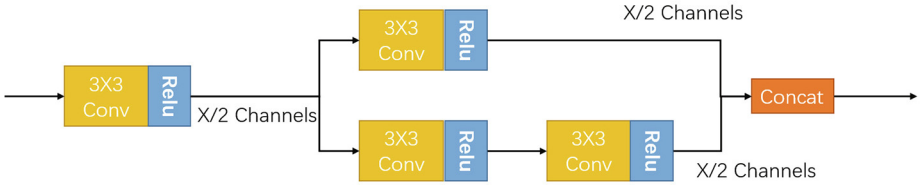

Fig. 6. The SSH context module.

Table 2
Parameters setting for training.

Learing rate	Stride	Learning rate scheduler	Loss function type	Epoch	Bitch size
0.005	16	poly	cross-entropy	25	2

Table 3
Per-class Intersection over Union (IoU) value of the ApolloScape Dataset.

category	DeepLabv3+	our method
Dividing	99.63%	**99.66%**
Guiding	83.45%	**84.10%**
Stopping	75.03%	**76.44%**
Zebra	61.86%	**65.69%**
Turn	79.13%	**80.57%**
Reduction	82.52%	**84.72%**
Others	71.83%	**75.13%**
mIoU	79.06%	**80.90%**

classes instead of 35 classes in ApolloScape dataset. The details of our selected classes are shown in Table 1.

2) Parameters Setting. Our algorithm was implemented using Intel (R) Xeon (R) Silver 4210 CPU @ 2.20GHZ × 40 for Lane-DeepLab. We keep the original image dataset resolution of 3384×2710 as input. The parameters in our experiments are learning rate, stride, learning rate scheduler, loss function type, epoch and batch size. Some parameters are determined empirically, such as learning rate scheduler, epoch and loss function type. For other parameters, we have made several attempts using GridSearchCV. Specifically, learning rate can be set to 0.005, 0.01. Stride can be set to 8, 16, 32. Batch size can be set to 2, 4. We listed the optimal combination of main parameters for the pipeline of Lane-DeepLab in Table 2.

3) Metrics. In order to evaluate the performance of the methods on the dataset, many literatures follow the PASCAL VOC [41] intersection-over-union (*IoU*) metric, which is denoted as $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels, respectively. *IoU* can be considered as the intersection of the target area predicted by the network model and the real area marked in the ground truth. In literature [42], the authors proposed a segmentation accuracy value, i.e mean Intersection over Union (*mIoU*), which is similar to *IoU*, but it can calculate the average *IoU* in different classes.Since we divided the dataset into 7 classes, *mIoU* is satisfied our preference.

We follow the literature [43], the *mIoU* is denoted as Eq. (7).

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}p_{ij} - p_{ii}} \tag{7}$$

where k represents the number of categories, there are $k+1$ categories if background is included. i represents the true value, j represents the predicted value. p_{ii} represents the total number of pixels whose category i is predicted as i, p_{ii} means pixels that are correctly classified. p_{ij} represents the total number of pixels whose category i is predicted as j, p_{ji} represents the total number of pixels whose category j is predicted as i. p_{ij} and p_{ji} indicate pixels that are incorrectly classified.

4) Evaluation Results. To the best of the authors' knowledge, no related works have been trained on the whole ApolloScape lane segmentation dataset. In our experiment, we trained the classic DeepLabv3+ model as a baseline. Meanwhile, we give the experimental results of our own method. Table 3 shows *mIoU* value and *IoU* values of our selected 7 types of lane and road markings. In our method, each category has increased in various degrees. We can get up to 80.90% *mIoU* the state-of-the-art result on our test set.

5) Ablation Study. We investigate the effects of different factors, e.g., encoder optimization: the attention mechanism (ARM-ASPP); decoder optimization: the high-level features and the low-level features fusion (SFM), and add context information (SSH). We perform experiments to obtain the *mIoU* value from our dataset of different factors (see Table 4). '√' represents that we add the module. For a clear view, we give a qauantity analysis about atrous spatial pyramid pooling (ASPP). Baseline means DeepLabv3+ model without ASPP module. As shown in the table, we can see that the performance of our proposed method, and we can get the state-of-the-art result.

Encoder Optimization: We have experimented with the proposed atrous spatial pyramid pooling (ASPP) scheme, described

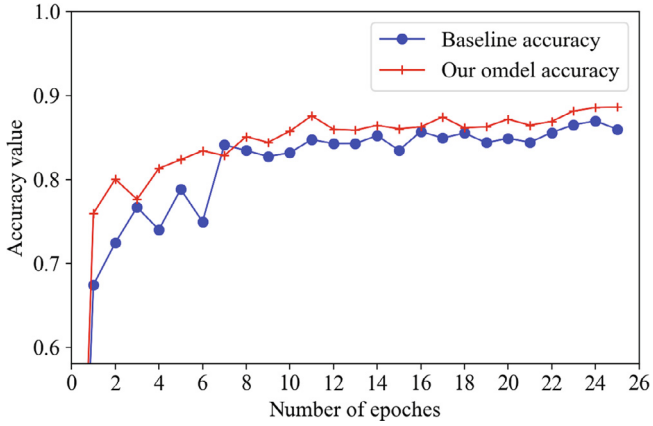

Fig. 8. The training curve of class accuracy.

Table 4
mean Intersection over Union (*mIoU*) value of the ApolloScape dataset.

Baseline	Encoder Optimization		Decoder Optimization		*mIoU*
	ASPP	ARM-ASPP	SFM	SSH	
√					78.52%
√	√				79.06%
√		√			79.29%
√	√		√		79.88%
√	√			√	80.42%
√			√	√	80.80%
√		√	√	√	**80.90%**

in Section 3.2. As shown in Table 4), our ARM-ASPP has improved the effect by 0.23%. This is due to ARM-ASPP can compute an attention vector to guide the feature learning.

Decoder Optimization: It is clear that SFM model, described in Section 3.3 improves the performance by 0.82% while the baseline model without SFM only obtains marginal gain, since the original simple combination ignores the two level information. SFM may be helpful for feature maps to embed more semantic information. Furthermore, SSH module increases nearly 1.36% effect. SSH leads to more efficient context modeling by using simple convolutional layers to achieve the same larger window effect.

Besides, we experimentally trained multiple types of lane line semantic segmentation tasks. Here, we give the class accuracy training curve of lane lines of multiple classes. As shown in Fig. 8, in our ablation experiment, we run 25 epochs during training. For the baseline model, the highest class accuracy is 86.92%, and the highest class accuracy of our method is 88.47%. In addition, we compare classic semantic segmentation network U-net [23], DeepLabv3+, our method Lane-DeepLab model. The visual figures of the different methods show as Fig. 9. Compared with other methods, our method generates a better high-quality semantic segmentation results. Moreover, as the result of visual display,

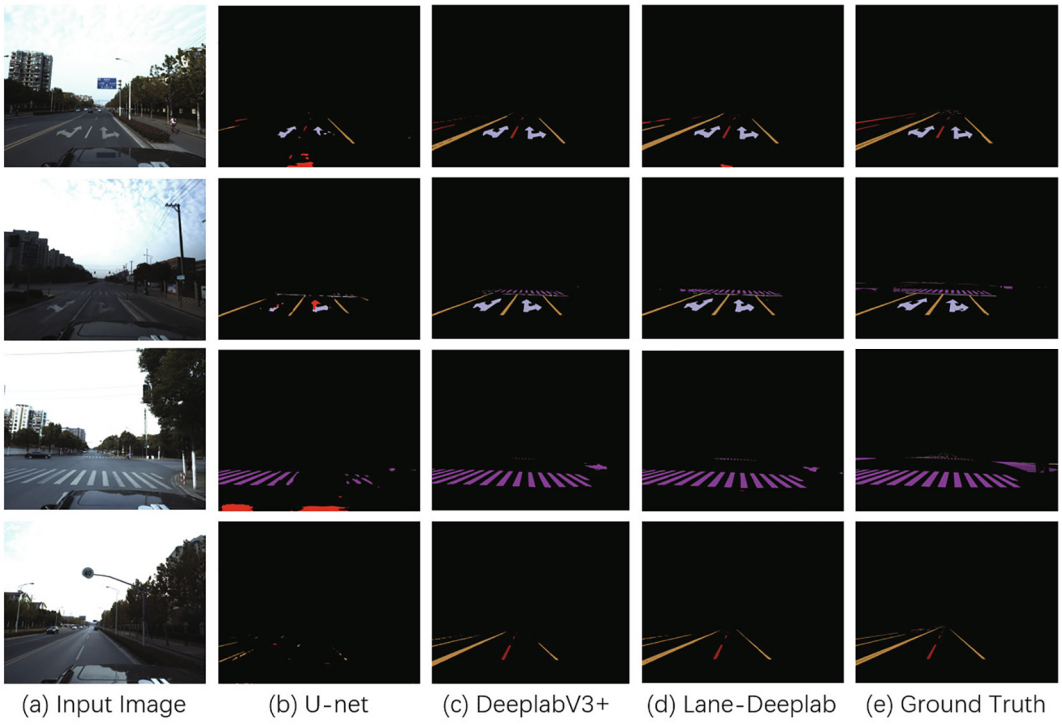

Fig. 9. Example results comparing with other methods.

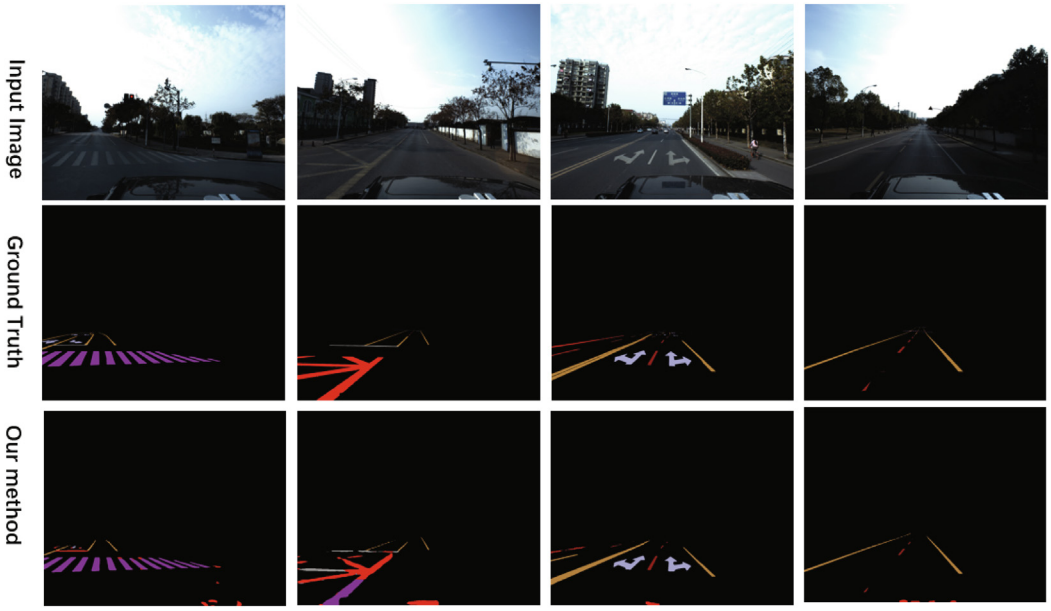

Fig. 10. Some failure cases segmentation results.

our model still performs well when dealing with changeable road types, complex environmental backgrounds and changing weather which also confirms our model is robust.

6) Failure Cases Analysis. The semantic segmentation network model has great challenges in the lane line multi-classification segmentation task, such as classification error, inaccurate target edge segmentation, lack of target details, etc. Although our model has achieved the state-of-the-art results, there are still some problems that we should explore. As shown in Fig. 10, we have summarized two common errors that are classification errors and long-distance performance degradation. As shown in the first two columns of Fig. 10, traffic intersections are prone to segmentation errors due to excessive and complex line types. In addition, at a long distance, the results of lane line segmentation show a downward trend. The lane line segmentation results are not continuous. The main reason is that the proportion of pixels in the lane line decreases at a long distance. As a result, the edges and details of the object are missing. These issues are still worthy of our continued exploration.

Table 5
Comparing with state-of-the-art methods on public Cityscapes test set without coarse annotation dataset.

Method	Backbone	*mIoU*
BiSeNetV1 [44]	ResNet18	74.8%
SwiftNet [45]	ResNet18	75.4%
DeepLabv2 [26]	ResNet101	71.4%
DUC [46]	ResNet101	77.6%
PSPNet [47]	ResNet101	78.4%
PAN [48]	ResNet101	78.6%
DeepLabv3+ [14]	Xception65	78.8%
Our method	ResNet101	**79.23**%

4.2. Cityscapes benchmark

1) Dataset. In this section, we experiment Lane-DeepLab on the Cityscapes dataset, which focuses on semantic understanding of urban street scenes from a car perspective. The dataset contains high quality pixel-level annotations of 5000 images (2975 for the training, 500 for validation, and 1525 as test sets respectively).

2) Evaluation Results. We do not use coarse data in our experiments. Our crop size of image is 768×768. We evaluate our models with 1024×2048 resolution input in the model. Backbone indicates the backbone models pre-trained on the ImageNet dataset. (see Table 5). Despite the general semantic segmentation task, our method achieves comparable results.

4.3. Real road scene results

Finally, we provided the detection results for the real dataset from our autonomous driving car, the visual images are selected from the 3 km video, we select eight images including the road lines, the signs on the roads, and zebra lines, the double solid lines and so on. From Fig. 11, we can see our Lane-DeepLab achieved superior performance on the real dataset and showed the efficiency of our proposed method.

Lane marking detection is an important part of unmanned road scene analysis which is one of the key technologies to realize modern assisted and autonomous driving systems. Accurate high-definition maps with lane markings are commonly used as the navigation backend of all commercial autonomous vehicles. Currently, most high-definition maps are manually constructed by manual labelers. Auto-assisted multi-category lane detection for generating high-definition maps for autonomous driving is a major chal-

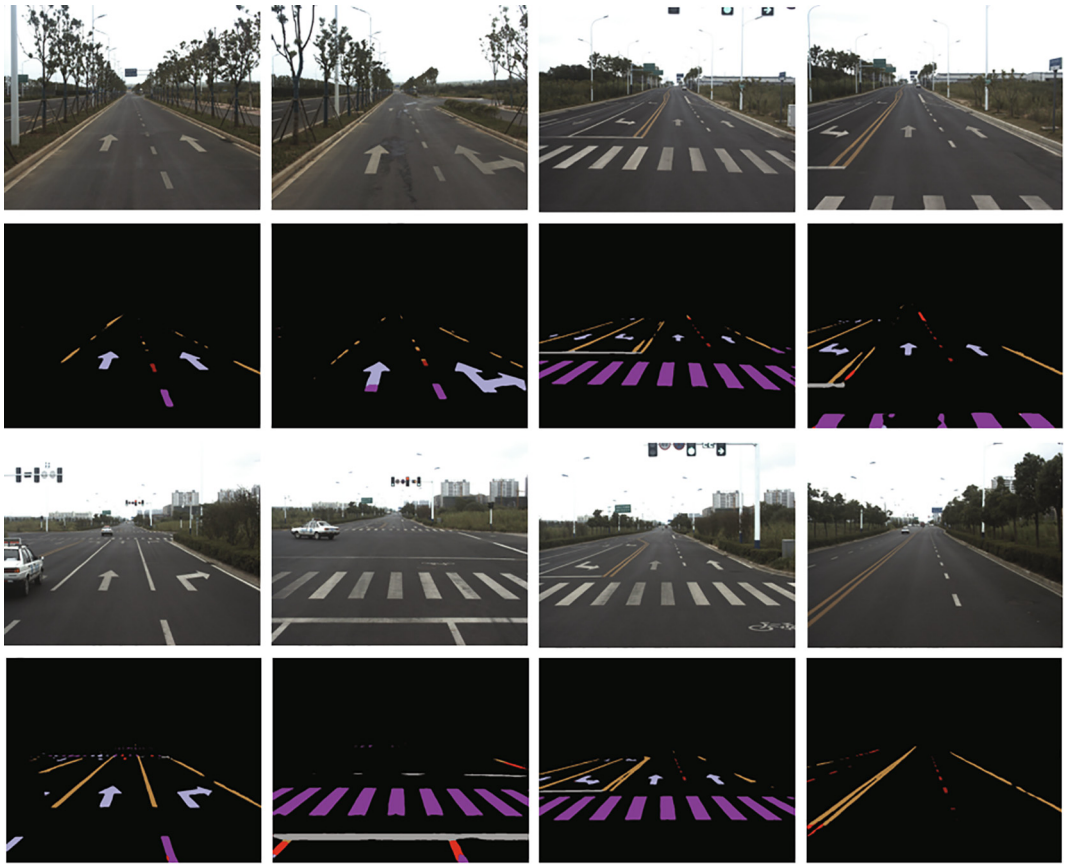

Fig. 11. Example results from real structure roads.

lenge for us. In this challenge, we extract all the basic road elements from RGB image frames, and the segmentation results can be directly used in the construction or update process of high-definition maps.

5. Conclusion

We proposed a novel Lane-DeepLab model to obtain the accuracy lane detection results for high-definition maps. The proposed method has two innovations: 1) It adds an attention module to the ASPP module to optimize the encoder structure; 2) It employs the SEB to combine the high-level and low-level semantic information to get more abundant feature.

Moreover, we summarize our contributions in the following three points: 1) For autonomous driving scenarios, most high-precision maps currently rely on manual labeling, which is time-consuming and labor-intensive. In this paper, we propose a multi-class, high-precision and offline lane line segmentation network to assist in the construction of high-precision maps for autonomous driving; 2) Most of the previous lane line network concerns timeliness and has low accuracy. And the specific category of lane lines is not given, a two-class network, which is the lane line or background. In this work, we give multi-category lane line output results. The proposed model use the attention mechanism and contextual semantics to fuse information, increase the robustness of the model, so that in the lane line environment in complex scenes and changing weather still have better results than the original model; 3) The construction of a multi-category high-precision lane line detection network in a real scene is given, and the network output results can use to construct a high-precision map in an autonomous driving scene.

From the experiments, which show the excellent performance of our Lane-DeepLab, our model can be used in high-definition maps to get high accuracy and multi-classes labels for displaying. Next, we will focus on the challenging work to get the high-definition maps via real-time.

CRediT authorship contribution statement

Jingyu Li: Conceptualization, Methodology, Software. **Fengling Jiang:** Data curation, Writing - original draft, Writing - review & editing. **Jing Yang:** Visualization, Investigation. **Bin Kong:** Project administration, Supervision. **Mandar Gogate:** Validation, Software. **Kia Dashtipour:** Writing - review & editing. **Amir Hussain:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (No.91320301), the Technological Innovation Project for New Energy and Intelligent Networked Automobile Industry of Anhui Province, the Innovation Research Institute of Robotics and Intelligent Manufacturing(CAS), the Natural Science Foundation of Education Bureau of Anhui Province (KJ2020A0111) and the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation (MMC202007).

References

[1] Z. Qin, H. Wang, X. Li, Ultra fast structure-aware deep lane detection, arXiv preprint arXiv:2004.11757..
[2] R.S. Mamidala, U. Uthkota, M.B. Shankar, A.J. Antony, A. Narasimhadhan, Dynamic approach for lane detection using google street view and cnn, in: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), IEEE, 2019, pp. 2454–2459.
[3] S.-Y. Lo, H.-M. Hang, S.-W. Chan, J.-J. Lin, Multi-class lane semantic segmentation using efficient convolutional networks, in: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6..
[4] Q. Zou, h. H. Jiang, Q. Dai, Y. Yue, L. Chen, Q. Wang, Robust lane detection from continuous driving scenes using deep neural networks, IEEE Transactions on Vehicular Technology 69 (1) (2019) 41–54..
[5] W. Van Gansbeke, B. De Brabandere, D. Neven, M. Proesmans, L. Van Gool, End-to-end lane detection through differentiable least-squares fitting, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.
[6] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[7] H. Zhou, J. Zhang, J. Lei, S. Li, D. Tu, Image semantic segmentation based on fcn-crf model, in: 2016 International Conference on Image, Vision and Computing (ICIVC), IEEE, 2016, pp. 9–14.
[8] Y. Lu, Y. Chen, D. Zhao, J. Chen, Graph-fcn for image semantic segmentation, in: International Symposium on Neural Networks, Springer, 2019, pp. 97–105.
[9] S. Zhang, A. E. Koubia, K. A. K. Mohammed, Traffic lane detection using fcn, arXiv preprint arXiv:2004.08977..
[10] N. Zakaria, H. Zamzuri, M. Ariff, M. Shapiai, S. Saruchi, N. Hassan, Fully convolutional neural network for malaysian road lane detection, International Journal of Engineering & Technology 7 (4.11) (2018) 152–155.
[11] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.
[12] X. Ou, P. Yan, Y. Zhang, B. Tu, G. Zhang, J. Wu, W. Li, Moving object detection method via resnet-18 with encoder–decoder structure in complex scenes, IEEE Access 7 (2019) 108152–108160.
[13] H. Wu, B. Zhang, A deep convolutional encoder-decoder neural network in assisting seismic horizon tracking, arXiv preprint arXiv:1804.06814..
[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV, 2018..
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[16] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, R. Yang, The apolloscape open dataset for autonomous driving and its application, IEEE Transactions on Pattern Analysis and Machine Intelligence..
[17] Y. Cui, J. Wu, H. Xu, A. Wang, Lane change identification and prediction with roadside lidar data, Optics & Laser Technology 123 (2020) 105934.
[18] J. Wu, H. Xu, J. Zhao, Automatic lane identification using the roadside lidar sensors, IEEE Intelligent Transportation Systems Magazine 12 (1) (2018) 25–34.
[19] J. Niu, J. Lu, M. Xu, P. Lv, X. Zhao, Robust lane detection using two-stage feature extraction with curve fitting, Pattern Recognition 59 (2016) 225–233.
[20] A. Abramov, C. Bayer, C. Heller, C. Loy, Multi-lane perception using feature fusion based on graphslam, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 3108–3115.
[21] W. Song, M. Fu, Y. Yang, M. Wang, X. Wang, A. Kornhauser, Real-time lane detection and forward collision warning system based on stereo vision, in: 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2017, pp. 493–498..
[22] X. Yan, Y. Li, A method of lane edge detection based on canny algorithm, in: 2017 Chinese Automation Congress (CAC), IEEE, 2017, pp. 2120–2124..
[23] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241..
[24] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147..
[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062..
[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4) (2017) 834–848.
[27] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587..
[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[29] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans, L. Van Gool, Towards end-to-end lane detection: an instance segmentation approach, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 286–291..
[30] Y. Sun, L. Wang, Y. Chen, M. Liu, Accurate lane detection with atrous convolution and spatial pyramid pooling for autonomous driving, in: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2019, pp. 642–647.
[31] L. Ding, H. Zhang, J. Xiao, C. Shu, S. Lu, A lane detection method based on semantic segmentation, Computer Modeling in Engineering & Sciences 122 (3) (2020) 1039–1053.
[32] Y. Hou, Z. Ma, C. Liu, C. C. Loy, Learning lightweight lane detection cnns by self attention distillation, arXiv preprint arXiv:1908.00821..
[33] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, X. Liu, An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips, IEEE Transactions on Nanotechnology 18 (2019) 819–829.
[34] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, X. Liu, A dynamic neighborhood-based switching particle swarm optimization algorithm, IEEE Transactions on Cybernetics..
[35] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing 425 (2021) 173–180.
[36] H. Xu, S. Wang, X. Cai, W. Zhang, X. Liang, Z. Li, Curvelane-nas: Unifying lane-sensitive architecture search and adaptive point blending, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, 2020, pp. 689–704..
[37] V. Sze, Y.-H. Chen, T.-J. Yang, J.S. Emer, Efficient processing of deep neural networks: A tutorial and survey, Proceedings of the IEEE 105 (12) (2017) 2295–2329.
[38] M. Najibi, P. Samangouei, R. Chellappa, L.S. Davis, Ssh: Single stage headless face detector, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4875–4884.
[39] Z. Zhang, X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: Enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284.
[40] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[41] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, International Journal of Computer Vision 111 (1) (2015) 98–136.
[42] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99..
[43] F. Liu, M. Fang, Semantic segmentation of underwater images based on improved deeplab, Journal of Marine Science and Engineering 8 (3) (2020) 188.
[44] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 325–341.
[45] M. Orsic, I. Kreso, P. Bevandic, S. Segvic, In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12607–12616.
[46] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1451–1460..
[47] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[48] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, arXiv preprint arXiv:1805.10180..

Jinyu Li received his B.S. degree from Anhui University of Technology, Ma'anshan, China, in 2018. He is currently working toward the Ph.D. degree at University of Science and Technology of China, Hefei, China. His research interests include computer vision, multimodal research.

Fengling Jiang received the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2021. He is currently a Lecturer at Hefei Normal University. His research interests include computer vision, saliency detection and deep learning.

Jing Yang received the Ph.D. degree from Hefei University of Technology, Hefei, China, in 2009. She is currently an associate professor at Hefei University. She mainly engaged in research on machine vision, intelligent computing, complex networks, etc.

Bin Kong received the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005. She is currently a Professor at Institute of Intelligent Machines, Chinese Academy of Sciences. Her research interests include image processing, computer vision and deep learning.

Mandar Gogate obtained his B.Eng. in Electronics (with the highest 1st Class Honours with distinction) from BITS Pilani, India, in 2016. During 2015-16, he worked as a Research assistant at ENSTA ParisTech – École Nationale SupErieure de Techniques AvancE¸es, Paris, France where he researched deep learning models for Multimodal Robotic sensor fusion and Incremental learning. He is currently a full-time IMPACT funded Ph. D. researcher in the Cognitive Big Data and Cybersecurity (CogBID) Research Lab at Edinburgh Napier University, in Scotland, UK. He is working on multimodal big data fusion using deep and incremental learning, in collaboration with the University of Oxford, for solving a number of challenging real-world problems, including cybersecurity, speech separation, and sentiment and opinion mining and 5G-IoT applications.

Kia Dashtipour obtained his Honour Degree from Edinbrugh Napier University, UK, 2011. During 2015-2017 he was doing a Master in Computer Advanced System Development in University West of Scotland. He is currently a full-time Ph.D. researcher in the University of Stirling, Scotland, UK. He is working on sentiment analysis using deep learning.

Amir Hussain obtained his B.Eng. (with the highest 1st Class Honors) and Ph.D. (in novel neural network architectures and algorithms) from the University of Strathclyde in Glasgow, Scotland, UK, in 1992 and 1997 respectively. Following postdocoral and academic positions at the University of West of Scotland (1996-98), University of Dundee (1998-2000), and University of Stirling (2000-2018) respectively, He joined Edinburgh Napier University, in Scotland, UK, in 2018, currently as Professor of Computing Science, and founding Director of the Cognitive Signal Image and Control Processing Research (COSIPRA) Big Data and Cybersecurity (CogBiD) Research Laboratory at the University of Stirling in Scotland, UK. His research interests are inter crossdisciplinary and industry focussed, and include secure and context-aware 5G-IoT driven AI, and multi-modal cognitive and sentic computing techniques and applications. He has published over more than 270 400 papers, including over a dozen books and around 8150 journal papers. He has led major national, European and international projects and supervised more than 30 Ph.D. students, He is the founding Editor-in-Chief of the two leading journals: Cognitive Computation (Springer NatureNeuroscience, USA), and BMC Big Data Analytics (BioMed Central),; and Chief-Editor of the Springer Book Series on Socio-Affective Computing, and Springer Briefs on Cognitive Computation Trends. He is anhas been appointed invited Associate Editor of several prestigeous journals, including the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Emerging Topics in Computational Intelligence, and (Elsevier) Information Fusion. He is a member of several Vice-Chair of the Emergent Technologies Technical Committees of the IEEE Computational Intelligence Society (CIS), and founding publications co-Chair of the IINNS Big Data Section and its annual INNS Conference on Big Data, and Chapter Chair of the IEEE UK and RI Industry Applications Society.