Deep Layer Aggregation with Cross Attention for Lane Detection

Ming Lu\textsuperscript{,a} Duan Liu\textsuperscript{,b} Yongteng Sun\textsuperscript{,c} Hao Duan\textsuperscript{,d}

\textsuperscript{a}School of Information Electrical and Engineering, Hunan University of Science and Technology
Xiangtan 411201

\textsuperscript{b}15377321599, China
37812588@qq.com

\textsuperscript{c}18373286630, China
761514724@qq.com

\textsuperscript{d}15050575922, China
281745148@qq.com

\textsuperscript{d}18482165878, China
289593063@qq.com

ABSTRACT. As an important part of driverless vehicle research and high-precision map making, lane detection technology needs to be improved urgently. However, it is difficult for current networks to obtain long-distance semantic information, which leads to the confusion of categories in narrow lanes. This paper adds cross attention based on deep layer aggregation to obtain long-distance semantic information, and constructs time series filter to filter in time domain, the method is simple and robust. Our encoder is based on resnet-50. According to the experimental results, slightly changing resnet-50 can achieve better results. Our method is evaluated on Baidu ApolloSpace land segmentation dataset, increases 3.4% relative to DeeplabV3+, and cross attention with time series filter contribute more than 1% mIoU accuracy.

1. INTRODUCTION

In the field of autonomous vehicle research, high-precision map with lane attributes is a key link of commercial driverless vehicle. Up to now, most high-precision maps are completed by labelling manually. Manual labelling is not only inefficient, but also costly and unsuitable for large-scale commercial applications. In the production process of high-precision map, how to segment different types of lane lines from images is very important [1]. At the same time, accurate segmentation of lane lines will also provide help for future updates of high-precision maps.

At present, many scholars have conducted in-depth research on lane detection and proposed a series of detection methods. The lane lines detection based on computer vision can be divided into artificial feature-based algorithm and depth neural network-based algorithm. Artificial feature-based algorithm [2-4] relies on highly specialized artificial features and heuristic combination to identify lane lines. Literature [2] proposes a lane detection algorithm based on ROI model. In ROI, Sober edge detection operator and improved Hough transform are used to detect possible lanes. Literature [3] visualizes the straight line based on the parallel coordinates. Thus, it gets a faster lane detection algorithm than traditional Hough detection. The algorithm based on deep neural network [5-7], which uses deep neural
network to extract lane features, has highly accuracy and robustness. Literature [5] proposes an end-to-end lane detection model, which includes LaneNet and H-Net. LaneNet is responsible for segmentation of lanes in road images and regression fitting of all pixels in the same lane, and H-Net is responsible for predicting projection matrix. Literature [6] proposes a novel network, Spatial CNN, which transfers information in rows and columns of images. It is especially suitable for detecting targets with long-distance continuous shape or large-scale targets. S-CNN has a strong spatial relationship, but it has a large amount of computation. In this paper, we attempt to improve the accuracy of land detection, strengthen the use of long-distance semantic information to eliminate category confusion in narrow lane lines. We propose a new model which combines deep layers aggregation [8] and across attention [9] mechanism for lane detection. Our model can fuse on different scales feature map, and get long-range contextual information. We validate the proposed method through experiments on ApolloSpace land segmentation dataset. The main contributions of this paper are as follows:

1. We propose a new end-to-end model for land detection, which can fully use different layer feature map information and get long-range contextual information.
2. We use the cross-attention mechanism instead of the S-CNN method to obtain long-distance semantic information, thus effectively preventing lane line category confusion.
3. Innovative time-series filter module, which uses simple and efficient timing information to solve specific problems encountered in lane detection, thereby improving by nearly 1.6% mIoU accuracy.

2. Related Work

In recent years, many researchers are considering how to better model Context information. Context information refers to consider the feature information of adjacent pixels for each pixel to be segmented, or even the feature information of more distant pixels, also known as long-range dependency.

PSP-Net [10] and DenseASPP [11] combines multi-scale features to effectively enlarge the receptive field of the convolution layers for segmentation tasks. Deformable CNNs [12] achieves the similar outcome by further learning offsets for the convolution sampling locations. Squeeze-and extension Networks [13] (SE-Net) uses global average pooling to incorporate an image-level descriptor at every stage. Nonlocal Networks [14], self-attention Mechanism [15] and Double Attention Networks (A2-Net) [16] tries to deliver long-range information from one location to another.

The object of lane segmentation is relatively narrow, and the semantic information only exists around the lane lines, which leads to the loss of semantics in network communication. S-CNN’s scheme of propagating between rows and columns effectively solves the problem of long-distance semantic loss, but it has the disadvantage of large amount of computation, and it is difficult to achieve real time operation. DeepLabV3 + [17] adds decoder module on the basis of Deeplab V3 [21], which makes the edge detection more accurate. However, the speed of complex network detection is slow, and it is difficult to obtain long-distance semantic information.

3. Approach

In this section, we give the details of DLAnet with Cross attention for lane detection. At first, the DLAnet-CA, which consists of deep layer aggregation network and a cross attention, is used to capture long-range contextual information in horizontal and vertical direction. Then the cross attention module is introduced. At the last, in order to filter inference of the eighth class in ApolloScape lane detection dataset, we design a time series filter as a post-processing method.
Figure 1. The DLAnet with cross attention model (DLAnet-CA), which consists of encoder module, attention module, decoder module and time series filter module (TSF). The attention model only uses to capture long-range contextual information in high feature map. The TSF module uses time series DLAnet-CA result to get final result.

3.1 Network architecture
The structure of proposed DLA-CA network is shown in Figure 1. Our decoder module is similar to Resnet-50, as shown in Figure 2. In contrast to the application of large filters (7*7) for the first convolution layer in other studies [18], three small convolution filters (3*3) are cascaded to construct a deeper network with the same receptive field. The resconv1, resconv2, resconv3 and resconv4 are the basic residual blocks [18] for learning the feature extraction. The cross attention module, as shown in Figure 1, is then applied to gather long range contextual information from resconv3. In decoder model, our structure is based on [19], which designs architecture to fuse information iteratively and hierarchically. Finally, TSF module is designed to filter class 8 inference, which provides time series information efficiently.

Figure 2. ResNet-50-change: in Resconv 0, the convolution of \{3 \times 3 \times 64\} \times 3 is used to replace the convolution of 7 \times 7 \times 64.

3.2 Cross Attention Module
The lane is very long, and the size of the input image is 1024*1024, the output of the resconv3 size is 64*64, 3*3 convolution for such a dimension can be observed only partial information, resulting in lane local interference resistant ability is very weak. Based on the above problems, in order to model long-range contextual dependencies over local feature representations using lightweight computation and memory [9], we introduce a cross attention module.
Cross attention module takes feature maps $L$ as input, which extracted from resconv3 and resconv4, output feature maps $L'$, for any position $\mu$ at feature map $L'$ and any position $\theta$ at feature map $H$. As shown in Figure 3, feature maps $Q$ and $K$, we further generate attention maps $A \in \mathbb{R}^{(H-W+1)\times W}$ via affinity operation, where $W$ and $H$ are the width and height of image $L$. So the relationship between $L'$ and $L$ is shown as:

$$L' \leftarrow f(A, \mu, \theta)L \quad (1)$$

At each position $\mu$ in spatial dimension of feature maps $Q$, we can get a vector $Q_{\mu} \in \mathbb{R}^C$. Meanwhile, we can obtain the set $\Omega_{\mu}$ by extracting feature vectors from $K$ which are in the same row or column with position $\mu$. Thus, $\Omega_{\mu} \in \mathbb{R}^{(H-W+1)\times C}$, $\Omega_{\mu,i} \in \mathbb{R}^C$ is $i$th element of $\Omega_{\mu}$. The affinity operation is defined as follows:

$$d_{\mu,i} = Q_{\mu}^{T} \Omega_{\mu,i} \quad (2)$$

In which $d_{\mu,i} \in D$ denotes the degree of correlation between feature $Q_{\mu}$ and $\Omega_{\mu,i}$, $i = [1,...,|\Omega_{\mu}|]$, $D \in \mathbb{R}^{(H-W+1)\times W\times H}$. Then, we apply a softmax layer on $D$ along the channel dimension to calculate the attention map. Although the attention mechanism based on Non-local proposed in [9] can obtain long-range context information, it is easily cause context confusion, so we add another branch Short cut to concatenate the input and $L'$ feature maps.

### 3.3 Time Series Filter Module

According to the classification standard of ApolloScape lane detection dataset, the reduction lane has obvious characteristics as shown in Figure 4 (c), it can be clearly distinguished in Figure 4 (a), but the reduction lane in Figure 4 (b) is easily confused with the driving lane, that is to say, the different situations contained in the reduction lane are easily confused with other categories.

The TSF module we designed solves the problem that reduction lane is easily confused with other categories by using time series information. Reduction lane has distinct entry and exit frames in time series, Figure 4 is a sign to enter the reduction lane. At the same time, in order to enhance the robustness of the algorithm, we first transform the image to bird’s-eye view, as Figure 5. (a) represents entering a time frame with reduction lane type, and (b) represents exiting a time frame with reduction lane type. Thus, we can construct a filter of time series by simple algorithm to filter reduction lane type frames. The algorithm flow is shown in Figure 6.
Figure 4. An example of reduction lane type in data sets: (a) Entering the reduction lane section; (b) In the reduction lane section; (c) Characteristic map of deceleration lane.

Figure 5. Perspective transformation: (a) Entering the reduction lane section; (b) Exiting the reduction lane section.

Figure 6. The flow chart of TSF.
4. Experiment

4.1 Datasets and evaluation metrics
The ApolloSpace land segmentation dataset (ALSD): the ApolloSpace land segmentation dataset is an accurate High Definition (HD) Maps with lane markings usually serves as the back-end for all commercial auto-drive vehicles for navigation. This large-scale dataset contains a diverse set of stereo video sequences recorded in street scenes from different cities, with high quality pixel-level annotations of 110 000+ frames. Each image size is 3384*1710. For ApolloSpace land segmentation dataset, the evaluation metrics is pixel-wise mean intersection over union (mIoU), IoU is defined as:

\[
\text{IoU} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive} + \text{FalseNegative}}
\]

Therefore, the mean IOU is calculated by averaging all classes. The data set of land segmentation is divided into eight grades, and mIoU is the average of eight grades.

4.2 Implementation details
In the entire experiment, the original image is transformed to get the bird's-eye view. The sky part is cut off. After the perspective transformation (PT), the image size is 1024*1024, and then the image is put into the DLA-CA model. In training phase, we used BCE (binary cross entropy) + dice coefficient loss as loss function and chose Adam as our optimizer. For the first 3 epochs, we decreased learning rate by a factor of two, when validation loss did not improve for two epochs. Then for another 6 epochs we used cyclic learning rate, oscillating between 1e-3 and 5e-5 on schedule: 5e-5, 3e-4, 1e-3, 3e-4, 5e-5, with 1 epoch in each cycle. We employ 2*1080TI GPUs for training and the batch size is 6.

Table 1. The comparison of accuracy among different classes.

| Method            | Driving | Guiding | Stopping | Chevron | Parking | Turn | Reduction | mIoU |
|-------------------|---------|---------|----------|---------|---------|------|-----------|------|
| Unet[20]          | 0.5453  | 0.5312  | 0.5007   | 0.3130  | 0.6775  | 0.6810| 0.3824    | 0.576|
| S-CNN[6]          | 0.3753  | 0.5762  | 0.3090   | 0.3217  | 0.6763  | 0.6650| 0.4536    | 0.594|
| LaneNet+Hnet[5]   | 0.5842  | 0.5814  | 0.4966   | 0.3297  | 0.6683  | 0.6570| 0.3629    | 0.590|
| DeeplabV3+[17]    | 0.5617  | 0.5594  | 0.5305   | 0.3273  | 0.7163  | 0.7006| 0.4208    | 0.601|
| DLANet            | 0.5705  | 0.5750  | 0.5178   | 0.3362  | 0.6927  | 0.6526| 0.4328    | 0.597|
| DLANet-CA         | 0.5797  | 0.5722  | 0.5222   | 0.4100  | 0.6930  | 0.6531| 0.4520    | 0.609|
| DLANet+PT+CA+TSF  | 0.5989  | 0.5899  | 0.5350   | 0.4370  | 0.7095  | 0.6572| 0.4576    | 0.635|


Figure 7. Test results of different models on ALSD datasets.

Table 2. Contrast Scoring of Different Modules in Network.

| Model          | PT   | CA   | TSF   | Backbone         | Mean_IoU |
|----------------|------|------|-------|------------------|----------|
| ResUnet        | ✔    | ✔    | ✔     | ResNet-50        | 0.599    |
|                | ✔    | ✔    | ✔     |                   | 0.610    |
|                | ✔    | ✔    | ✔     | ResNet-50        | 0.621    |
| DeeplabV3+     | ✔    | ✔    | ✔     | ResNet-50        | 0.601    |
|                | ✔    | ✔    | ✔     |                   | 0.613    |
|                | ✔    | ✔    | ✔     | ResNet-50        | 0.626    |
| DLAnet         | ✔    | ✔    | ✔     | ResNet-50        | 0.608    |
|                | ✔    | ✔    | ✔     |                   | 0.619    |
|                | ✔    | ✔    | ✔     | ResNet-50-change | 0.630    |
|                | ✔    | ✔    | ✔     |                   | 0.635    |

4.3 Performance on ALSD

The experimental results are shown in Figure 7 and Table 1. Unet is the benchmark of the experiment. S-CNN has better scores in driving lane and guiding lane because of adding information transfer in rows. However, due to the lack of multi-scale network design, the fitting effect of some shorter turn lane types is not good. LaneNet+Hnet uses multi-task method to segment. Because the segmentation structure is not advanced enough, the segmentation structure is not ideal. However, the perspective transformation
method makes the inter-class interval more obvious and achieves good scores on driving lane and guiding lane. DeeplabV3 + is a well-designed network. It is difficult to obtain long-distance semantic information. It has poor segmentation effect in long-distance objects such as driving lane, but it has achieved very good results in compact objects such as turn lane. We designed DLAnet-CA network to fuse features of different scales. After adding cross attentions, the performance of DLAnet-CA network in driving lane and guiding lane categories increased by nearly 1% mIoU.

4.4 Ablation studies

4.4.1 The effect of perspective transformation (PT)
In DeeplabV3+, the perspective transformation increased by 1.2% mIoU, and increased by 1% mIoU in DLAnet-CA, as shown in Table 2. It is proved that perspective transformation can widen the distance between classes in lane detection, which makes the network more easily fit.

4.4.2 The effect of cross attention module (CA)
To verify the validity of adding CA module, we compare DLAnet without CA module with DLAnet with CA module. From this comparison, we can see that the addition of CA module contributes 1% to the IoU score of the network. We add CA module to the top and penultimate layers of the network, and use residual mode to add input and cross-attention feature map, so that we can retain the original information and add long-distance semantic information. When we design the network, we find that only adding CA module in the last layer will contribute less than 0.5% to the overall accuracy. We find that the single-layer CA module can find the semantic relationship of a single narrow road, thus eliminating confusion. However, because the road is usually narrow and occupies a small proportion in the top-level feature map, we try to add CA module in the penultimate layer, and find that the effect is more obvious, 0.4% higher than that of the single layer, and 1% higher than that of DLA network without CA module.

4.4.3 The effect of time series filter module (TSF)
TSF is actually a post-processing method. The goal is to introduce temporal information. In order to ensure the robustness of the algorithm, we transform the original image into a bird's-eye view, and then use Unet, DeeplabV3 + and DLA networks to segment lanes. The features of this kind of image have very obvious features of entering and going out the road, but the characteristics of the lane are not clear, so the final result not only leads to recognition errors due to the ambiguity between classes, but also leads to confusion on other types of roads. In order to solve these problems, we propose a simple and intuitive way to solve these problems, the results is that the IOU score increased by 1.6% and the single-class score increased from 0.456 to 0.536.

5. CONCLUSION
In this paper, we propose a new neural network architecture DLAnet-CA for lane detection, which innovatively integrate deep layers aggregation and cross attention model, aiming to obtain accurate locations and perceive of the entire lane. The proposed Time Series Filter Module can effectively classify reduction lane based on time series information, and filter out other kinds of interference. Extensive experimental results show that the proposed algorithm can extract lane lines more quickly and accurately, comparing with previous models.

ACKNOWLEDGMENT
This work was supported by the National Natural Science Foundation of China (Grant No.61603132, 61672226).
REFERENCES

[1] Rui S, Hui C, Zhiguang X, et al. 2017. Lane detection algorithm based on geometric moment sampling. Scientia Sinica, 47(4): 455-467. DOI= https://doi.org/10.1360/N11 2016-00174

[2] QIAN Jide, CHEN Bin, QIAN Jiye, et al. 2018. Fast Lane Detection Algorithm Based on Region of Interest Model. Journal of University of Electronic Science and Technology of China, 47(3): 356-361.

[3] WANG Xuchen, LU Xinchen, ZHANG Hengsheng, et al. 2018. A Lane Detection Method Based on Parallel Coordinate System. Journal of University of Electronic Science and Technology of China, 47(3): 362-367.

[4] SON J, YOO H , KIM S, et al. 2015. Real-time illumination invariant lane detection for lane departure warning system. Expert Systems with Applications. 42(4): 1816–1824. DOI= https://doi.org/10.1016/j.eswa.2014.10.024

[5] Neven D, De Brabandere B, Georgoulis S, et al. 2018. Towards End-to-End Lane Detection: an Instance Segmentation Approach. In: 2018 IEEE Intelligent Vehicles Symposium (IV), China.

[6] Pan X, Shi J, Luo P, et al. 2017. Spatial As Deep: Spatial CNN for Traffic Scene Understanding.

[7] GURGHIAN A, KODURI T, BAILUR S V, et al. 2016. Deepplanes: End-to-end lane position estimation using deep neural networksa. Proceedings of the Computer Vision and Pattern Recognition. Las Vegas: IEEE Conference, pp. 38-45. DOI=10.1109/CVPRW.2016.12

[8] Yu F, Wang D, Shelhamer E, et al. 2017. Deep Layer Aggregation. In: Computer Vision and Pattern Recognition (CVPR) 2018.

[9] Huang Z, Wang X, Huang L, et al. 2018. CCNet: Criss-Cross Attention for Semantic Segmentation. In: Computer Vision and Pattern Recognition.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. 2017. Pyramid scene parsing network. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on.IEEE. pp. 6230–6239. DOI=10.1109/CVPR.2017.660

[11] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. 2018. Denseaspp for semantic segmentation in street scenes. In: Computer Vision and Pattern Recognition.

[12] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. 2017. Deformable convolutional networks. In: ICCV. DOI=10.1109/ICCV.2017.89

[13] J. Hu, L. Shen, and G. Sun. 2018. Squeeze-and-excitation networks. In: CVPR. DOI=10.1109/TPAMI.2019.2913372

[14] X. Wang, R. Girshick, A. Gupta, and K. He. 2018. Non-local neural networks. In: Computer Vision and Pattern Recognition (CVPR).

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 6000–6010.

[16] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng. 2018. A2-nets: Double attention networks. In: NIPS.

[17] Chen L C, Papandreou G, Schroff F, et al. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. In: Computer Vision and Pattern Recognition.

[18] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778. DOI=10.1109/CVPR.2016.90

[19] Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018. Deep layer aggregation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2403-2412.

[20] Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmen-tation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, Proceedings, Part III (pp.234-241). DOI=10.1007/978-3-319-24574-4_28

[21] Chen L C, Papandreou G, Schroff F, et al. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.