Engineering Applications of Artificial Intelligence 127 (2024) 107260

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Image semantic segmentation approach based on DeepLabV3 plus network with an attention mechanism

Yanyan Liu [a], Xiaotian Bai [b], Jiafei Wang [a], Guoning Li [b,**], Jin Li [c,*], Zengming Lv [b]

[a] Department of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun, 130022, China
[b] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences (CIOMP), Changchun, 130033, China
[c] School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, 100191, China

A B S T R A C T

Image semantic segmentation is a technique that distinguishes different kinds of things in an image by assigning a label to each point in a target category based on its "semantics". The Deeplabv3+ image semantic segmentation method currently in use has high computational complexity and large memory consumption, making it difficult to deploy on embedded platforms with limited computational power. When extracting image feature information, Deeplabv3+ struggles to fully utilize multiscale information. This can result in a loss of detailed information and damage to segmentation accuracy. An improved image semantic segmentation method based on the DeepLabv3+ network is proposed, with the lightweight MobileNetv2 serving as the model's backbone. The ECAnet channel attention mechanism is applied to low-level features, reducing computational complexity and improving target boundary clarity. The polarized self-attention mechanism is introduced after the ASPP module to improve the spatial feature representation of the feature map. Validated on the VOC2012 dataset, the experimental results indicate that the improved model achieved an MIoU of 69.29% and a mAP of 80.41%, which can predict finer semantic segmentation results and effectively optimize the model complexity and segmentation accuracy.

1. Introduction

The emergence of artificial intelligence (AI) has dramatically changed every aspect of our lives. The concept of semantic segmentation is easy to understand. When people see a picture, it is easy to understand the content of the picture. Semantic segmentation allows the machine to understand the content of the picture. The application, in reality, is also increasingly extensive, for example, scene recognition of automatic driving technology, for surgical navigation in the field of medical image segmentation, and advertising recommendations. The wide application of image semantic segmentation has high practical value (Iftikhar et al., 2022, 2023).

To date, many different semantic segmentation algorithms have been proposed, including traditional and deep learning semantic segmentation. From the traditional methods, such as threshold (Otsu, 1979), histogram-based bundling, region-grow (Nock and Nielsen, 2004), k-means clustering (Dhanachandra et al., 2015), and watersheds (Najman et al., 1994), to more advanced algorithms such as active contours (Dhanachandra et al., 2015), graph cut (Najman et al., 1994), conditional and Markov random fields (Kass et al., 2004), and sparsity-based methods (Boykov et al., 2001; Plath et al., 2009). To compensate for the lack of traditional methods, the semantic segmentation methods of deep learning mainly have two types of classification from the model structure: based on information fusion and based on coder-decoder(Minaee et al., 2021). Based on the information fusion method, the model utilization is improved by increasing the number of layers of the network (Starck et al., 2005; Minaee et al., 2017). The representative algorithms include the full convolutional network (FCN) algorithm and a series of improved algorithms (Biao et al., 2018), such as FCN–32S, FCN–16S, and FCN–8S. Based on the coder-decoder method (Liu et al., 2018; Fu et al., 2022), the accuracy of the network is improved by adopting different backbone network forms and pyramid pooling modules. The representative algorithms include the pyramid scene parsing network (PSPNet)(Sun and Wang, 2018) and DeepLabv series. The current method based on Deeplabv3+ has high computational complexity and large memory consumption, and it is difficult to deploy on embedded platforms with limited computational power. Deeplabv3+ cannot fully utilize the multiscale information when extracting the image feature information, and it is easy to cause the loss of detail information and lead to damage of segmentation accuracy. To further improve the ability

* Corresponding author.
** Corresponding author.
E-mail addresses: liuyanyan@cust.edu.cn (Y. Liu), baixiaotian19@mails.ucas.edu.cn (X. Bai), 2020100813@mails.cust.edu.cn (J. Wang), liguoning@ciomp.ac.cn (G. Li), jl11269@buaa.edu.cn (J. Li), lzm232@163.com (Z. Lv).

https://doi.org/10.1016/j.engappai.2023.107260
Received 3 May 2023; Received in revised form 15 September 2023; Accepted 3 October 2023
Available online 10 October 2023

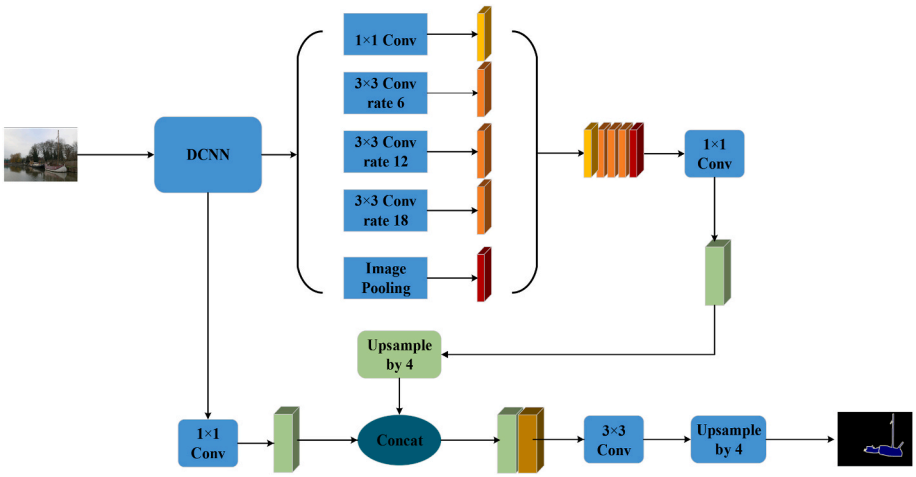


Fig. 1. Deeplabv3 plus model.

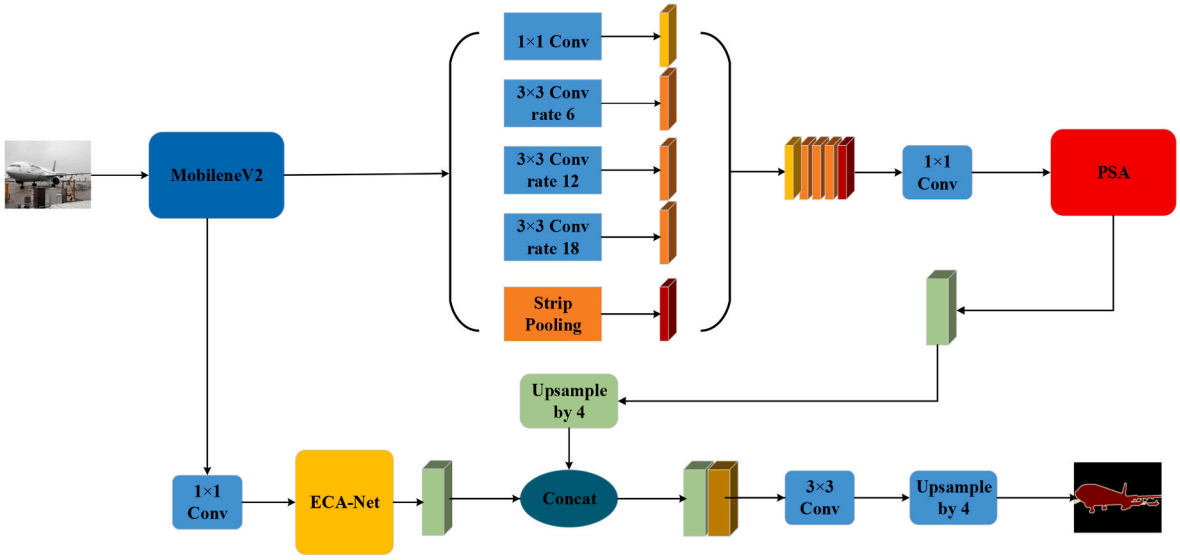


Fig. 2. Improved DeepLabv3 plus.

of the DeepLabv3 plus network to obtain key category information, improvements are mainly made based on DeepLabv3 plus. The main contributions of this paper are summarized as follows.

1. The DeepLabv3+ network is improved to make it suitable to fit the needs of realistic scenarios. The original feature extraction network parameter amount is too large, and the model adopts the lightweight MobileNetV2 as the backbone network, based on which it is further optimized to solve the problems of spatial detail loss and insufficient feature extraction.
2. In DeepLabv3+, the polarized self-attention mechanism (PSA-P, PSA-S) is added after the ASPP module to increase the ability of the feature map to extract detailed information to improve the accuracy performance of semantic segmentation. A channel attention mechanism (ECA-Net) is added after the MobileNetv2 low-level features to recover clearer segmentation boundaries.
3. Stripe pooling is utilized in the ASPP module instead of the original global average pooling to effectively capture long-range dependencies, and hybrid pooling is utilized instead of the original global average pooling to effectively capture short-range and long-range interdependencies between different locations, thus improving the efficiency and reliability of the system.

2. DeepLabv3 plus network

The DeepLabv3 plus network (Yang et al., 2020) is shown in Fig. 1. The role of the backbone network is to extract feature semantic information (Zhao et al., 2017). The function of ASPP is to extract feature information from the backbone network again to obtain sufficient feature information. DCNN is generally a deep convolutional neural network. The ASPP module is mainly composed of 5 parts, 1×1 Convolution and void ratio are 6, 12, and 18 times, respectively 3×3 Convolution and global average pooling. These five parts are in parallel and together constitute the ASPP part. Backbone network low-level feature postaccess 1×1. The convolution and ASPP are then connected to the 4 times downsampling part for feature fusion and then connected to the 3×3 convolution and 4 times downsampling to recover the size of the image.

3. Improved DeepLabv3 plus network

The DeepLabv3 plus model is taken as the main body for improvement. In image semantic segmentation based on the DeepLabv3 plus network, this paper uses lightweight MobileNetV2 as the backbone network. Then, ASPP is used to extract multiscale information from the

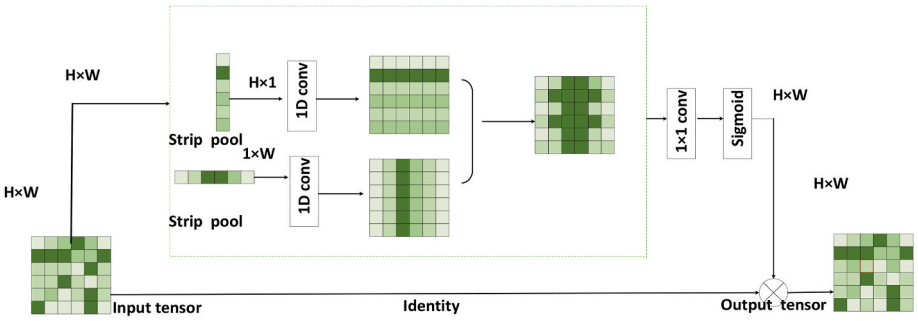


Fig. 3. Structure of strip pooling.

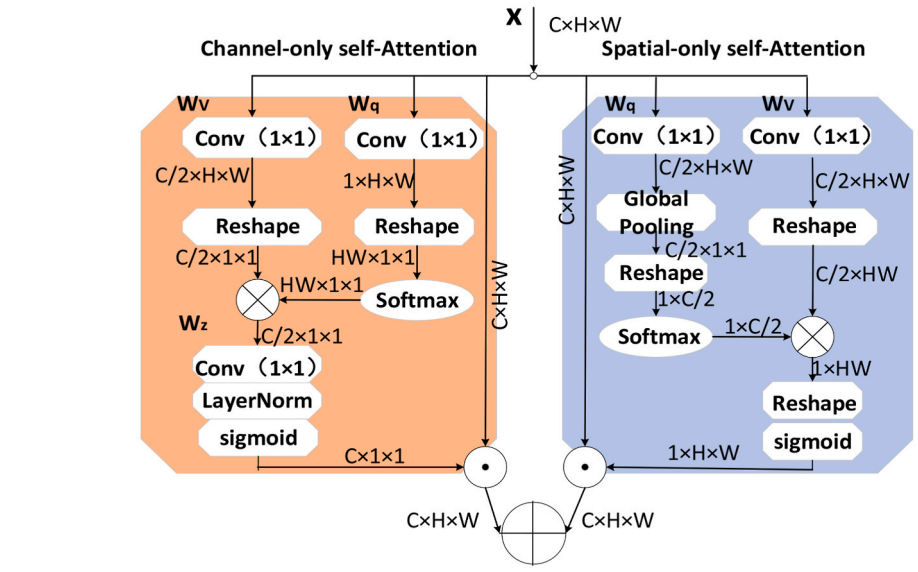


Fig. 4. PSA in parallel.

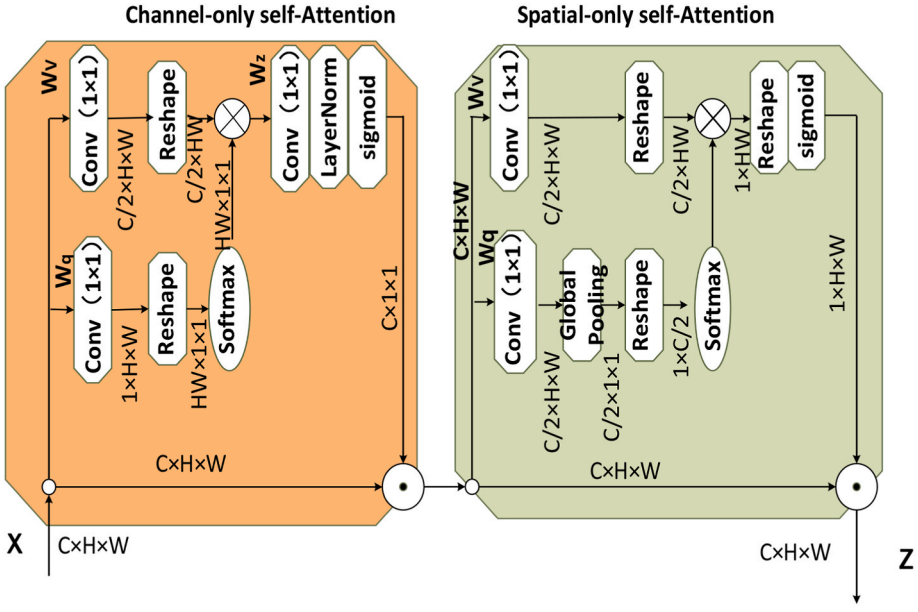


Fig. 5. PSA in series.

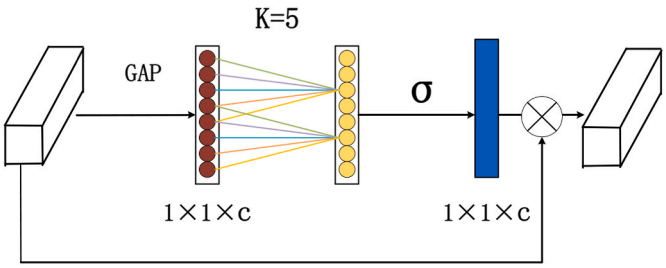


Fig. 6. ECA-Net diagram.

Table 1
Comparison results of ASPP improvement experiments.

Algorithm	Backbone	MIoU	mAP
Deeplabv3 plus	MobileNetV2	66.16%	78.75%
Deeplabv3 plus-SP		67.6%	78.6%

Table 2
Comparison of different attention mechanisms.

Backbone	Attention	MIoU	mAP
MobileNetV2	ECA-Net	66.95%	79.64%
MobileNetV2	PSA_p	67.3%	80.34%
MobileNetV2	PSA_s	67.74%	81.3%

Table 3
Comparison of network segmentation accuracy by integrating different modules.

Group	SP	PSA_p	PSA_S	ECA-Net	MIoU	MAP
①	×	×	×	×	66.16%	78.75%
②	✓	✓	×	×	68.67%	80.34%
③	✓	×	✓	×	69.05%	79.65%
④	✓	✓	×	✓	68.74%	79.01%
⑤	✓	×	✓	✓	69.77%	79.29%

feature maps obtained in the backbone network while using strip pooling instead of global pooling to retain more detailed information. Introduce the attention mechanism and add a polarization self-attention mechanism to weigh the feature maps obtained by the ASPP module. ECA-Net was added to fuse shallow features of MobileNetV2 and improve image segmentation performance. The improved model is shown in Fig. 2.

3.1. Strip pooling

The pooling window of global average pooling is square, which has certain limitations, and it is difficult to obtain the correlation of graph scales in different directions. Strip pooling has more advantages than global average pooling. The pooling window of strip pooling is rectangular, and the design of strip pooling can obtain global information from horizontal and vertical dimensions, expanding the scope of obtaining feature information (Hou et al., 2020).

Different from the global average pooling calculation method, strip pooling is performed simultaneously according to the horizontal and vertical spatial dimensions. In addition, when two spatial dimensions are pooled, the eigenvalues of a column or row are weighted averages. The model structure is shown in Fig. 3 below.

For the input image, the calculation formula of the row vector output is as follows:

$$y_i^h = \frac{1}{W} \sum_{0 \le j < w} X_{ij} \tag{1}$$

The calculation formula of the column vector output is as follows:

$$y_i^v = \frac{1}{H} \sum_{0 \le i < H} X_{ij} \tag{2}$$

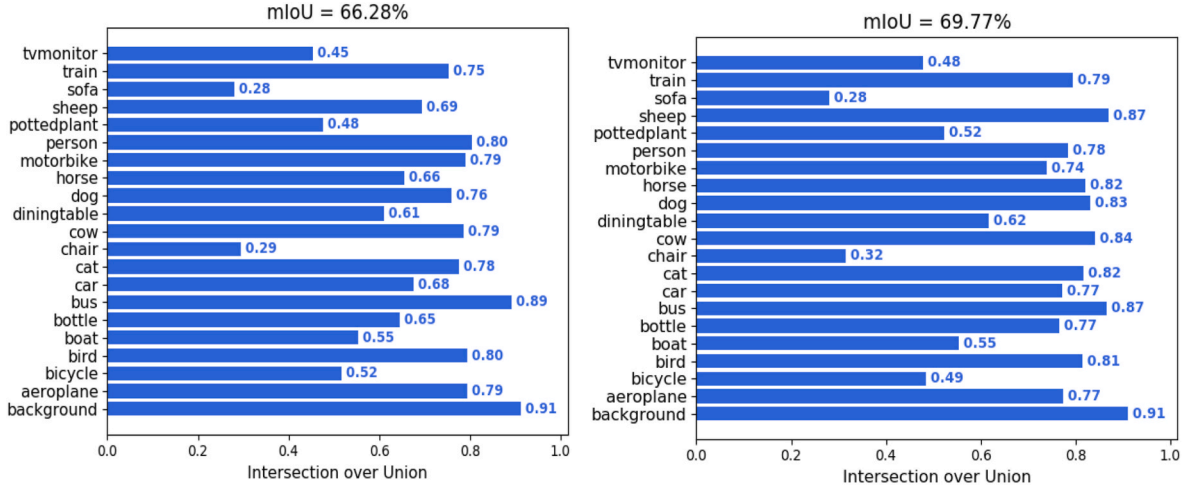
For an input $X \in R^{C \times H \times W}$, where C refers to the number of channels, H and W represent the height and width, respectively. X enters the horizontal and vertical paths for pooling, and the outputs in the vertical and horizontal directions are $y^h \in R^{C \times H}$ and $y^v \in R^{C \times W}$, respectively. After combining the two, the output is calculated as follows:

$$y_{c,i,j} = y_{c,j}^h + y_{c,j}^v \tag{3}$$

The convolution and sigmoid function will obtain the characteristic image, which will be fused with the original image to obtain the output z. The output z calculation formula is:

$$z = Scale(X, \sigma(f(y))) \tag{4}$$

In the above formula, scale () represents multiplication, σ represents the sigmoid function, and f represents 1×1 convolution.



(a) Before modification (b) After modification

Fig. 7. Comparison chart of category segmentation accuracy.

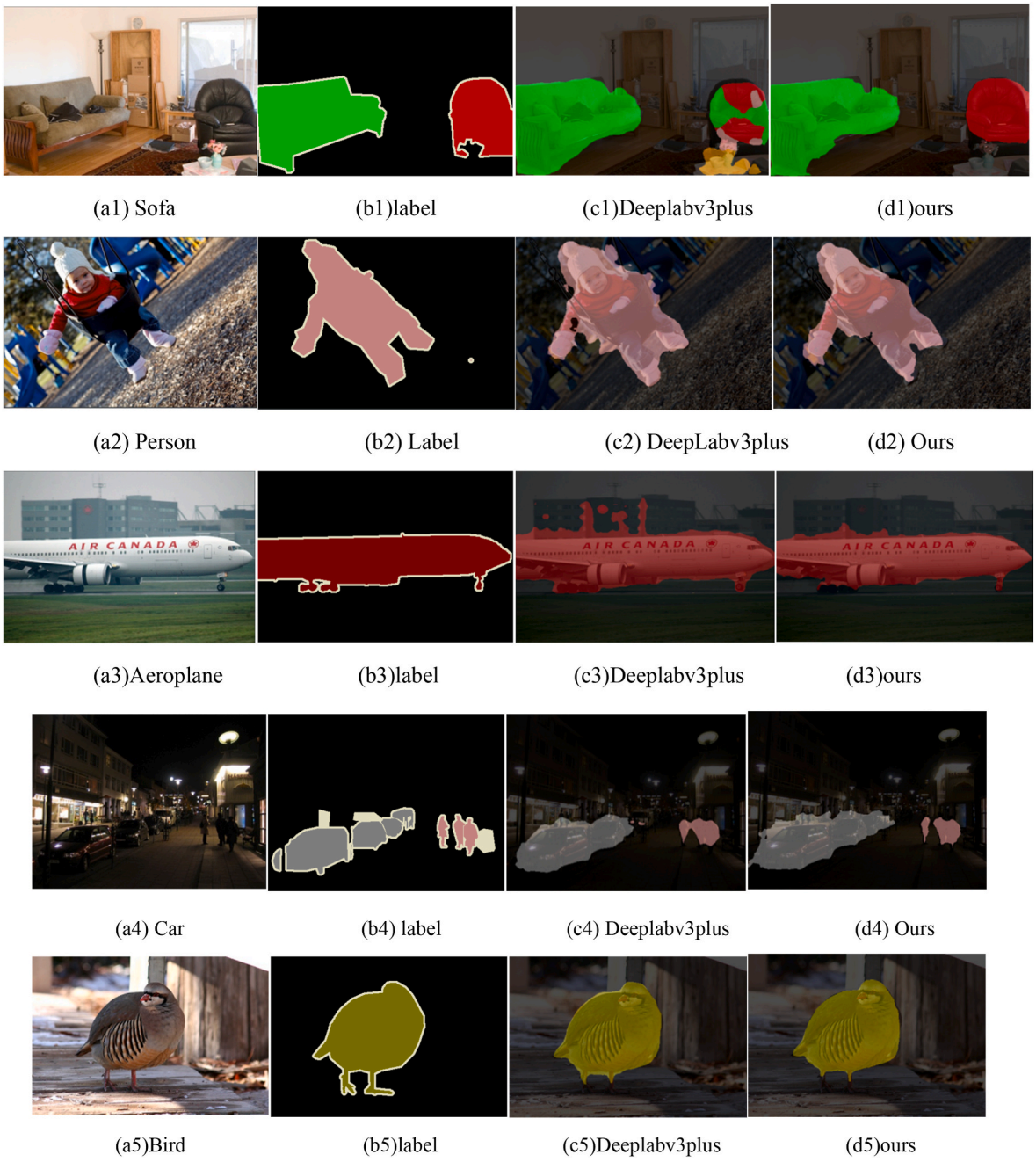


Fig. 8. Comparison of PASCAL VOC 2012 dataset segmentation results.

3.2. Polarized self-attention mechanism

We are all familiar with the concept of attention (Zeng et al., 2020). People cannot pay attention to the whole picture when they watch a picture. It must be that the eyes tend to be more interested in the part of the painting, and people will ignore the part that they are not interested in. Based on such characteristics, the attention mechanism in the neural network takes advantage of this, that is, to screen out effective information from complex information (Chen et al., 2017a). For image processing, the target will be locked in one part of the image while ignoring other areas, which can improve the efficiency of image processing and save unnecessary trouble. With the rapid development of attention mechanisms, an increasing number of neural network models have added attention mechanisms (Zhang et al., 2020; Honarbakhsh et al., 2023) to improve the efficiency of the model, which has shown a good effect. This paper mainly adds polarization self-attention and channel attention mechanisms to the DeepLabv3 plus network. The two attention mechanisms are added at different locations in the network, and both show good performance.

The polarized self-attention mechanism (Hridoy et al., 2021; Liu et al., 2021) has two main forms, series and parallel. The serial form refers to the serial form of the channel self-attention mechanism and spatial self-attention mechanism. The parallel form refers to the parallel form of the channel self-attention mechanism and spatial self-attention mechanism. The two ways together constitute the polarized self-attention mechanism. After inserting the polarization self-attention mechanism into the ASPP module (Yang, 2020; Zhu et al., 2019), the model can increase the extraction of important information and improve the utilization of the model. PSA_p and PSA_s can maintain high resolution in the channel and spatial dimensions, which is why they are

increasingly widely used in deep learning networks. The model diagram is shown in Figs. 4 and 5 below.

The series and parallel forms of the polarization self-attention mechanism are formally divided into two branches: channel branches and space branches.

The channel weight calculation formula is as follows:

$$A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}\left(\left(\sigma_1(W_v(X)) \times F_{SM}\left(\sigma_2\left(W_q(X)\right)\right)\right)\right)\right] \tag{5}$$

where and $\sigma_1\sigma_2$ represent the 1×1 convolution. F_{SM} represents the softmax function part. $W_{Z|\theta 1}$ Representing 1×1 convolution and LN elevates the dimension of C/2 on the channel to C. F_{SG} represents the sigmoid function.

The spatial weight calculation formula is as follows:

$$A^{sp}(X) = F_{SG}\left[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X))))) \times (X))\right] \tag{6}$$

where $\sigma_1\sigma_2$ and σ_3 represent the 1×1 convolution. F_{SM} represents the softmax function. F_{GP} represents global pooling. F_{SG} Represents the sigmoid function.

The above formula shows the calculation formula for two branch weights. The polarization self-attention mechanism is fused based on the branching weight. Parallel and series are just two simple calculations for shunt weights, similar to addition and multiplication.

3.3. ECA attention mechanism

The advantage of ECA-Net (Liu, 2020) is that it utilizes global pooling to transform spatial matrices into one-dimensional vectors .(see Fig. 6) Then, the size of the one-dimensional convolutional kernel can be obtained based on the number of network channels. Then, an adaptive size convolution kernel is used for the convolution operation, and the feature map of the input image is obtained through a weighted form. Finally, the input image is multiplied by the feature map obtained after convolution calculation to extract the information of interest. Due to the pretraining method of the backbone network adopted by the network, inserting ECA-Net into MobileNetV2 damages the network structure of the backbone network. Therefore, inserting ECA-Net into the shallow features of MobileNetV2 can improve the segmentation effect without damaging the network.

4. Experiments

4.1. Datasets

The PASCAL VOC2012 dataset is widely used and can be effectively utilized in the field of image processing. A dataset that can be used for image semantic segmentation. There are four main types in this dataset: indoor furniture, people, vehicles, and common animals. There are 21 categories in four categories, and 3200 images are randomly selected and divided into 9:1:1. A total of 2616 images are used as the training set, 292 images are used as the validation set, and 292 images are used as the testing set.

4.2. Experimental equipment and evaluation indicators

The operating system is Ubuntu 20.04, using the Python 1.2.0 deep learning open source framework and CUDA version 10.0. The programming language is Python 3.6, and the hardware configuration is as follows: The CPU is i7-9600, and the GPU is NVIDIA 3060-Ti. The average intersection to union ratio (MloU) and average pixel accuracy (mAP) are used as performance evaluation coefficients for image semantic segmentation. Where k represents k categories, P_{ij} indicates that the true value is i and the predicted value is j; P_{ji} indicates that the true value is j, the predicted value is i, and P_{ii} indicates that the true and predicted values are i. The calculation formulas for MloU and mPA are:

$$MIoU = \frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij} + \sum_{j=0}^{k}P_{ij} - P_{ii}} \tag{7}$$

$$\mathrm{m}PA = \frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K}p_{ij}} \tag{8}$$

4.3. Experimental comparison

The algorithm proposed in this paper is based on the original DeepLabv3 plus model (Sun et al., 2019; Badrinarayanan et al., 2017). The ASPP module is redesigned, and an attention mechanism is introduced to make the shallow and deep features of the model pay more attention to important semantic information(He et al., 2016; Chen et al., 2017b, 2018; Sehar and Naseem, 2022)-(He et al., 2016; Chen et al., 2017b, 2018; Sehar and Naseem, 2022). The fitting effect can be achieved by training the algorithm for 100 epochs using the Adam network model optimizer. The training was divided into two phases: the freezing phase and the unfreezing phase. A learning rate of 0.005 is used in the freezing phase, and the batch size is set to 8. A learning rate of 0.0005 is used in the unfreezing phase, and the batch size is set to 4. To prevent overfitting, the weight decay rate is set to 0.005. Epoch refers to the process of all the data entering the network to complete the forward computation and backpropagation once, and the number of epochs is set to 100, with 50 rounds in the freezing phase and 50 rounds in the unfreezing phase. Phase of 50 rounds and the unfreezing phase of 50 rounds. Before and after improvement. This article adopts the MloU and MAP evaluation index system and conducts ASPP module optimization, attention mechanism addition, and mutual fusion experiments on PASCAL VOC2012 to verify the performance of the model.

4.3.1. ASPP improvement experiment

The stripe pooling module (SP) is introduced in the ASPP module, where Deeplabv3 plus-sp represents using stripe pooling instead of global pooling in the ASPP module. To demonstrate the applicability of stripe pooling, MloU improved the DeepLabv3 plus network by 1.09% before and after improvement. As shown in Table 1 below.

4.3.2. Introduction of different attention experiments

Based on the MobileneV2 backbone network and ASPP module, different attention mechanisms are introduced. The polarization self-attention mechanism in series and parallel forms was introduced after the ASPP module. ECA-Net is introduced after the shallow layer of MobileneV2. MloU increased by 0.79% after joining PSA and ECA-Net. PSA_s has a better performance than PSA_p. In particular, MloU increased by 1.68% after adding PSA_s. As shown in Table 2 below.

4.3.3. Comparative experiments of different models

To demonstrate the effectiveness of the stripe pooling module, polarization self-attention mechanism module, and ECA Net module and to verify the accuracy of the improved algorithm, five control experiments were established. Among them, ① refers to the DeepLabv3 plus network. ② It refers to changing the global average pooling to stripe pooling in the ASPP module of DeepLabv3 plus and adding a polarization self-attention mechanism in parallel after the ASPP module. It refers to changing global pooling to stripe pooling in the ASPP module. DeepLabv3 plus, and adding a polarization self-attention mechanism in a concatenated form after the ASPP module. ④ It refers to changing global pooling to stripe pooling in the ASPP module of DeepLabv3 plus, adding a parallel form of polarization self-attention mechanism after the ASPP module, and adding the ECA-Net module after the shallow features of MobileneV2. ⑤ It refers to changing global pooling to stripe pooling in the ASPP module of DeepLabv3 plus, adding a concatenated form of the polarization self-attention mechanism after the ASPP module, and

adding the ECA-Net module after the shallow features of MobileneV2.

Table 3 compares ① and ② and ① and ③ of table. By using stripe pooling instead of global average pooling and introducing a polarization self-attention mechanism, Mlou improved by 2.51% and 2.89%, respectively. Compare ① and ④, ① and ⑤ of the table. In the ASPP module, stripe pooling replaces global average pooling, and the polarization self-attention mechanism and ECA-Net are introduced, resulting in increases of 2.58% and 3.61% in MloU, respectively. By analyzing the above table, it has been verified that all modules have played a role, and all the improvements mentioned above can greatly improve the accuracy of the algorithm.

4.4. Comparison of segmentation results for different categories

The most important evaluation indicator for accuracy in semantic segmentation is the average intersection-to-union ratio, which can be seen from the graph among the 21 categories. The modified model only has 6 categories that are lower than the original algorithm, and the accuracy of the 6 lower categories is not significantly different from the original algorithm. The remaining 15 categories are all higher than those of the original algorithm. Especially for categories such as houses, dogs, cats, trains, sheep, etc., showing better advantages. After adding the attention mechanism, the accuracy of key categories is improved, which can to some extent improve the accuracy of the original algorithm. The category segmentation results are shown in Fig. 7.

To see the effects before and after the improvement more clearly, the segmentation prediction maps of the DeepLabv3 plus network and the improved DeepLabv3 plus network were compared. Where (a) represents the original image, (b) represents the image label, (c) represents the DeepLabv3 plus segmentation image, and (d) represents the improved DeepLabv3 plus segmentation image. From the results, it can be seen that the model segmentation that integrates stripe pooling and introduces the attention mechanism is relatively smoother and more complete. The original DeepLabv3 plus network has problems with misclassification and discontinuous segmentation. The optimized network has improved the semantic segmentation effect, better resolution, refined the segmentation boundary of the target and achieved better accuracy. The selected segmentation prediction diagram is shown in Fig. 8.

5. Summary

This article proposes a DeepLabv3 plus network based on the attention mechanism. Changing global pooling to stripe pooling in the ASPP module captures global contextual information, while the addition of the polarization self-attention mechanism enhances the utilization of image spatial features. Finally, by adding ECA-Net after the low-level features of MobileNetV2, the acquisition of shallow features improves. The experimental results show that embedding the attention module into DeepLabv3 plus as a network can improve the accuracy of key categories and effectively improve the segmentation accuracy of objects in images by the network. The objective indicator MIoU improved by approximately 2%. Our work improves the performance of image semantic segmentation, which provides new ideas for autonomous driving, medical imaging, and other fields and provides direction for the field of computer vision.

Although the improved algorithm has made good improvements, there are still shortcomings. Since the introduction of the attention mechanism increases the model complexity to some extent, further research is needed in terms of model complexity and parameter quantity. In the future, we will consider using model compression methods to optimize the network so that the model can balance high accuracy and light weight.

CRediT authorship contribution statement

Yanyan Liu: Conceptualization, Methodology, Experiments. **Xiaotian Bai:** Experimental results analysis, Writing – review & editing. **Jiafei Wang:** Conceptualization, Methodology, Experiments. **Guoning Li:** Supervision. **Jin Li:** Supervision, Writing – review & editing. **Zengming Lv:** Experimental results analysis.

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.

Data availability

No data was used for the research described in the article.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Biao, W., Yali, G., Qingchuan, Z., 2018. Research on Image Semantic Segmentation Algorithm Based on Fully Convolutional HED-CRF[C]//2018 Chinese Automation Congress (CAC). IEEE, pp. 3055–3058. https://doi.org/10.1109/CAC.2018.8623459.

Boykov, Yuri, Veksler, Olga, Zabih, Ramin, 2001. Fast approximate energy minimization via graph cuts. In: Proceedings of the Seventh IEEE International Conference on Computer Vision 1, pp. 377–384, 1.

Chen, L.C., Papandreou, G., Kokkinos, I., et al., 2017a. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, L.C., Papandreou, G., Schroff, F., et al., 2017b. Rethinking Atrous Convolution for Semantic Image segmentation[J]. https://doi.org/10.48550/arXiv:1706.05587 arXiv preprint arXiv:1706.05587.

Chen, L.C., Zhu, Y., Papandreou, G., et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Dhanachandra, Nameirakpam, Manglem, Khumanthem, Yambem Jina Chanu, 2015. Image segmentation using K -means clustering algorithm and subtractive clustering algorithm. Procedia Comput. Sci. 54, 764–771.

Fu, J., Yi, X., Wang, G., et al., 2022. Research on ground object classification method of high resolution remote-sensing images based on improved DeeplabV3+[J]. Sensors 22 (19), 7477. https://doi.org/10.3390/S22197477.

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition[C]. Proc. IEEE Conf. on Comput. Vision and Pattern Recogn. 770–778. https://doi.org/10.3390/APP12188972.

Honarbakhsh, V., Siahkoohi, H.R., Rezghi, M., et al., 2023. SeisDeepNET: an extension of Deeplabv3+ for full waveform inversion problem[J]. Expert Syst. Appl. 213, 118848 https://doi.org/10.1016/J.ESWA.2022.118848.

Hou, Qibin, Zhang, Li, Cheng, Ming-Ming, Feng, Jiashi, 2020. Strip pooling: rethinking spatial pooling for scene parsing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 4003–4012.

Hridoy, R.H., Habib, T., Jabiullah, I., et al., 2021. Early recognition of betel leaf disease using deep learning with depthwise separable convolutions[C]. In: 2021 IEEE Region 10 Symposium (TENSYMP). IEEE, pp. 1–7. https://doi.org/10.1109/TENSYMP52854.2021.9551009.

Iftikhar, S., Asim, M., Zhang, Z., et al., 2022. Advance generalization technique through 3D CNN to overcome the false positives pedestrian in autonomous vehicles. Telecommun. Syst. 80, 545–557. https://doi.org/10.1007/s11235-022-00930-1.

Iftikhar, Sundas, Asim, Muhammad, Zhang, Zuping, Muthanna, Ammar, Chen, Junhong, El-Affendi, Mohammed, Ahmed, Sedik, Ahmed, A., Abd El-Latif, 2023. Target detection and recognition for traffic congestion in smart cities using deep learning-enabled UAVs: a review and analysis. Appl. Sci. 13 (6), 3995. https://doi.org/10.3390/app13063995.

Kass, Michael, Witkin, Andrew P., Terzopoulos, Demetri, 2004. Snakes: active contour models. Int. J. Comput. Vis. 1, 321–331.

Liu, M z, 2020. Research on Image Semantic Segmentation Algorithm Based on Self-Attention Mechanism [D] Dalian. Dalian Univ. Technol. 20–35. https://doi.org/10.26991/d.cnki.gdllu.2020.001777.

Liu, A., Yang, Y., Sun, Q., Xu, Q., 2018. A deep fully convolution neural network for semantic segmentation based on adaptive feature fusion. In: 2018 5th International Conference on Information Science and Control Engineering (ICISCE), pp. 16–20. https://doi.org/10.1109/ICISCE.2018.00013. Zhengzhou, China.

Liu, H., Liu, F., Fan, X., et al., 2021. Polarized self-attention: toward high-quality pixelwise regression[J]. arXivpreprintarXiv:2107.00782. https://doi.org/10.48550/arXiv.2107.00782.

Minaee, Shervin, Wang, Yao, 2017. An ADMM approach to masked signal decomposition using subspace representation. IEEE Trans. Image Process. 28, 3192–3204.
Minaee, S., Boykov, Y., Porikli, F., et al., 2021. Image segmentation using deep learning: a survey[J]. IEEE Trans. Pattern Anal. Mach. Intell. 44 (7), 3523–3542.
Najman, Laurent, Schmitt, Michel, 1994. Watershed of a continuous function. Signal Process. 38, 99–112.
Nock, R., Nielsen, F., 2004. Statistical region merging. IEEE Trans. Pattern Anal. Mach. Intell. 26 (11), 1452–1458. https://doi.org/10.1109/TPAMI.2004.110.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man, and Cybern. 9 (1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076.
Plath, Nils, Toussaint, Marc, Nakajima, Shinichi, 2009. Multiclass image segmentation using conditional random fields and global classification. International Conference on Machine Learning.
Sehar, U., Naseem, M.L., 2022. How deep learning is empowering semantic segmentation: traditional and deep learning techniques for semantic segmentation: a comparison[J]. Multimed. Tool. Appl. 81 (21), 30519–30544. https://doi.org/10.1007/S11042-022-12821-3.
Starck, J.-L., Elad, M., Donoho, D.L., 2005. Image decomposition via the combination of sparse representations and a variational approach. IEEE Trans. Image Process. 14 (10), 1570–1582. https://doi.org/10.1109/TIP.2005.852206.
Sun, W., Wang, R., 2018. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM[J]. Geosci. Rem. Sens. Lett. IEEE 15 (3), 474–478. https://doi.org/10.1109/LGRS.2018.2795531.
Sun, Y., Jiang, Q., Hu, J., et al., 2019. Attention mechanism based pedestrian trajectory prediction generation model[J]. J. Comput. Appl. 39 (3), 668. https://doi.org/10.13203/j.whugis20200159.
Yang, X., 2020. An overview of the attention mechanisms in computer vision[C]// Journal of Physics: conference Series. IOP Publish. 1693 (1), 012173 https://doi.org/10.1088/1742-6596/1693/1/012173.
Yang, Z., Peng, X., Yin, Z., 2020. Deeplab_v3_plus-net for image semantic segmentation with channel compression[C]//2020 IEEE 20th international conference on communication technology (ICCT). IEEE 1320–1324. https://doi.org/10.1109/ICCT50939.2020.9295748.
Zeng, H., Peng, S., Li, D., 2020. Deeplabv3+ semantic segmentation model based on feature cross attention mechanism[C]. In: Journal of Physics: Conference Series. IOPPublishing, 012106. https://doi.org/10.1088/1742-6596/1678/1/012106, 1678(1).
Zhang, Z., Huang, J., Jiang, T., et al., 2020. Semantic segmentation of very high-resolution remote sensing image based on multiple band combinations and patchwise scene analysis[J]. J. Appl. Remote Sens. 14 (1) https://doi.org/10.1117/1.JRS.14.016502, 016502-016502.
Zhao, Hengshuang, Shi, Jianping, Qi, Xiaojuan, Wang, Xiaogang, Jia, Jiaya, 2017. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR) 2881–2890.
Zhu, Z.L., Rao, Y., Wu, Y., et al., 2019. Research progress of attention mechanism in deep learning[J]. J. Chin. Inf. Process. 33 (6), 1–11. https://doi.org/10.13374/j.issn2095-9389.2021.01.30.005.