Enet Semantic Segmentation Combined with Attention Mechanism

Wei Bai (✉ nxbwei@163.com)

Ningxia Normal University

Research Article

Keywords: Semantic segmentation, Convolutional neural network, ENet, Attention mechanism, Computer vision

Posted Date: April 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-425438/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Enet Semantic Segmentation Combined with Attention Mechanism

Wei Bai*

School of mathematics and computer science, Ningxia Normal University, Ningxia, Guyuan, 756000, China

Corresponding author: Wei Bai; nxbwei@163.com

Abstract Image semantic segmentation is one of the core tasks of computer vision. It is widely used in fields such as unmanned driving, medical image processing, geographic information systems and intelligent robots. Aiming at the problem that the existing semantic segmentation algorithm ignores the different channel and location features of the feature map and the simple method when the feature map is fused, this paper designs a semantic segmentation algorithm that combines the attention mechanism. Firstly, dilated convolution is used, and a smaller downsampling factor is used to maintain the resolution of the image and obtain the detailed information of the image. Secondly, the attention mechanism module is introduced to assign weights to different parts of the feature map, which reduces the accuracy loss. The design feature fusion module assigns weights to the feature maps of different receptive fields obtained by the two paths, and merges them together to obtain the final segmentation result. Finally, through experiments, it was verified on the Camvid, Cityscapes and PASCAL VOC2012 datasets. Mean intersection over union (MIoU) and mean pixel accuracy (MPA) are used as metrics. The method in this paper can make up for the loss of accuracy caused by downsampling while ensuring the receptive field and improving the resolution, which can better guide the model learning. And the proposed feature fusion module can better integrate the features of different receptive fields. Therefore, the proposed method can significantly improve the segmentation performance compared to the traditional method.

Keywords Semantic segmentation; Convolutional neural network; ENet; Attention mechanism; Computer vision

1 Introduction

Computer vision mainly includes tasks such as classification, detection, and segmentation. The classification task is mainly for the whole picture, the detection task is for the part of the image, and the semantic segmentation aims to give each pixel on the input picture a corresponding semantic label^[1]. Semantic segmentation is the basis of image processing, and its segmentation results are often used in subsequent image processing^[2]. At the same time, compared with image classification or target detection, semantic segmentation can realize more fine-grained reasoning on images^[3-5]. However, traditional image segmentation technology faces the problem of difficult feature extraction and training, and requires a lot of artificial feature design^[6].

In recent years, artificial intelligence (AI) has developed rapidly, driven by new theoretical technologies such as cloud computing, big data, brain science, the Internet of Things, blockchain, and the strong demand for social and economic development^[7-8].

The Chinese government has successively issued important policy documents such as "Made in China 2025" and "New Generation Artificial Intelligence Development Plan" that constitute the core of China's

artificial intelligence strategy. At the same time, the U.S. government has successively released the "National Artificial Intelligence Research and Development Strategic Plan", "Artificial Intelligence, Automation and Economy" and "American AI Plan" and other artificial intelligence development strategies^[9-10]. Artificial intelligence has become a new strategic highland for many powerful countries to compete^[11].

Artificial intelligence has become a decisive technological force leading the future, and it will strongly affect the country's comprehensive competitiveness, economy, security, and social ethics^[12]. These phenomena and measures highlight the opportunities and challenges brought about by the development of artificial intelligence technology, and at the same time trigger the world's investment in artificial intelligence.

As a branch with great development potential and influence in the field of artificial intelligence, deep learning has achieved remarkable results in cutting-edge tasks such as computer vision, natural language processing, machine translation, and autonomous driving^[13-15]. In many application scenarios, computer vision technology is a very important part. Such as intelligent driving, face recognition, beauty selfie, product selection and traffic flow control and other applications^[16-17]. Deep

learning has also greatly promoted the development of image semantic segmentation technology in computer vision^[18].

On the other hand, a large number of application scenarios require precise and efficient semantic segmentation technology. The society's demand for it is becoming stronger and more widely used. Such as the following application areas:

(1) Unmanned driving field. Semantic segmentation is widely used in the field of unmanned driving. Sensor devices such as on-board cameras and optical radars perceive the surrounding environment and then input images into the neural network. The computer can automatically segment and classify the images to accurately locate roads and avoid obstacles such as pedestrians and vehicles.

(2) Geographic Information System. Through training the neural network, the machine can input satellite remote sensing images to effectively identify bridges, roads, houses, farmland, mountains, rivers and lakes. Different scenes can be automatically marked.

(3) Medical image analysis. In recent years, more and more medical researches apply deep learning to the field of medical diagnosis. In the field of intelligent medicine, semantic segmentation can be applied to tasks such as tumor image segmentation and cardiovascular and cerebrovascular diagnosis. And it is stronger than radiologists in analysis ability, which can greatly reduce the time required for diagnosis.

(4) Smart robot. Intelligent robots need functions such as movement, obstacle avoidance and recognition. In particular, when performing tasks such as medical care, high-precision detection and recognition, industrial production, high-speed or complex trajectory movement, the understanding of the scene often needs to reach a pixel-level fineness. Semantic segmentation technology can help intelligent robots deeply understand the surrounding environment.

2 Overall structure and basic concepts

Semantic segmentation is a basic task in the field of computer vision. It is required to mark each pixel in the image with a category according to context information and spatial structure information^[19]. Image semantic segmentation has important applications in various fields. In the field of health and medical care, semantic segmentation technology can accurately extract lesions from patients' X-ray or B-ultrasound images to improve diagnosis efficiency. In the field of autonomous driving,

specific information about obstacles can be obtained through semantic segmentation technology to help cars avoid obstacles. In the security field, semantic segmentation technology can accurately segment dangerous goods from security inspection images, realize automatic detection of dangerous goods, and improve security inspection efficiency. With the development of deep learning technology and the improvement of graphics processing unit (GPU) computing power, training large-scale, deep-level neural networks has become the mainstream mode in the field of semantic segmentation.

AlexNet is a proposed deep convolutional neural network model^[20]. In the 2012 ILSVRC (ImageNet large scale visual recognition challenge) competition, AlexNet achieved the first place with a top-5 accuracy rate of 84.6%, while the accuracy rate of the second place was only 73.8%. The emergence of AlexNet has aroused the attention of many scholars and unveiled the curtain of deep learning. Reference [21] proposes fully convolutional networks (FCN). By changing the fully connected layer in the last part of the original classification network VGG16 (visual geometry group 16-layer net)^[22] to a convolutional layer. The purpose of combining the semantic information of the deep and coarse network layer with the surface information of the shallow and fine network layer is realized, so as to achieve precise segmentation. Reference [23] proposed that the semantic segmentation network SegNet uses de-pooling in the decoder to upsample the generated feature maps, so that some subtle edge information is preserved intact. At the same time, in the feature extraction part, SegNet uses the same full convolutional network as FCN, which greatly reduces the number of parameters. The U-Net proposed in [24] adopts a U-shaped structure, splicing the feature map of the encoder stage to the feature map of the corresponding size in the decoder stage, and fusion of spatial semantic information and pixel information. The decoder learns the information lost in the downsampling and improves the segmentation effect. The DeepLabV1 proposed in [25] uses hole convolution to reduce the size of the feature map while keeping the resolution unchanged. And at the end of the network, the conditional random field (CRF) model is used to restore the edge information and achieve accurate positioning. Reference [26] proposed ENet for

tasks with low latency and high speed, which greatly improved the speed of the model. It is 18 times faster than the original network model, 79 times less than FLOPs (floating-point operations per second), and the number of parameters is 79 times less. The accuracy is close to the original model. Reference [27] proposed PSPNet (pyramid scene parsing network), which uses expanded convolution to modify the basic residual network Res-Net architecture. After the initial pooling layer, the same resolution is used in the entire encoder network. In addition, PSPNet added an additional loss function to the overall learning and achieved the best results. Reference [28] proposed a new type of decoder structure. First, the deep features and shallow features are cascaded, and then through the self-attention mechanism, the final decoded features are obtained. Reduce the accuracy loss caused by decoding upsampling. Reference [29] proposed a new type of semantic segmentation algorithm. The dense layer structure is added to the network, and grouped convolution is used to speed up the calculation, and an attention mechanism is added to improve the segmentation effect.

Although many achievements have been made in the field of semantic segmentation, there are still too simple fusion methods that directly cascade or add shallow features and deep features. Ignoring the specificity of different features, resulting in loss of segmentation accuracy. In addition, information is lost in the process from down-sampling to up-sampling, resulting in a decrease in segmentation accuracy.

In response to the above problems, this paper designs a semantic segmentation algorithm combined with attention mechanism. The spatial information path is used to use dilated convolution, and a smaller down-sampling multiple is used to maintain the resolution of the image and obtain the detailed information of the image. Introducing the attention mechanism module to assign weights to different parts of the feature map, which reduces the accuracy loss. The design feature fusion module assigns weights to the feature maps of different perception fields obtained by the two paths, and merges them together to obtain the final segmentation result.

3 Method

3.1 ENet network structure

ENet network is a lightweight image semantic

segmentation network proposed by Paszke [30], which can achieve target pixel-level semantic segmentation. It has the characteristics of fewer parameters and fast calculation speed. At the same time, ENet network also has a certain degree of plasticity. Based on this, this paper optimizes and improves the pruning and convolution of the ENet network. A semantic segmentation algorithm based on improved ENet network is proposed.

The ENet network uses the current general encoder-decoder network structure. This is a lightweight network structure, which has huge advantages in model size and parameter amount. The ENet network changes the previous encoder-decoder symmetry structure. The convolution operation is reduced in the decoder, thereby greatly speeding up the processing speed. The ENet network has an initialization operation for the input image, as shown in Fig. 1. Its main purpose is to generate feature maps, and merge the feature maps generated by pooling and convolution operations. The convolution operation has a total of 13 3×3 filters with a sliding step of 2, and a total of 13 feature maps are obtained. Max Pooling is a non-covered 2×2 sliding window to obtain 3 feature maps. Finally, a total of 16 feature maps were obtained after fusion.

Fig.1 Initialization operation

In addition, a Bottleneck convolution structure is also used in the ENet network, and this module runs through the ENet network. Its main application is in the encoder-decoder, the concrete structure is shown as in Fig. 2. Each Bottleneck contains 3 convolutional layers. In Fig. 2, from top to bottom, there are 1 1×1 projection mapping for reducing dimensionality, 1 main convolutional layer, and 1 1×1 ascending dimension. And perform normalization and PReLU activation operations between convolutional layers. Bottleneck is not static and will change accordingly according to specific operations. If it is a down-sampled Bottleneck, the 1×1 projection

mapping is replaced by a Max Pooling layer with a kernel size of 2×2 and a step size of 2, and it is filled with 0 to match the size of the feature map. Conv is a 3×3 conventional convolution, expansion convolution or full convolution, and sometimes 1×5 and 5×1 asymmetric convolution are used instead. Regularization uses Spatial Dropout to solve the problem of model overfitting.

Fig.2 Bottleneck structure

The overall architecture of the ENet network includes 5 parts between initialization and final full convolution. In the first part, there is 1 down-sampled Bottleneck, followed by 4 ordinary convolutional Bottleneck. In the second part, the first is the Maximum Pooled Bottleneck, followed by 8 ordinary Bottleneck. In the third part, repeat all operations in the second part except the maximum pooling Bottleneck operation. In the fourth part, there is 1 up-sampled Bottleneck, followed by 2 regular Bottleneck. Since the previous operation has extracted enough feature information, it is necessary to restore the resolution of the image to its original size and output it. Therefore, the fifth part is directly an up-sampled Bottleneck followed by an ordinary convolutional Bottleneck. Finally, full convolution outputs the final result of image semantic segmentation. The reason why the expansion convolution module is not used in the fourth and fifth parts is because the coding modules of the first three parts have already segmented the image completely, and there is no need to expand the field of view to extract feature information.

3.2 Attention mechanism module

In order to improve the segmentation speed of the network, this paper uses a lightweight network as the feature extraction network, but there is a certain loss in

accuracy. In order to reduce the loss of accuracy, an attention module is added to the network of this article. The attention feature helps to enhance the feature expression of the model, integrate different information, and improve the understanding of the model. This is similar to the attention mechanism of human vision. Human visual attention mechanism is divided into two types: bottom-up data-driven attention mechanism, top-down task-driven target attention mechanism. Both mechanisms can learn the part needed by the task from a large amount of data. This paper uses a bottom-up data-driven attention mechanism. The attention module starts from the relationship between the feature channels and considers the interdependence factors between the feature channels. Through the self-learning of the network, it effectively suppresses the features that have little influence on the current segmentation effect and enhances the weight of beneficial features. The attention module first performs global average pooling on the feature map of each channel to obtain a $1 \times 1 \times K$ vector, and then performs two FC layer conversions. In order to suppress the complexity of the model, dimensionality reduction and dimensionality increase were performed between the two FC layer conversions, and the Sigmoid and ReLU activation functions were used. The feature extraction network with the attention module is shown in Fig. 3.

Fig.3 Feature extraction network with attention module

In the attention module, global average pooling is first used to shrink the data output from the dense layer from $W \times H \times K$ to $1 \times 1 \times K$. The specific formula is as follows:

$$Z_c = F_{sq}(u_c) \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (1)$$

Among them, F_{sq} represents the global average pooling function.

Next, perform two FC operations on the network. C represents the dimensionality reduction coefficient, which can be adjusted according to the specific network. The best performance is obtained when $C=16$ in this experiment, and the formula is as follows.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (2)$$

Among them, σ and δ respectively represent Sigmoid and ReLU activation functions, $W_1, W_2 \in R^{\frac{K^2}{C}}$.

Then carry out the scale operation, and based on the number of channels unchanged, change the data to $W \times H \times K$.

$$X_c = F_{scale}(u_c, s_c) = u_c \cdot s_c \quad (3)$$

Before the network joins the attention module, it takes 42ms for a forward propagation for input pictures with batches of 256 and 224×224 . After adding the attention module, it takes 47ms. After reducing the complexity through dimensionality reduction-upgrading, there is still time to increase, but it can be ignored compared to the improvement of segmentation accuracy.

Finally, each layer of the network is connected by a skip connection. The jump connection in this paper is different from DenesNet in two points. One is that DenesNet only has connections in the blocks between downsampling, and in this article, there are connections in all layers. The second is in the way of pooling. This article adopts the atrous spatial pyramid pooling (ASPP). ASPP provides a model of multi-scale information. It adds void convolutions with different expansion rates on the basis of the spatial pyramid pooling (SPP) to capture a wide range of language territory. Global average pooling (GAP) is combined with image features to increase the global context. A total of 4 parallel operations, a 1×1 convolution and three 3×3 convolutions are included, and batch normalization is added.

3.3 Data augmentation

This article chooses to conduct comparative experiments on the CamVid dataset. The data is collected

and annotated by the University of Cambridge. The data set consists of a group of road and street view images, and the resolution size of the data set is 960×720 . The number of label categories in the data set is uniformly set to 11 categories. The data set consists of three parts, namely training set, validation set, and test set. For the CamVid data set, in order to further expand the number of images to prevent over-fitting, all data preprocessing in the comparative experiment stage in this paper chose the same method as SEGNET to enhance the data set. Including Gaussian smoothing, random addition of noise points, color dithering, image rotation and scaling and other operations. There is no data enhancement for the test set and validation set.

For the Cityscapes dataset and PASCALVOC2012 dataset, due to the scarcity of data and the limitation of computing resources, we performed the following operations on the dataset. First, the five remote sensing images contained in the data set are divided, the first four are used as training sets, and the last one is used as validation sets. Then segment the first four images (overlap allowed) into 256×256 image blocks. Finally, data enhancement is performed on the obtained image block, including operations such as adding salt and pepper noise, color dithering, image rotation and scaling.

4 Results and discussion

4.1 Introduction to the experimental environment and network parameter settings

All experiments in this article are built under the TensorFlow1.9 framework, using cuDNN7.5 kernel calculation, workstation configuration Intel CoreTMi7-6800K CPU@3.4GHz, GTX1080Ti graphics card, memory 128GB. The batch size during training is 256, the initial learning rate is set to 0.01, the learning rate attenuation is 10^{-6} , and Nesterov momentum is used. The initialization method is Variance Scaling, which can adjust the scale according to the weight size, and the activation function uses soft plus. The process of determining the model in this paper is as follows: First, a dense layer is added to the 18-layer feature extraction network. It is found that the network segmentation performance is improved compared to the original 18-layer network, and then the attention module is added to obtain a better segmentation effect, but the speed cannot meet the real-time requirements. In order to obtain

better segmentation performance and faster segmentation speed, a feature extraction network is added layer by layer. Observe the effect of the number of feature extraction network layers on accuracy and speed. Finally, it is determined that the 5-layer feature extraction network structure can maintain the segmentation performance and speed at a high level.

4.2 Evaluation index

Many standards are usually used in image segmentation to measure the accuracy of the algorithm.

Pixel accuracy (PA) is the simplest accuracy measure for semantic segmentation, and it represents the proportion of pixels that are correctly labeled to the total pixels.

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (4)$$

Where, p_{ii} represents the correct number of segmentation. p_{ij} represents the number of pixels that originally belonged to category i but were divided into categories j . p_{ji} represents the number of pixels that originally belonged to category j but were divided into category i . There are $k + 1$ categories (including k categories and an empty category or background category).

Mean Pixel Accuracy (MPA) is a simple improvement of PA. Calculate the proportion of pixels that are correctly segmented in each class, and then find the average of all classes.

$$MPA = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}} \quad (5)$$

Mean intersection over union (MIOU) is a metric for semantic segmentation technology, which calculates the ratio of intersection and union of two sets. Calculate the IOU within each pixel category, and then calculate the average.

$$MIOU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (6)$$

4.3 CamVid dataset

CamVid is the earliest semantic segmentation data set used in the field of autonomous driving. At first, five video sequences with a resolution of 960×720 pixels were shot on the dashboard of the car, and the shooting angle of view was basically the same as that of the driver. Using image annotation software, 700 images were

continuously annotated in the video sequence, including 32 categories such as buildings, trees, traffic lights, sky, roads, pedestrians, motorcycles, cars, and buses. The annotation of an image in the dataset is shown in Fig. 4.

Fig. 4 Schematic diagram of annotated images of CamVid data set

Under the hardware conditions described in section 4.1 of this article, the CamVid data set is verified, and the models obtained are MPA and MIOU on the CamVid test set, as shown in Fig. 5:

Fig.5 Comparison of experimental results (CamVid)

Comparing the results in Fig. 5, the method proposed in this paper surpasses the original model, FCN model, and DeepLabV3 model in all indicators. And the indicators are slightly higher than the data indicators of DeepLabV3.

In order to more intuitively reflect the improvement of pixel category consistency by adding the attention mechanism module structure, we visualized the predicted values of the original SEGNET model and the attention mechanism module structure, as shown in Fig. 6. Comparing Fig. 6(a) and Fig. 6(b), we can see that compared with the traditional model, after adding the attention mechanism module structure, the misdetection of pixel categories contained in the same target is greatly reduced.

(a)Original image (b)Traditional model (c) Proposed method (d)Manual marking

Fig.6 Segmentation results comparison

4.4 Cityscapes dataset

Cityscapes is a large-scale data set of 5,000 high-quality images collected from street scenes in 50 different cities. The training set, validation set and test set of the Cityscapes data set consist of 2975, 500 and 1525 images respectively. Including 19 categories such as ground, building, road signs, nature, sky, people and vehicles. The data set also provides 20,000 coarsely segmented images for training the performance of the classification network based on weakly supervised learning. The annotation map of the images in this dataset is shown in Fig. 7.

Fig.7 Schematic diagram of annotated images of Cityscapes data set

In order to verify the effectiveness of the improved algorithm, the segmentation results of DeepLabV3 and the proposed algorithm on Cityscapes dataset are compared. The comparison results of MIoU values on Cityscapes dataset before and after the improvement are shown in Table 1.

Tab.1 Comparison of MIoU values of the algorithm on the Cityscapes dataset

Classification	DeepLab V3	Improved algorithm
Road	0.913	0.922
Sidewalk	0.845	0.867
Building	0.900	0.912
Wall	0.501	0.518
Sky	0.918	0.932
Train	0.911	0.933
Person	0.942	0.940
Bicycle	0.728	0.786
Fence	0.923	0.923
Traffic light	0.965	0.966
Bus	0.854	0.866

It can be seen from Table 1 that the improved algorithm has improved the accuracy of most object

recognition. It shows that the improved network helps to improve the segmentation ability of the network model.

The prediction results of the improved algorithm on the Cityscapes data set are shown in Fig. 8. The algorithm in this paper effectively segmented the missing pedestrians in DeeplabV3+ and improved the segmentation ability of small-scale targets.

(a) Original image (b) DeeplabV3+ (c) Proposed method (d) Manual marking

Fig.8 Segmentation results of algorithm on Cityscapes dataset

4.5 PASCAL VOC2012 dataset

PASCAL VOC is a challenge competition for object classification, recognition and detection in the field of computer vision. It provides standard labeled data sets and evaluation systems for detection algorithms and network learning performance. The training set, validation set, and test set of the PASCAL VOC 2012 data set include 1464, 1449, and 1452 images respectively, with a total of 21 categories. Including humans, animals (birds, cats, cows, dogs, horses, sheep), vehicles (airplanes, bicycles, boats, buses, cars, motorcycles, trains), indoor objects (bottles, chairs, tables, potted plants) , Sofa, TV) and background. The annotation of an image of this dataset is shown in Fig. 9.

Fig. 9 Schematic diagram of annotated images of PASCAL VOC2012 data set

The visualization result of the algorithm on the PASCAL VOC2012 data set is shown in Fig. 10.

(a) Original image (b) DeeplabV3+ (c) Proposed method (d) Manual marking

Fig.10 Segmentation results of algorithm on PASCAL VAL 2012 dataset

It can be seen from the first row of Fig. 10 that the algorithm in this paper has identified the pixels lost in the DeeplabV3+ network segmentation process, and the prediction result is more accurate. In the result of the second row of segmentation, the DeeplabV3+ network model incorrectly recognizes the chair as another object. The algorithm proposed in this paper avoids this error.

5 Conclusion

This paper proposes a semantic segmentation algorithm combined with attention mechanism. Through dilated convolution, a certain resolution is ensured while downsampling obtains a larger receptive field. At the same time, an attention mechanism is added to the network, so that the network can learn weights adaptively and improve the segmentation accuracy. Finally, a feature fusion module is designed to better integrate the features of different receptive fields. The algorithm in this paper is tested on the Camvid, Cityscapes and PASCAL VOC2012 data sets, and the final results are compared with various algorithms. The algorithm in this paper has achieved better results. Next, we will further improve the attention mechanism module to explore the specificity and commonalities between different locations and channel characteristics. Extract more effective feature information to improve the algorithm's segmentation accuracy on small target objects.

Ethics approval and consent to participate

Our manuscript does not involve research manuscripts of human participants, human data or human tissues, so our manuscript does not require the statement of ethical approval and ethical consent.

Consent for publication

Our manuscript does not contain any individual person's data in any form, so we do not need the consent of others.

Data Availability

The data included in this paper are available without any restriction.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Funding Statement

This research is supported by School-level scientific

research project funding of Ningxia Normal University (No. NXSFDZDA2104).

Abbreviations

Not applicable.

Authors' contributions

The main idea of this paper is proposed by Wei Bai. The algorithm design and experimental environment construction are completed by Wei Bai. The experimental verification was completed by Wei Bai. The writing of the article is completed by Wei Bai. And the writing guidance, English polish is completed by Wei Bai.

Acknowledgement

We wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

References

- [1] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: A brief review[J]. Computational intelligence and neuroscience, 2018, 8(23): 13-26.
- [2] Khan S, Rahmani H, Shah S, et al. A guide to convolutional neural networks for computer vision[J]. Synthesis Lectures on Computer Vision, 2018, 8(1): 187-207.
- [3] O'Mahony N, Campbell S, Carvalho A, et al. Deep learning vs. traditional computer vision[C]//Science and Information Conference. Springer, Cham, 2019: 128-144.
- [4] Nanni L, Ghidoni S, Brahnam S. Handcrafted vs. non-handcrafted features for computer vision classification[J]. Pattern Recognition, 2017, 7(02): 158-172.
- [5] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.
- [6] Garcia A, Orts S, Oprea S, et al. A review on deep learning techniques applied to semantic segmentation[J]. arXiv preprint arXiv:1704.06857, 2017.

-
- [7] Lu H, Li Y, Chen M, et al. Brain intelligence: go beyond artificial intelligence[J]. *Mobile Networks and Applications*, 2018, 23(2): 368-375.
- [8] Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence[J]. *California management review*, 2019, 61(4): 5-14.
- [9] Liew C. The future of radiology augmented with artificial intelligence: a strategy for success[J]. *European journal of radiology*, 2018, 10(3): 152-156.
- [10] Cath C, Wachter S, Mittelstadt B, et al. Artificial intelligence and the 'good society': the US, EU, and UK approach[J]. *Science and engineering ethics*, 2018, 24(2): 505-528.
- [11] Collins G, Moons KGM. Reporting of artificial intelligence prediction models[J]. *The Lancet*, 2019, 393(11): 1577-1579.
- [12] Yu K, Beam A, Kohane S. Artificial intelligence in healthcare[J]. *Nature biomedical engineering*, 2018, 2(10): 719-731.
- [13] Kamilaris A, Prenafeta-Boldú X. Deep learning in agriculture: A survey[J]. *Computers and electronics in agriculture*, 2018, 14(7): 70-90.
- [14] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare[J]. *Nature medicine*, 2019, 25(1): 24-29.
- [15] Chen J, Ran X. Deep Learning With Edge Computing: A Review[J]. *Proceedings of the IEEE*, 2019, 107(8): 1655-1674.
- [16] Naik N, Kominers D, Raskar R, et al. Computer vision uncovers predictors of physical urban change[J]. *Proceedings of the National Academy of Sciences*, 2017, 114(29): 7571-7576.
- [17] Kim Y, Kim H. Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study[J]. *Information Processing & Management*, 2018, 54(6): 1101-1114.
- [18] Zhang J, Zhao X, Chen Z, et al. A review of deep learning-based semantic segmentation for point cloud[J]. *IEEE Access*, 2019, 7: 179118-179133.
- [19] Jiang F, Gu Q, Hao H. Summary of content-based image segmentation methods[J]. *Journal of Software*, 2017, 28 (1):160-183
- [20] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.
- [21] Sun W, Wang R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM[J]. *IEEE Geoscience and Remote Sensing Letters*, 2018, 15(3): 474-478.
- [22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*[J]. *arXiv preprint arXiv:1409.1556*, 2018.
- [23] Badrinarayanan V, Handa A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling[J]. *Computer Science*, 2015, 23(2):2006-2017.
- [24] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2015.
- [25] Zhang S, Ma Z, Zhang G, et al. Semantic image segmentation with deep convolutional neural networks and quick shift[J]. *Symmetry*, 2020, 12(3): 427-440.
- [26] Lo S, Hang HM, Chan SW, et al. Efficient dense modules of asymmetric convolution for real-time semantic segmentation[M]//*Proceedings of the ACM Multimedia Asia*. 2019: 1-6.
- [27] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 2881-2890.
- [28] Ma Z, Gao H, Lei T. Semantic segmentation algorithm based on enhanced feature fusion decoder [J]. *Computer Engineering*, 2020, 46(05): 254-258.
- [29] Cheng X, Zhao L, Hu Q. Fast semantic segmentation based on dense layer and attention mechanism [J]. *Computer Engineering*, 2020, 46(04): 247-252.

[30] Paszke A, Chaurasia A, Kim S, et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation[J]. arXiv 1606.02147, 2016.

Wei Bai received his M.S. degree in Education from Ningxia University in Ningxia, China. He is currently worked in Ningxia Normal University, associate professor. His research interest in the area of Computer Science and Technology. He has published several research papers in scholarly journals in the above research areas.

Figures

Figure 1

Initialization operation

Figure 2

Bottleneck structure

Figure 3

Feature extraction network with attention module

Figure 4

Schematic diagram of annotated images of CamVid data set

Figure 5

Comparison of experimental results (CamVid)

Figure 6

Segmentation results comparison

Figure 7

Schematic diagram of annotated images of Cityscapes data set

Figure 8

Segmentation results of algorithm on Cityscapes dataset

Figure 9

Schematic diagram of annotated images of PASCAL VOC2012 data set

Figure 10

Segmentation results of algorithm on PASCAL VAL 2012 dataset