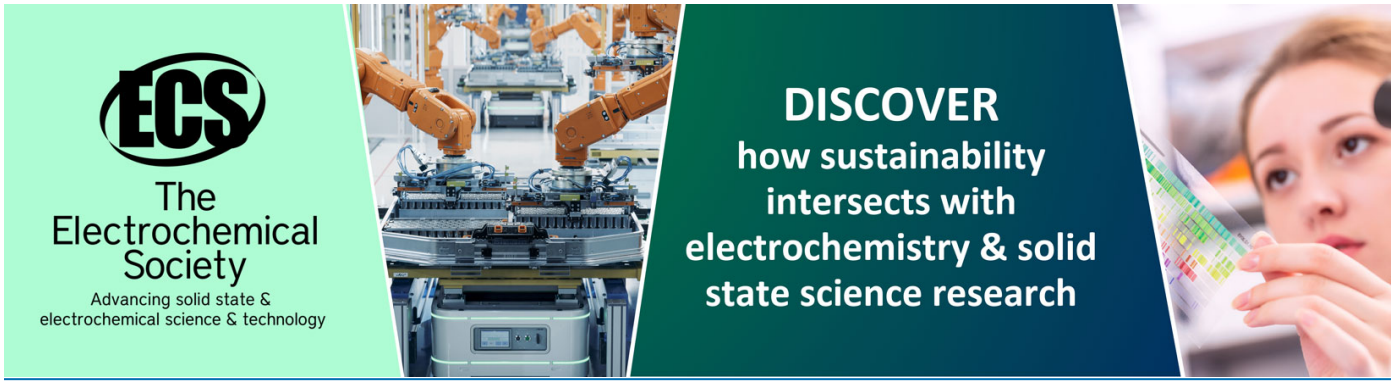
Research on Semantic Segmentation of Portraits Based on Improved Deeplabv3 +



IOP Conf. Series: Materials Science and Engineering **806** (2020) 012057 doi:10.1088/1757-899X/806/1/012057

Research on Semantic Segmentation of Portraits Based on Improved Deeplabv3 +

Kaihui Zhang[1*], Xianhui Liu[1] and Yufei Chen[1]

[1]College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China

[*]Corresponding author's e-mail: 418277390@qq.com

Abstract. At present, semantic segmentation is one of the hottest research directions in the field of computer vision, and the Deeplab series is one of the best neural network models in semantic segmentation. Based on the original Deeplab-v3 + neural network model, this paper makes a series of improvements to make it perform better on semantic segmentation of portraits: 1. Add channel attention module to speed up the convergence speed of model training and improve segmentation accuracy. 2. Improve the original AtrousSpatial Pyramid Pooling (ASPP), adjust the receptive field of the network, and improve the segmentation accuracy. Finally, this paper compares the segmentation performance of different semantic segmentation network models on human-parsing dataset and proves that the proposed neural network structure has better segmentation effect.

1. Introduction

Portrait semantic segmentation is to segment the human body from the image containing the human body and perform detailed segmentation processing on each part of the human body (Figure 1). It has high requirements for the accuracy and real-time of image segmentation,so it is a challenging task. As the basis of human behavior analysis and understanding, the quality of the semantic segmentation of portraits directly determines the effectiveness of subsequent work, such as 3D model of human bodies, motion recognition, detection and tracking. Due to these challenges of portrait semantic segmentation, the research and application of portrait semantic segmentation are still in the exploratory stage[1].

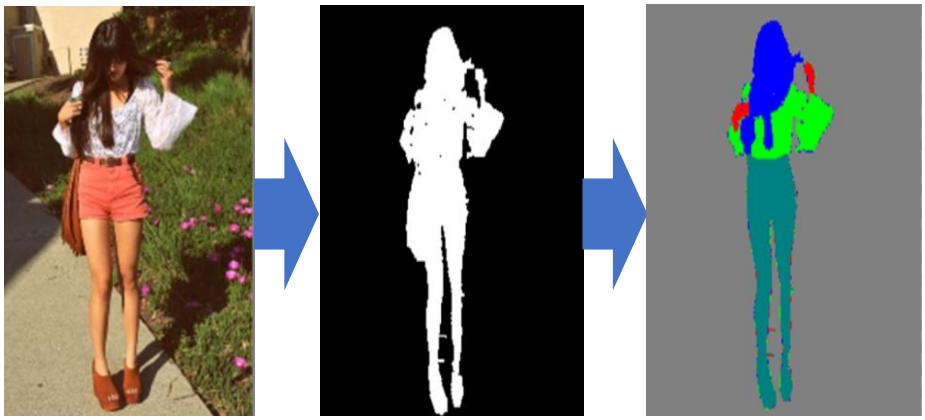


Figure 1.Portrait semantic segmentation

The essence of portrait segmentation is to divide each pixel in the image into a foreground and a

background, and continue to subdivide the pixels that belong to the foreground into sections such as the head, upper body, arms, and legs. There are two research directions in the field of semantic segmentation: one is image semantic segmentation based on image graphics, and the other is image semantic segmentation based on deep learning.

The methods based on image graphics generally use the pixel values in the image for analysis. For example, the colour component Red-Green-Blue(RGB) is used as the feature input of the colour feature rough classifier to calculate the colour similarity feature grayscale map. Sobel or Canny operators are used to find target boundaries to achieve image semantic segmentation [2].

The emergence of convolutional neural networks(CNN) provides a new research direction for the study of semantic segmentation. CNN have powerful feature extraction capabilities and have achieved satisfactory segmentation accuracy on human semantic segmentation. With the introduction of fully convolutional networks(FCN)[3], the research on end-to-end semantic segmentation methods has made a major breakthrough, which is a milestone progress. Subsequently, Seg-Net[4], U-Net[5], Deeplab series and other neural networks for semantic segmentation appeared, which further improved the accuracy of segmentation.

The rest of this thesis consists of the following: The section2 introduces the related research and work before the experiment. In section3, the thesis introduces the improved principles and methods of the algorithm in detail. In section4, we carry out comparative experiments and analyse the experimental results. Finally, we conclude this paper in section 5.

2. Related Work

The Deeplab series is one of the most popular CNN architectures in the field of semantic segmentation. Since 2015, v1[6], v2[7], v3[8] and v3 + series have been successively proposed. When traditional CNNs are used for feature extraction, spatial information is often lost during down-sampling and pooling processing, which causes the problem of reduced resolution. In response to this problem, Deeplabv1 proposed atrousconvolution to expand the perceptual field to obtain more contextual information and reduce the loss of location information. Inspired by SPP-Net[9], the v2 network proposed an AtrousSpatial Pyramid Pooling(ASPP) structure, which used multiple atrousconvolution to extract features in parallel and then fused the features. In addition, v2 replaced the backbone network from VGG to resnet-101. v3 is a further improvement of ASPP. It uses deep-level atrousconvolution and adds BN layers, and then the obtained feature map is restored by bilinear interpolation.

Deeplabv3+ is currently the latest neural network structure of the Deeplab series, which is mainly improved based on Deeplabv3. This network mainly borrows the traditional encoder-decoder architecture,expands a simple and effective module for recovering boundary information. Based on the above-mentioned encoder-decoder architecture, inspired by work such as Xception[10], deep separation convolution was applied to ASPP and decoder modules to quickly calculate and maintain the powerful learning ability of the model.

Based on the original Deeplabv3 + architecture, this paper makes the following improvements to the problem of portrait semantic segmentation: introduces the attention module of the channel, improves the accuracy of model segmentation, and accelerates the model's convergence speed. At the same time, the ASPP structure is improved, a smaller convolution kernel is used, and the size of the rate is reduced to improve the accuracy of network segmentation in thin parts.

3. Methods

In this section, we first introduce the working principle of the channel-attention module and its application in this research.Second, we explain in detail the application and improvement of the ASPP structure.Finally, we give the final Improved Deeplabv3 + network structure.

3.1. Channel Attention Module

The essence of channel-attention is to adopt a new "feature recalibration" strategy. Specifically, the

importance of each feature channel is automatically obtained through learning, and then according to this importance, useful features are promoted and features that are not useful for the current task are suppressed. This paper introduces the Squeeze-and-Excitation (SE) block in SE-Net for channel-attention learning, as shown in Figure 2:

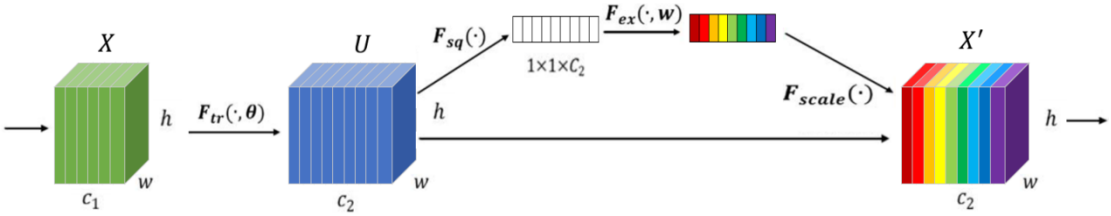


Figure 2.Squeeze-and-Excitation (SE) block

According to figure2, a feature map X is first given, and the number of feature channels is c_1. After a series of general transformations, a feature map U with the number of feature channels c_2 is obtained:

$$F_{tr}: X \to U, X \in R^{c_1 \times h \times w}, U \in R^{c_2 \times h \times w} (1)$$

The formula of F_{tr} is shown in (3) (convolution operation, v_c represents the c-th convolution kernel, and x^s represents the s-th input):

$$u_c = v_c * X = \sum_{s=1}^{c_2} v_c^s * x^s (2)$$

Next, we need to perform the squeeze operation, that is, perform the global_average_pooling operation: generate a vector with global receptive fields:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) (3)$$

The next step is excitation operation, as shown in (4):

$$s = F_{ex}(z, W) = \sigma\big(g(z, W)\big) = \sigma(W_2 \delta(W_1 z)) (4)$$

It is the core of the entire attention mechanism. First multiply z by W_1, which is a fully connected layer operation. The dimension of W_1 is C / r * C. This r is a scaling parameter. In this paper, r is 16. The purpose of this parameter is to reduce the number of channels and thus reduce the amount of calculation. And because the dimension of z is 1 * 1 * C, the result of W_1 z is 1 * 1 * C / r; then, after passing through a relufunction, the output dimension is unchanged; then multiply by W_2 and multiply by W_2. It is also a fully connected layer process, where the dimension of W_2 is C * C / r, so the dimension of the output is 1 * 1 * C; finally, it passes the sigmoid function to obtain s.

After s is obtained, channel-wise multiplication can be performed, as shown in (5). Where u_c is a two-dimensional matrix, and s_c is a number, which is the weight, so it is equivalent to multiplying each value in the s_c matrix by s_c. That corresponds to F_{scale} in figure 2[11].

$$X' = F_{scale}(u_c, s_c) = s_c \cdot u_c (5)$$

3.2. ASPP

Because in the semantic segmentation problem, there are large and small instances that need to be segmented. If the convolution kernels of the same size are used, there may be a problem that the receptive field is not large enough, and the accuracy of segmentation of large objects may decrease. In response to this problem, atrous convolution emerged at the historic moment, that is, the receptive field of the convolution kernel was changed by adjusting the size of the dilation rate.

However, the effect of using atrous convolution on a branch convolutional neural network is not good. For example, in portrait segmentation, we use large atrous convolutions to obtain the upper body information of the portrait, but it is not effective for the feature extraction of smaller hands and feet. If we continue to use smaller atrous convolutions to re-acquire the information of small objects, there will be a lot of redundancy.

ASPP uses the dilation rate of different sizes to capture multi-scale information on different scales on the network decoder. Each scale is an independent branch. It is merged at the end of the network and then a convolutional layer is output to predict the label. Infigure 3. This design effectively avoids

the acquisition of redundant information on the encoder and directly focuses on the correlation with the object.

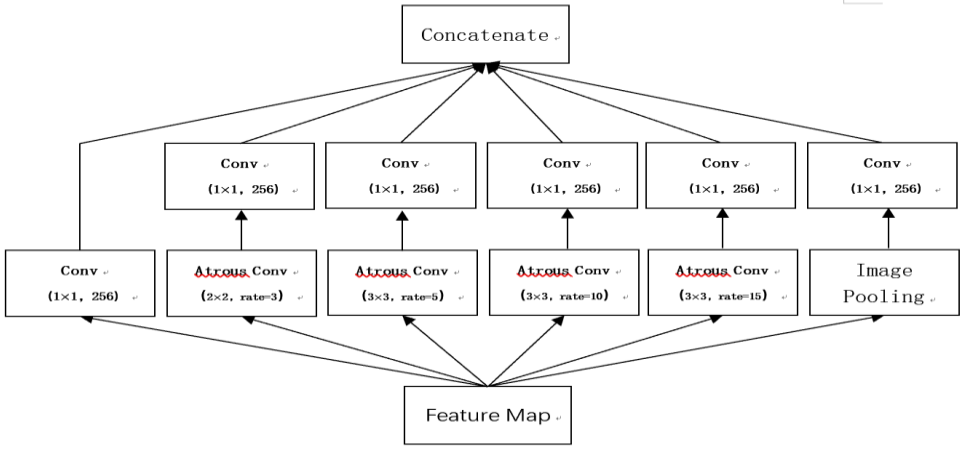


Figure 3.ASPP Structure

Among them, after each convolution, BatchNormalization is used, and relu is used as the activation function. Finally, the results obtained by branch convolution are Concatenated.

3.3. Improved Deeplabv3+

This paper is based on the original Deeplabv3+ network[12], of which BackBone uses an improved Xception network. This paper mainly makes 2 improvements based on the original network:

1. Improve the original ASPP, add 2 × 2 hole convolution, and adjust the dilation rate to better detect the portrait part.

2. The feature map spliced after ASPP branch convolution, because the channel is very large. The paper uses channel attention to process to improve the segmentation accuracy and speed up the model convergence speed.

After these two improvements, the resulting Improved Deeplabv3 + network structure is shown in Figure 4:

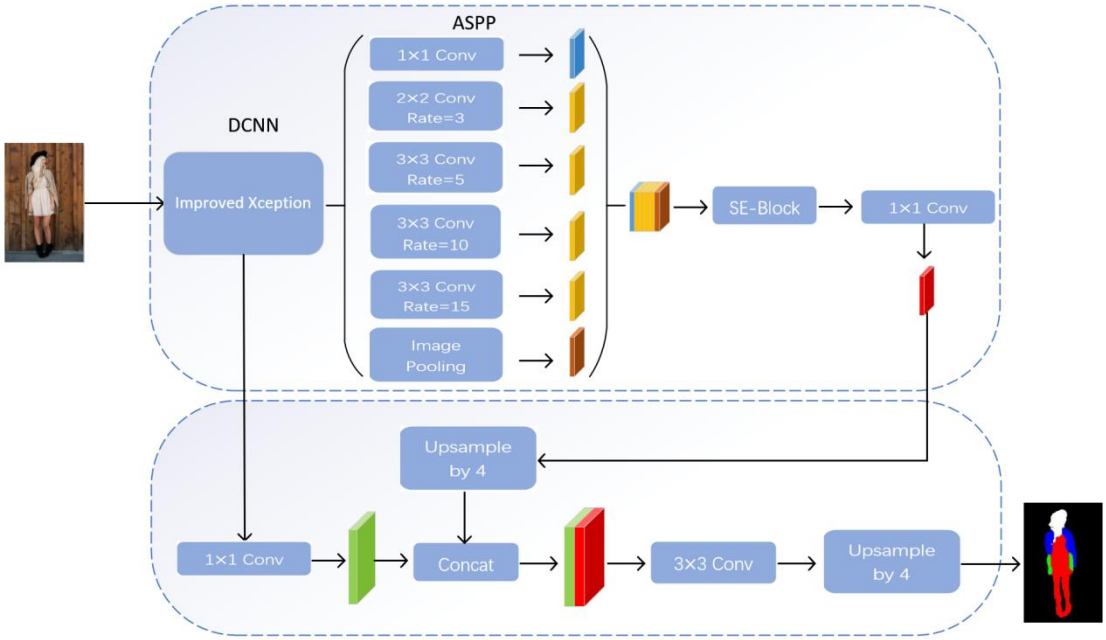


Figure 4. Improved Deeplabv3 + network structure

4. Experiment

4.1. dataset

This paper uses HumanParsing-Dataset data set for experiments. The dataset contains 17,706 images, including a total of 16 + 1 object classes. The dataset is segmented for portrait details, and hats, tops, and pants are segmented. In this experiment, the segmented regions are merged and divided into 4 + 1 categories, which are background, head, upper body, hands, and lower body. In this project, 12,706 pictures were selected. After data enhancement, it was expanded to 63,539 pictures as the training set, 3000 pictures as the verification set, and the rest as the test set.

4.2. loss function

In this experiment, categorical_crossentropy is used as the loss function during training. Suppose there are K label values, and the probability that the i-th sample predicts the k-th label value is$p_{i,k}$, that is, $p_{i,k} = \Pr(t_{i,k} = 1)$. If there are N samples in total, the loss function of the data set is:

$$L_{log}(Y, P) = -logPr(Y|P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} log p_{i,k} \quad (6)$$

4.3. results

This experiment uses mIOU (Mean Intersection over Union) as an index to measure the quality of model segmentation, and uses Seg-Net, U-net, Deeplabv3 + and the Improved Deeplabv3+ in this paper for comparison experiments. As shown in Table 1, the Improved Deeplabv3+ has significantly improved the segmentation effect.

Table 1.Performance of each segmented network model on the human-parsing dataset

Method	Head	Upperbody	Both hands	Lower body	Mean
SegNet	0.404	0.261	0.027	0.379	0.268
U-net	0.665	0.486	0.270	0.665	0.522
Deeplabv3+	0.708	0.667	0.499	0.845	0.679
Improved Deeplabv3+	**0.715**	**0.685**	**0.512**	**0.854**	**0.691**

Deeplabv3+has two kinds of Backbone networks: ResNet-101 and Xception. The experimental results are shown in Table 2.

Table 2. Performance of the two Backbone networks on the human-parsing dataset

BackboneNetwork	mIOU
Improved Deeplabv3+(ResNet-101)	0.673
Improved Deeplabv3+(Xception)	**0.691**

In order to prove that the improved ASPP does improve the accuracy of portrait semantic segmentation, this article also did a comparison experiment using the improved ASPP and the original ASPP, as shown in Table 3.

Table 3. Performance of the two ASPP on the human-parsing dataset

BackboneNetwork	mIOU
Improved Deeplabv3+(Original ASPP)	0.657
Improved Deeplabv3+(Improved ASPP)	**0.691**

By observing the training process of the model, we also found that adding the channel-attention module can speed up the model's convergence speed, as shown in Figure 5.

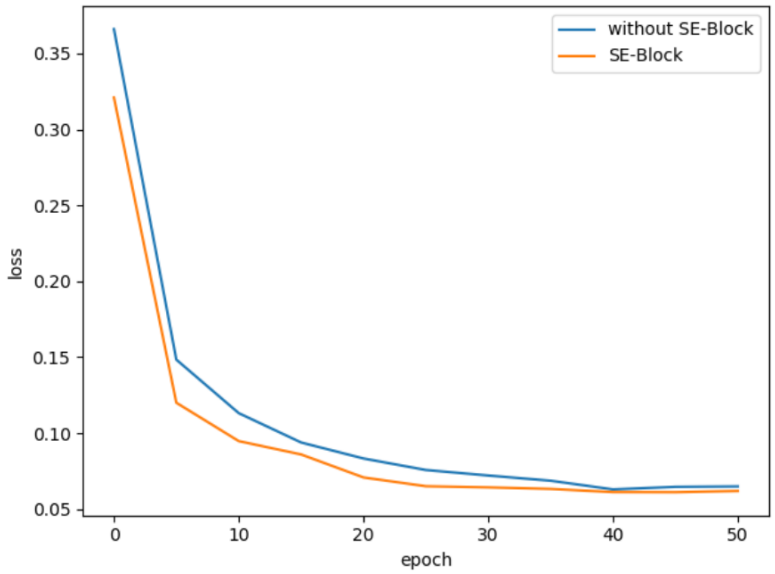


Figure 5. Comparison of Improved Deeplabv3+(SE-Block and without SE-Block)

5. Conclusion

In this paper, we have made a series of improvements based on the original Deeplabv3 + to improve the convergence speed of the training model and the accuracy of segmentation. In addition, a series of comparative experiments have been done in this article. By comparing the experimental results, it has been proved that the use of Xception as the backbone, Improved Deeplabv3+ and ASPP, as well as SE-block have the best segmentation effect on the human-parsing dataset.

Acknowledgment

This paper was financially supported by National Key R&D Program of China (2017YFB0304102) and Science and Technology Innovation Project of Shanghai (18511107400).

References

[1] Ma Z , Wang N , Gao X , et al. Semantic Segmentation Based Automatic Two-Tone Portrait Synthesis[J]. 2017.
[2] Aissa A A , Duchesne C , Rodrigue D . Polymer powders mixing part II: Multi-component mixing dynamics using RGB color analysis[J]. Chemical Engineering Science, 2010, 65(12):3729-3738.
[3] Long J ,Shelhamer E , Darrell T . Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[4] BadrinarayananV , Kendall A , Cipolla R . SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. 2015.
[5] RonnebergerO , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. 2015.
[6] Chen L C , Papandreou G , Kokkinos I , et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4):357-361.
[7] Chen L C , Papandreou G , Kokkinos I , et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4):834-848.
[8] Chen L C , Papandreou G , Schroff F , et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017.

[9] He K , Zhang X , Ren S , et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.
[10] Chollet, François. Xception: Deep Learning with Depthwise Separable Convolutions[J]. 2016.
[11] Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
[12] Chen L C ,ZhuY ,Papandreou G , et al. Encoder-Decoder with AtrousSeparableConvolution for Semantic Image Segmentation [J]. 2018.