DA-Res2Net: a novel Densely connected residual Attention network for image semantic segmentation

Xiaopin Zhao[1], Weibin Liu[1*], Weiwei Xing[2] and Xiang Wei[2]
[1] Institute of Information Science, Beijing Jiaotong University
Beijing 100044, China
[2] School of Software Engineering, Beijing Jiaotong University
Beijing 100044, China
*Corresponding author: Weibin Liu, wbliu@bjtu.edu.cn

Received April 5, 2020; revised July 19, 2020; accepted November 4, 2020; published November 30, 2020

Abstract

Since scene segmentation is becoming a hot topic in the field of autonomous driving and medical image analysis, researchers are actively trying new methods to improve segmentation accuracy. At present, the main issues in image semantic segmentation are intra-class inconsistency and inter-class indistinction. From our analysis, the lack of global information as well as macroscopic discrimination on the object are the two main reasons. In this paper, we propose a Densely connected residual Attention network (DA-Res2Net) which consists of a dense residual network and channel attention guidance module to deal with these problems and improve the accuracy of image segmentation. Specifically, in order to make the extracted features equipped with stronger multi-scale characteristics, a densely connected residual network is proposed as a feature extractor. Furthermore, to improve the representativeness of each channel feature, we design a Channel-Attention-Guide module to make the model focusing on the high-level semantic features and low-level location features simultaneously. Experimental results show that the method achieves significant performance on various datasets. Compared to other state-of-the-art methods, the proposed method reaches the mean IOU accuracy of 83.2% on PASCAL VOC 2012 and 79.7% on Cityscapes dataset, respectively.

Keywords: Semantic segmentation, Densely connected, Attention network, Channel-Attention-Guide module, Feature fusion.

1. Introduction

Semantic segmentation [1] which involves taking some raw data such as a flat image as input and converting them into a mask with a highlighted region of interest is a typical computer vision problem. The goal of Semantic segmentation is to assign each pixel in the image a category label and identify the content and position in the image by finding all the pixels that belong to it. Semantic segmentation has great prospects in geological monitoring, automatic automatic driving, facial segmentation and other fields. With the development of convolutional neural networks, fully convolutional network (FCN) [2] has achieved outstanding results in the field of semantic image segmentation. Compared to traditional non-parametric methods, these methods basing on fully convolutional neural network improve the accuracy of the scene analysis task to a considerable extent. The performance improvement comes mainly from multiple convolutional and nonlinear activation layers that learn the specific local features of the data and assign each pixel a category label on the local area. However, as the convolution and pooling operations will inevitably fill out some detailed information of the image, which resulting in intra-class inconsistency [3] as shown in **Fig. 1** and inter-class indistinction [3] as shown in **Fig. 2**.

(a) **(b)** **(c)**

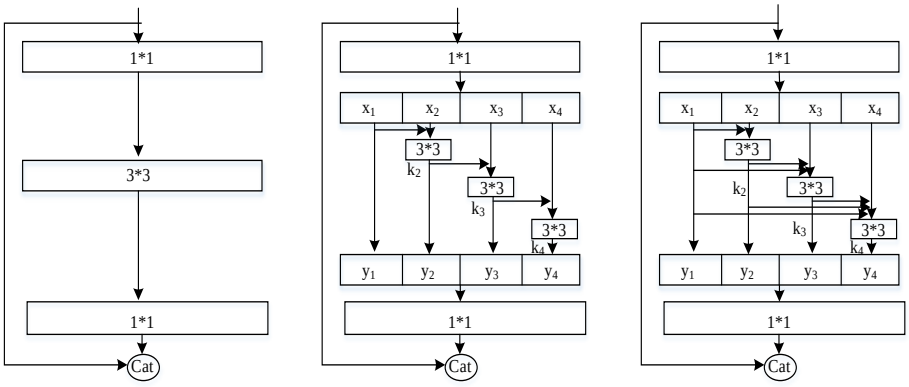
Fig. 1. This is the **Intra-class inconsistency**. The computer case has the similar blue light and black shell with the computer screen (a)Image, (b)Segmentation result, (c)Ground truth.

(a) **(b)** **(c)**

Fig. 2. This is the **Inter-class indistinction**. The left bottom corner of the cow is recognized as a horse. (a)Image, (b)Segmentation result, (c)Ground truth.

From our analysis, the main reason of the two problems is that the multi-scale information extracted from image is not sufficient. Suppose a multi-scale representation is to be obtained in a visual task, the acceptance domain of the feature extractor will be large enough to describe objects/parts/contexts of different scales. The convolutional neural network gradually learns

from coarse features to fine features by using convolution operations, so CNN's multi-scale feature extraction methods (such as residual network [4]) are used in many visual tasks. Then, how to design a more efficient network architecture is the key to improve multi-scale representation.



(a) (b) (c)

Fig. 3. Network structure: (a)Bottleneck block, (b)Res2Net module, (c)The proposed Dense-Res2Net module.

Based on the Res2Net network [5] of the **Fig. 3(b)**, we propose Dense-Res2Net as shown in **Fig. 3(c)** which is an excellent multi-scale variance of the existing network architecture and extracts tiny features of different scales in a more intensive way. The Dense-Res2Net merges the features of different scales together to enhance the contextual relationship between different regions of the image. Furthermore, the Channel-Attention-Guide (CAG) module is added to the network to improve the feature selecting capabilities of the network. We evaluate our proposed network on the PASCAL VOC 2012 and Cityscapes dataset, and compare it with PSPNet [6], Deeplab v2-CRF [7], DFN [3]and so on.

The main contributions of the paper are:

- A new densely connected residual network is proposed as a feature extractor, which fusions the features of different scales to enhance the contextual connection between different areas of the images.
- The Channel-Attention-Guide module (CAG) module is proposed to change the problem of equal weighting of each channel feature and select informative features more efficiently.
- We combine the above two components to design a Densely connected residual Attention network (DA-Res2Net). We verify the effectiveness of the proposed network through

sufficient research which achieves optimal results on two scenes analysis benchmarks including Cityscapes and PASCAL VOC 2012.

The rest of the structure of this paper is organized as follows: The second part introduces the classical methods of semantic segmentation. The third part introduces our proposed segmentation network and theoretical explanation. The fourth part is an experiment comparing our proposed network with other classical methods, which proves that our improvement is effective. The last part of the paper gives the conclusion.

2. Related Work

In this section, we summarize the general methods for solving the semantic segmentation problem into two types. One is to find a way to fuse multi-scale information, and the other is a probability map model.

2.1 Multi-scale based model

Generally, convolutional neural network (CNN) is connected to a number of fully connected layers after convolution, and the feature map generated by the convolution layer is mapped into a fixed length feature vector. The general CNN structure is suitable for image level classification and regression tasks as they ultimately expect a probability of classification of the input image. Different from the classic CNN in the convolutional layer using the fully connected layer to obtain fixed-length feature vectors for classification, Long et al. propose fully convolutional network (FCN) [2] that can accept input images of any size and use the deconvolution layer to up-sample the feature map of the last volume base layer to return the same size as the input image. Therefore, it can generate a prediction for each pixel while preserving the spatial information in the original input image. Another major innovation of FCN is the skip structure combined with the results of different deep layers. This idea also plays an important role in the subsequent networks.

As objects often occur at multi-scale in natural senses. For the same object, if there is a lack of macro understanding of things, it is very likely that the target category will be misjudged. Therefore, it is necessary to learn more distinguishing and effective features. However, most of the current segmentation algorithms cannot fully utilize the multi-scale information of the image. In a general convolutional network, each layer extracts local feature information of an object, but local information is difficult to provide a higher level of semantic information. In order to reduce the dimension of the feature maps and increase the perceived domain of subsequent layers, Rabinovich et al. proposes Parsenet [7] that combines global features that improve segmentation performance. In many segmentation papers, the concept of multi-scale information is also used such as PSPNet [6] and DeeplabV3 [8].

Attention model [Li, 2019] has been widely used in various fields of deep learning in recent years such as image processing, speech recognition or natural language processing which obtains the target area that needs to be focused on by quickly scanning the global image. For example, in the Seq2seq [10] model, the encoding process of the original encoder-decoder model generates an intermediate vector C which is used to store the semantic information of the original sequence. Due to the length of the vector is fixed, the vector C cannot save all

semantic information. Contextual semantic information is limited so as to limit the understanding of the model. The attention mechanism whose principle is to calculate the degree of matching between the current input sequence and the output vector breaks the limitation of the original encoder-decoder model on fixed vectors. Shen et al. propose a squeeze-and-excitation networks (SENet) [11] which uses a new feature recalibration strategy to explicitly model the interdependencies between feature channels. SENet learns how important each feature channel is by learning while it enhances useful features and suppresses features that are of little use to the current task. The accuracy of the segmentation can be significantly improved by embedding the SE module in the building block unit of the original network structure. Wang et al. propose a discriminative feature network (DFN) [3] which involves two components: the Smooth Network (SN) and the Border Network (BN). The SN is combined with the adjacent phase feature where advanced features provide semantics to guide the selection of low-level features to achieve more discriminative features.

2.2 CRF / MRF based model

Markov Random Field (MRF) [12] and its variant conditional random fields are widely used in the classical framework of semantic segmentation, which improves the classification of pixels often resulting in unsatisfactory results and the actual image of visual characteristics does not match. CRF describes the relationship between pixels and pixels, which encourages similar pixels to assign the same label while pixels with larger differences assign different labels. Therefore, the CRF can split the image as much as possible at the boundary, treating each pixel as a CRF node. This structure is typically added to the network as a post-processing module that combines the scores from the classifier with locally captured information to complete the pixel's interaction with the edge or super-pixel

Krahenbuhl et al. propose a fully connected CRF (DenseCRF) [13] model, in which pairwise edge potentials are defined by a linear combination of Gaussian kernels. Falong Shen et al. propose a joint FCN and CRF model (SegModel) [14], which integrates segment-specified features that constitute high-level context and boundary guidance for semantic segmentation. Shen et al. propose a Deep Parsing Network (DPN) [15] which models a unitary item and a pair of items (ie, a mixture of high-order relationships and label contexts) in a single CNN and achieves high performance by extending the VGG network and adding some layers for modeling pairwise terms. Using the deep information as a conditional random supplemental information, Jiang et al. propose a depth-sensitive full-join condition combined with a full convolutional network (DFCN-DCRF) [16] for the random field, whose idea is to completely add deep information to the expanded FCN. These methods couple a fully connected CRF with a DCNN to produce accurate predictions and detailed segmentation to improve performance.

3. DA-Res2Net

In this section, we will detail our DA-Res2Net model. In subsection 3.1, we present the overall architecture of the network, as shown in Fig 1, which clearly shows the entire implementation process of the algorithm. In subsections 3.2 and 3.3, we will describe in detail two important components of our architecture: the dense residual network and the Channel-Attention-Guide

(CAG) module, which are the most noteworthy parts of this paper.

3.1 Overview

In this subsection, we give our overall network architecture as shown in **Fig. 4**, which consists of (a) (b) (c) (d) parts. It is noticed that (b) module is the highlight of this paper.

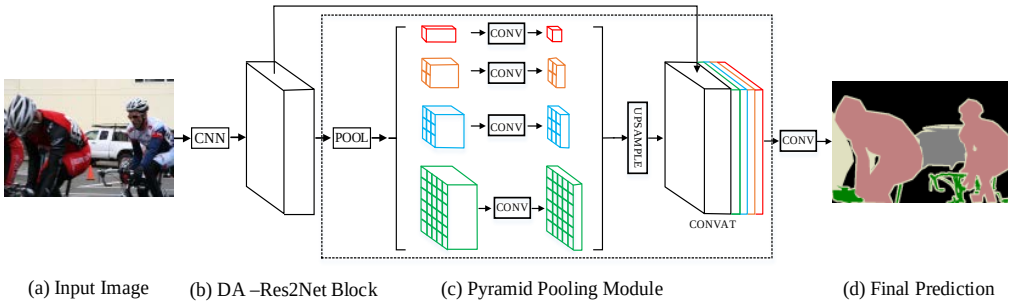


Fig. 4. Overall network architecture.

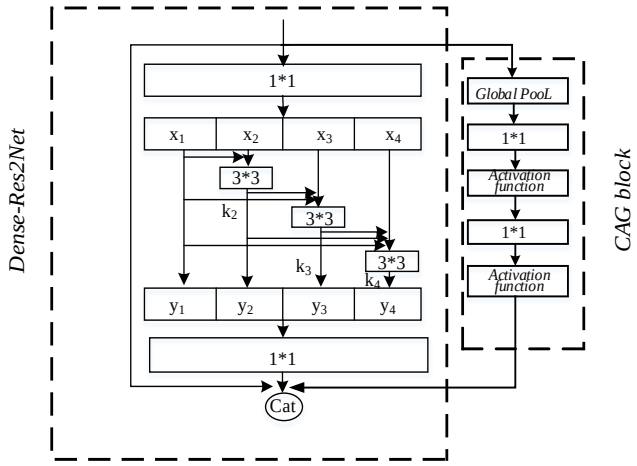


Fig. 5. The detail of DA-Res2Net block

(a)Input image: We take PASCAL VOC 2012 and Cityscapes dataset as input.

(b)DA-Res2Net: As shown in **Fig. 5**, DA-Res2Net is a Densely connected residual Attention network which makes the connection between these small filters more densely and the extracted features are more plentiful. By embedding the CAG module into the dense residual network, we can get a new and more efficient feature extractor which can enhance useful features and suppress useless features. Therefore, the feature recalibration on the spatial channel is completed.

(c)Pyramid Pooling Module: The pyramid pooling module combines features from four different scales. The global pooling is to obtain the overall information of the image, which is a rough expression. The following three pyramid level convolution kernels gradually become larger which extract image information more subtle. The features after these four dimensions are adjusted to the same size by 1*1 convolution and connected to form the final feature representation of the input image. This low-dimensional feature is up-sampled or deconvoluted to obtain a segmented image of the same size as the input image.

(d)Final Prediction: The final output image is a predictive segmentation map, where each type of object is represented by a color.

3.2 Dense-Res2Net

In this subsection, we introduce the network structure of Dense-Res2Net, which is a more fine-grained feature fusion. The connection between the different small residual blocks becomes denser and the receptive field is further enlarged to extract more subtle multi-scale information.

The bottleneck structures shown in **Fig. 3 (a) and (b)** are the basic components of the existing convolutional neural network, for example, ResNet [4], ResNext [17], Res2Net [5]. Unlike the general hierarchical multi-scale representation, Res2Net improves the representation of multi-scale in a more granular way. Based on Res2Net, we add more connections that can learn identity mapping to enhance feature reusability. We replace the n channel $3*3$ filters with a smaller set of filters that are internally connected in a densely hierarchical residual-like style. Benefiting from the DenseNet [18] network, this dense connection can get a bigger receptive field. Moreover, it also helps to alleviate the problem of gradient dispersion and makes training faster.

As shown in **Fig. 5**, the output after $1*1$ convolution is equally divided into s blocks according to the number of channels (In **Fig. 5**, s takes 4), and each part is donated by x_i, where $i \in \{1, 2, ...s\}$. Each x_i has a corresponding $3*3$ convolution which represented by $K_i()$. We take y_i to represent the output of $K_i()$. The current y_i is equal to the current x_i and the output of each previous stage $y_{i-1}, y_{i-2}, \ldots y_1$. Therefore, y_i can be described as:

$$y_i = \begin{cases} x_i & i=1 \\ k_i(x_i + y_{i-1} + y_{i-2} + ... + y_1) & 1 \le i \le s \end{cases} \quad (1)$$

According to formula (1), we replace the Res2Net residual block with the Dense-$3*3$ Group = c block except for x_i, the $3*3$ convolution under each of the other modules combines the current input with the previous output. Because of the dense connection and the splicing operation, the out of Dense-Res2Net module contains different number and different combinations of receptive field sizes/scales. The receptive field in the final output feature map becomes very large. In the Dense-Res2Net module, we use a dense connection strategy that

facilitates the extraction of multi-scale and local information from images. At the same time, each output of the Dense-Res2Net Module is $y_i = k_i(x_i + y_{i-1} + y_{i-2} + ...y_1)$ except for x_i and final output is $Y = H(y_1 y_2, y_3, y_4)$. If some measures are not taken to control the size of the model, the accumulation of multiple groups will make the input too large so as to the network is paralyzed. In order to control the depth of the model, we splice the operation in each group followed by a $1*1$ convolution. Furthermore, the $1*1$ convolution operation fuses information of different scales and also reduces the number of parameters. The split and concatenation strategy is a fashion of feature fusion.

Motivating the idea of the attention mechanism, we design a CAG module which uses the inter-dependencies between channels to recalibrate features. Firstly, the global pooling turns each two-dimensional feature map into a real number which are connected to form a global vector that can represent the global receptive field. Secondly, it is $1*1$ convolution and activation function.

3.3 Channel-Attention-Guide module

The purpose of the Channel-Attention-Guide (CAG) module is to solve the problem of equal weighting of each channel feature. The output of the convolution operation is a score graph that calculates the probability of each pixel on each category. The final predicted label for each pixel is the category with the highest probability. This probability value is calculated by default when the weights of different channels are equal. However, the characteristics of different channels have different degrees of discrimination resulting in different consistency of prediction. In order to obtain intra-class consistent, we should extract the discriminant features and suppress the non-discriminating features.

Motivating the idea of the attention mechanism, we design a CAG module which uses the inter-dependencies between channels to recalibrate features. Firstly, the global pooling turns each two-dimensional feature map into a real number which are connected to form a global vector that can represent the global receptive field. Secondly, it is 1*1 convolution and activation function. The diagram illustrating the structure of the CAG block is shown in **Fig. 5**.

We change the number of channels by the $1*1$ convolution global vector generated in the previous stage, and then pass a nonlinear activation function which enhances the nonlinear fitting ability of the vector. The final module output will be a ‘colored vector’ that is more satisfied with the actual phenomenon after nonlinear operation. The CAG module allows the network to learn the importance of each channel feature autonomously, which gives the useful features a relatively big weight and a small weight to features that are less useful.

In the segmentation model that does not have the attention mechanism, the output of the convolution operation is a probability map of each variable belonging to each category, and the final score is added to the final feature map. As shown in (2):

$$y_k = F\left(x; w\right) = \sum_{i=1,\, j=1}^{D} w_{i,j} x_{i,j} \tag{2}$$

Where x is the feature map output by the network. w represents the convolution operation. $k \in \left\{1, 2, \ldots k\right\}$, k is the number of channels. D represents the collection of all pixel positions. The formula implicitly indicates that the weight of each channel is the same. However, the characteristics of different channels in different stages have different characteristics, which may lead to inconsistencies in prediction.

$$\sigma_i\left(y_k\right) = \frac{\exp\left(y_k\right)}{\sum_{j=1}^{K} \exp\left(y_j\right)} \tag{3}$$

Where σ represents the predicted probability value. y represents the output of the network. It can be seen from formula (2) and formula (3) that the final predicted label has the highest probability value. Suppose that the predicted label for a pixel is y_0 and the true label is y_1 this section changes the highest probability value by introducing attention vector parameters. As shown in (4):

$$\overline{y} = \alpha y = \begin{bmatrix} \alpha_1 \\ \ldots \\ \alpha_K \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \ldots \\ y_K \end{bmatrix} = \begin{bmatrix} \alpha_1 w_1 \\ \ldots \\ \alpha_K w_K \end{bmatrix} \times \begin{bmatrix} x_1 \\ \ldots \\ x_K \end{bmatrix} \tag{4}$$

Where $\overline{y}$ represents the new predicted value. α is the output vector of the attention module. The above formula assigns channels through the channel attention module and selects features that are useful for the current task.

4. Experimental Classification Results and Analysis

In this section, we conduct the experimental results of our DA-Res2Net. In subsection 4.1, we will introduce the experimental setup. In subsections 4.2 and 4.3, we will test and analyze the performance of DA-Res2Net on the PASCAL VOC2012 and Cityscapes dataset**.**

4.1 Experimental Setup

Training: We train our network using an initial learning rate of 0.007 and decaying according to the following formula:

$$lr = lr \times \left(1 - iterations / total\_iterations\right)^{0.9} \tag{5}$$

The optimizer we used is stochastic gradient descent (SGD) with a momentum of 0.9. We take the cross entropy loss function to train the data. The model is trained on four 1080Ti of 12G memory, and we set batch size to 8. Model performance can be improved by increasing the number of iterations, where we set 50 epochs for PASCAL VOC2012 and 200 epochs for Cityscapes dataset. The evaluation criteria of the model are divided into two types. one is the visual effect of human beings and the other is the standard measure of mathematics. Now we focus on the common mathematical evaluation criteria-IOU for semantic segmentation. Intuitively, it is the ratio of the target area predicted by the model to the real area marked in ground truth. Mean Intersection-over-Union (mIoU) is a weighted average of the IoU for different categories.

Data augmentations: In order to resist overfitting, we adopt random mirror, random resize between 0.5 and 2, random rotation between -10 and 10 degrees, and random Gaussian blur.

4.2 Evaluation indicators and evaluation methods

The evaluation criteria of the model are divided into two types, one is the visual effect of human beings, and the other is the standard measure of mathematics. In this experiment we take these two evaluation methods. Now we focus on the common mathematical evaluation criteria for semantic segmentation.

$$\mathrm{IOU} = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}} \tag{6}$$

Intuitively, it is the ratio of the target area predicted by the model to the real area marked in ground truth. Mean Intersection-over-Union (mIoU) is a weighted average of the IoUs for different categories.

4.3 PASCAL VOC 2012

In this experiment, we use PASCAL VOC 2012 [19] for semantic segmentation, which contains a total of 1464 training data and 1449 test data, a total of 21 categories, one of which is background.

In the experiment, we take resnet50, resnet101, res2net50, res2net101 as our benchmark network structure and PSPNet as our segmentation baseline. All implementations are the same as PSPNet, except that the backbone network is replaced by res2net and DA-res2net. As shown in **Table 1**, res2net50 performs better than resnet50 by 1.1% and DAres2net50 is better than resnet50 by 1.6%. Res2net101 is better than resnet101 of 0.2% and DAres2net101 is better than resnet101 of 0.6%.

In addition, we design ablation experiments. When only changing the network structure and not introducing the attention mechanism module, the performance of the 50-layer and 101-layer networks is improved by 0.1% and 0.1% respectively compared to the res2net

performance. Without changing the network structure, when the attention mechanism module are introduced, the performance of the 50-layer and 101-layer networks is improved by 0.3% and 0.2% respectively compared to the res2net performance. Therefore, we can see that the improvement of model performance is mainly due to the introduction of attention mechanism.

Table 1. The performance of semantic segmentation on PASCAL VOC2012 val set.

Backbone	50-layer(mIoU)	101-layer(mIoU)
ResNet	69.3	82.6
Res2Net	70.4	82.8
Ours(without CAG)	70.5	82.9
Ours(without DenseRes2Net)	70.7	83.0
Ours	70.9	83.2

As shown in **Table 2**, We compare our proposed method with eight latest image semantic segmentation methods, namely FCN [2], ParseNet [7], DeepLabv2-CRF [13], DPN [15], Piecewise [20], SegModel [21], PSPNet [6], DFN [3]. FCN converts fully connected layers in traditional CNNs into convolutional layers and classifies images at the pixel level, thereby solving the problem of image segmentation at the semantic level. DeepLab uses atrous convolutions to expand the receptive field. The biggest contribution of ParseNet is that it uses global context for segmentation. ParseNet can perform global pooling directly on any layer in the network to obtain a feature map representing the features of the entire graph and use this feature map for segmentation. DeepLabv2 uses atrous convolution instead of downsampling in the last few largest pooling layers to calculate feature maps with higher sampling density. It proposes ASPP to sample atrous convolutions with different sampling rates for a given input to capture the image context at multiple scales. What's more, DeepLabv2's more efficient method is to use conditional random fields to enhance the model's ability to capture details. In DPN, the author uses a high-order RNN structure (HORNN) to link DenseNet and ResNet, and proves that DenseNet can extract new functions from the previous hierarchy, and ResNet is essentially a function reuse in the previous hierarchy. In Piecewise, authors use semantic information to improve semantic segmentation. The author studies the "patch-patch context" and "patchbackground context" between image regions In SegModel, the author proposes a joint goal that combines subdivision features, high-order fragments, and boundary guidance to achieve accurate semantic segmentation. The pyramid pool module proposed in PSPNet can aggregate context information of different regions, thereby improving the ability to obtain global information. DFN rethinks the task of semantic segmentation from a macro perspective, and considers semantic segmentation to assign the same semantic label to the same object area. It proposes a smooth network with global information and channel attention models to improve intra-class consistency. Our architecture is superior to other networks. Our architecture reaches 83.2% on the PASCAL VOC2012 dataset, which is 0.6% higher than the baseline.

Table 2. The performance comparison on PASCAL VOC2012

Backbone Method	**101-layer(mIoU)**
FCN-2015	62.2
ParseNet-2016	69.8
DeepLabv2-CRF-2016	71.6
DPN-2016	74.1
Piecewise-2016	75.3
SegModel-2017	81.8
PSPNet-2017	82.6
DFN-2018	82.7
Ours	**83.2**

Fig. 8(c) is the segmentation result of PSPNet. Since PSPNet only uses the spatial pyramid pooling module, it can be seen that there are still cases of category judgment errors. For example, there is an overlapping part between the horse and the railing in the second row, which results in the lack of the second half of the horse body in the segmentation diagram. The cattle in the third row cause the category judgment error due to the different colors; Cats on a row of sofas and cats on washstands, these small indoor targets are ignored by the PSPNet algorithm. For the segmentation algorithm in this paper, such as the overlapping target in the second line, DA-Res2Net supplements the second half of the horse's body to a certain extent because it retains picture-level features and richer context features. As shown in the cattle segmentation results in the third row, the DA-Res2Net dense feature extraction network corrects the misclassified pixel regions to some extent. For the segmentation of indoor small targets in the fourth, fifth, and last lines, the DA-Res2Net extracts a larger range of feature receptive fields so that a larger range of targets can be seen. The indoor table segmentation boundary is smoother and the chairs are also partially recognition. The cat next to the sofa and the cat on the washstand are both identified and segmented. Therefore, the algorithm has strong sensitivity to small target objects.

(a) (b) (c) (d)

Fig. 8. Visual improvements of VOC dataset based on DA-Res2Net. (a)Images, (b)Ground truth, (c)Baseline, (d)Ours.

4.4 Cityscapes

Cityscapes [22] is a released dataset for urban streetscapes that contains 5,000 high-quality pixel-level, accurate annotation images captured in 50 different cities without the seasons including 2,975 training charts, 500 verification drawings and 1525 test charts. It defines 19

categories containing both stuff and objects.

The statistics in the **Fig. 9** show that our network has good performance on Cityscapes compared with CRF-RNN [23], MultitaskLearning [24], TKCN [25] and other state-of-art methods. The test results on Cityscapes reaches 79.8% accuracy, better than to baseline 1.3%.

Table 3. Performance on Cityscapes

Method	Iou cla.	iIou cla.	Iou cat.
CRF-RNN-2015	62.5	34.4	82.7
FCN-2015	65.3	41.7	85.7
DPN-2015	66.8	39.1	86.0
DeepLab-2016	70.4	42.6	86.4
PSPNet-2016	78.4	56.7	90.6
MultitaskLearning-2017	78.5	57.4	89.9
TKCN-2018	79.5	61.3	91.1
Ours	**79.8**	**61.5**	**91.7**

The following images are a few examples from Cityscapes. Observed from the 'sidewalk' and 'car', our segmentation effect is more complete, supplementing the missing part of the bodywork in the baseline.

(a) (b) (c) (d)

Fig. 9. Segmentation results on Cityscapes dataset. (a)Original images, (b)Ground truth, (c) Baseline, (d)Ours.

5. Conclusion

In this paper, we propose a new DA-Res2Net deep network model for image semantic segmentation. The dense connection structure inside our model improves the receptive field of objects at the level of granularity, which extracts more multi-scale information and fuse different levels of features for image semantic segmentation. We also add a channel attention mechanism into the network that is used as a feature filter to enhance useful feature information and weaken those useless ones. We validate the proposed architecture segmentation performance on PASCAL VOC2012 and Cityscapes benchmarks.

Acknowledgments

This research is partially supported by National Natural Science Foundation of China (No.61876018, No.61976017, No.61906014).

References

[1] T.Hui.O and K.Ma.K, "Semantic image segmentation using oriented pattern analysis," in *Proc. of IEEE Conference on Information, Communications & Signal Processing*, pp. 13-16, 2011. Article (CrossRef Link)

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.Article (CrossRef Link)

[3] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. sang, "Learning a Discriminative Feature Network for Semantic Segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1857-1866, 2018. Article (CrossRef Link)

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. Article (CrossRef Link)

[5] S. H. Gao, M. M. Cheng, K. Zhao, et al., "Res2Net: A New Multi-scale Backbone Architecture," in *Proc. of IEEE TPAMI 2020*, arXiv: 1904.01169, 2019. Article (CrossRef Link)

[6] H. Zhao, J. Shi, X. Qi, Wang, X, J, Jia "Pyramid scene parsing network," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881-2890, 2017. Article (CrossRef Link)

[7] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv: 1506.04579*, 2015.

[8] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. of IEEE International Conference on Robotics and Automation*, 2020.

[9] . Li, X. Chen, Z. Zhu, et al., "Attention-guided Unified Network for Panoptic Segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7026-7035, 2019. Article (CrossRef Link)

[10] Y G. Cinar, H. Mirisaee, P. Goswami, et al., "Position-based Content Attention for Time Series Forecasting with Sequence-to-sequence RNNs," in *Proc. of International Conference on Neural Information Processing*, pp 533-544. 2017. Article (CrossRef Link)

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2017. Article (CrossRef Link)

[12] S. Z. Li, "Markov random field models in computer vision," in *Proc. of European Conference on Computer Vision*, pp 361-370, 1994. Article (CrossRef Link)

[13] P. Krahenbuhl. V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Proc. of the Advances in Neural information Processing Systems*, pp. 109-117, 2011.
[14] F. Shen, R. Gan S. Yan and G. Zeng, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, arXiv: 1511.00561, 2015.
[15] Y. Chen, J. Li, H. Xiao, S. Yan, "Dual path networks," in *Proc. of Neural Information Processing, Systems*, 2017.
[16] J. Jiang. Z. Zhang. Y. Huang. L Zheng, "Incorporating depth into both cnn and crf for indoor semantic segmentation," in *Proc. of 8th IEEE International Conference on Software Engineering and Service Science. IEEE*, pp. 525-530, 2007. Article (CrossRef Link)
[17] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, Kaiming He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492-1500, 2017.Article (CrossRef Link)
[18] G Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708, 2017.
[19] Sara Vicente, Joao Carreira, Lourdes Agapito, Jorge Batista, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41-48, 2014.Article (CrossRef Link)
[20] Guosheng Lin, Chunhua Shen, Anton van den Hengel, Ian Reid, "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks," in *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3194-3203, 2016.
[21] Falong Shen, Rui Gan, Shuicheng Yan, Gang Zeng, "Semantic Segmentation via Structured Patch Prediction, Context CRF and Guidance CRF," in *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1953-1961, 2017.Article (CrossRef Link)
[22] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele, The IT NowOxford University Press, pp. 10 – 10, 1997.
[23] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 1529-1537, 2015.Article (CrossRef Link)
[24] A. Kendall, Y. Gal, R. Cipolla, "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482-7491, 2017.Article (CrossRef Link)
[25] T. Wu, S. Tang, R. Zhang, et al, "Tree-structured Kronecker Convolutional Networks for Semantic Segmentation," in *Proc. of 2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019. Article (CrossRef Link)

Xiaopin Zhao received the B.S. degree in Communication Engineering at Hebei University in 2017. She is pursuing her M.S. degree in Signal and Information Processing at Beijing Jiaotong University, Beijing. Her research interests mainly include image semantic segmentation, image processing.

Weibin Liu received the Ph.D. degree in Signal and Information Processing from Institute of Information Science at Beijing Jiaotong University, China, in 2001. During 2001-2005, he was a researcher in Information Technology Division at Fujitsu Research and Development Center Co., LTD. Since 2005, he has been with the Institute of Information Science at Beijing Jiaotong University, where currently he is a professor in Digital Media Research Group. He was also a visiting researcher in Center for Human Modeling and Simulation at University of Pennsylvania, PA, USA during 2009-2010. His research interests include computer vision, computer graphics, image processing, virtual human and virtual environment, and pattern recognition. He is a member of the IEEE, ACM, IEICE and CCF.

Weiwei Xing received her B.S. degree in Computer Science and Technology and Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University, in 2001 and 2006 respectively. During 2011-2012, she was a visiting scholar at University of Pennsylvania. Currently, she is a professor at School of Software Engineering, Beijing Jiaotong University. Her research interests mainly include intelligent information processing and artificial intelligence.

Xiang Wei received his Ph.D degree in Software Engineering from Beijing Jiaotong University in 2019. During 2016-2017, he was a visiting scholar at University of Central Florida. From 2019 to now, he is currently a lecturer with the School of Software Engineering, Beijing Jiaotong University. His research interest focuses on intelligent information processing, especially for semi-supervised deep learning and GANs.