Excavating effective information in different stage of backbone to improve semantic segmentation results

Huihui Han*, Lei Fan
School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, Anhui, 230009, China
*Corresponding author’s e-mail: 1209061315@qq.com

Abstract To overcome the two difficulties in semantic segmentation task: 1) the presence of multi-scale objects, 2) the loss of spatial resolution, a new semantic segmentation model is designed in this paper, which can explore effective information from different stage of backbone. Extensive experiments conducted on two public benchmark datasets have proved the effectiveness of each module, and the new model achieves 73.62% mIoU on Cityscapes test set and 79.88% mIoU on PASCAL VOC 2012 test set.

1. Introduction
Semantic segmentation, which aims to provide detailed pixel-level classification of images, is extremely potential for various challenging applications, such as autonomous driving, medical image analysis, and 3D semantic segmentation. Though remarkable progress in semantic segmentation has been made, two difficulties are still existing: 1) the presence of multi-scale objects, 2) the spatial resolution loss.

A great many of works have devoted to solve above two difficulties. For capturing rich multi-scale information, atrous spatial pyramid pooling is applied in deeplabv2 [1], deeplabv3 [2], deeplabv3+ [3], and PSPNet [4]. And large kernel is utilized in [5] to capture richer global context information. For reducing the spatial resolution loss caused by the repeated combination of max-pooling and striding at consecutive layers in deep convolutional neural networks (DCNNs), deconvolutional layers and atrous convolution are utilized in many works. Moreover, some U-shape networks, namely Large Kernel Matters [5] and RefineNet [6], take advantage of low-level information to help high-level features recover objects detail. To further solve above two problems, we design a new model, which consists of Tree-structured Atrous Convolution Module (TACM), Enhanced Channel Attention Module (ECAM) and Global Spatial Attention Module (GSAM). Lots of experiments have proved the effectiveness and efficiency of our model.

2. Related work

2.1. Attention mechanism
There are two types of attention mechanism: channel attention and spatial attention, all of which are utilized to select discriminative and effective feature maps. On the one hand, channel attention is utilized in [7], OCNet [8], PSA [9] and DANet [10], which can make use of attention map to recalibrate the channel feature maps to select discriminative features. On the other hand, extensive works have explored the effectiveness of spatial attention. In CBAM [11], the spatial attention map is utilized to enhance the inter-spatial relationship of features. A self-attention mechanism is designed in GAN [12] to enable both
the generator and the discriminator to efficiently model relationships between widely separated spatial regions.

2.2. Multi-scale feature fusion
Multi-scale feature fusion plays an important role in semantic segmentation. To fuse feature maps effectively, different works design models starting from different perspective. Skip connections are applied in FCNs [13] to conduct the late fusion. Four parallel atrous convolutional layers with different rates are introduced in DeepLab-v2 [1], deeplab-v3 [2], and deeplab-v3+ [3] to fuse feature maps from multiple receptive field. Taking the advantage of arranging atrous convolution layers in parallel and in cascade, DenseASPP [14] generates features of more scales in a larger range. Short-range and long-range residual connections with identity mappings are designed in RefineNet [6], which promote the fusion of different-level feature maps and further improve the performance of the model greatly.

3. Materials and methods

3.1. The architecture of our model
To excavate more information in the backbone of Resnet 101 as much as possible to improve semantic segmentation results, we design our model from three perspectives. Firstly, as shown in Figure 1, the pretrained ResNet 101 with atrous principle, which utilizes atrous convolution with rate = 2 and rate = 4 in ‘Res-4’ and ‘Res-5’ respectively, is applied as our backbone to extract feature maps, which can reduce the spatial resolution loss to some extent. Moreover, GSAM is introduced to explore rich spatial information in shallow stage, which can effectively refine the objects details and enhance relationships between widely separated spatial regions. Secondly, TACM is appended on the top of backbone to capture multi-scale information in deep stage of the network. Thirdly, ECAM is designed to capture discriminative information by applying channel attention mechanism in middle stage of backbone. Finally, all feature maps obtained by TACM, ECAM and GSAM are aggregated together to form new feature maps space, which is sufficient in semantic information and spatial information. After the concatenation, the final prediction is obtained by an upsample operation.

3.2 Tree-structured Atrous Convolution Module (TACM)
TACM is designed to expand the receptive fields to capture hierarchical context information and represent multi-scale objects in complex background. As shown in Figure 1, based on stacking and
expansion principle, ACMs (consisting of atrous convolution, Batch Normalization (BN) and ReLU) are arranged in tree structure, which can effectively capture multi-scale feature maps. Moreover, atrous convolutions in different ACMs have different expanding rate, which can further promote the extraction of hierarchical contextual information and aggregate multi-scale feature maps.

The details of TACM can be defined as follows. The feature maps captured from backbone are taken as the input of TACM. There are three steps to expand the receptive field, in each one of which the input is divided into two information streams. In the first step, one branch of the input feature maps $F$ is flowed into ACM-1 to obtain a larger receptive field and another one is fed forward to $1 \times 1$ convolution layer, which preserves features of the current scale and reduces the number of channels to decrease the computation complexity. At the same time, the output features $T_1$ of ACM-1 are stacked with previous features through concatenation. In the same way, we can obtain $T_2$, $T_3$ in the second and third step respectively. Finally, features $T_0$, $T_1$, $T_2$ and $T_3$ of four different branches will be integrated to a new feature $T_f$, fusing more precise context features.

3.3 Enhanced Channel Attention Module (ECAM)

ECAM, as shown in Figure 1, can model and select the information among channels, which is designed to extract discriminative information. In ECAM, two convolutional layers with $3 \times 3$ and $5 \times 5$ convolution kernels respectively are utilized, which is able to provide different receptive fields and extract different features maps. Followed that one global average pooling layer is installed, which squeezes the features spatially to grab the global information among channels, then two $1 \times 1$ convolution layers named Conv3 and Conv4 generate a bottleneck. After that, a sigmoid activation layer to map the information between 0 and 1 is used to reweight the original output to generate a self-learned channel wise attention. Finally, the channel wise attention is applied to reweight the low-level feature maps to select discriminative feature maps.

3.4 Global Spatial Attention Module (GSAM)

To help the network gain enhanced details, we introduce a Global Spatial Attention Module [12], which can efficiently model relationships between widely separated spatial regions. The image features $h \in R^{c \times h \times w}$ extracted from the second block of backbone are first transformed into two feature spaces $A$ and $B$ to calculate the spatial attention $\phi_{j,i}$,

$$\phi_{j,i} = \frac{\exp(A(h)^{T} B(h_j))}{\sum_{j} \exp(A(h)^{T} B(h_j))}$$

where $A(h) = W_{a \times} \cdot h$, $B(h) = W_{a \times} \cdot h$, and $\phi_{j,i}$ indicates the extent to which the model attends to the $i$th location when synthesizing the $j$th region. After that the output $v = (v_1, v_2, \ldots, v_j, \ldots, v_d) \in R^{d \times d}$ of the attention layer can be obtained by equation (2):

$$v_i = \sum_{j=1}^{d} \phi_{j,i} \cdot Z(h_j)$$

where $Z(h) = W_{s \times} \cdot h$. In the above formulation, $W_{a} \in R^{c \times d}$, $W_{d} \in R^{c \times d}$, $W_{s} \in R^{d \times d}$ are the learned weight matrices. Besides, a scale parameter $\lambda$ is used to multiply the output of the attention layer, the result of which will be added with the input feature maps. Thus, the final output can be obtained by equation (3):

$$T_i = \lambda v_i + h_i$$

where $\lambda$ is initialized as 0. By introducing a small parameter, the network can first rely on the cues in the local position and then gradually learn to assign more weight to the non-local position.
4. Experiments and results
To testify the effectiveness of our proposals, a series of experiments are performed on two public benchmark datasets: Cityscapes dataset and PASCAL VOC 2012 dataset. Firstly, the datasets and implementation details are introduced. Secondly, extensive ablation studies are conducted to verify the contribution of each module. Finally, comparison experiments with the state-of-the-art works are conducted to demonstrate the efficiency and effectiveness of our new model.

4.1 The datasets and implementation details
4.1.1. The datasets. The Cityscapes dataset captured from 50 different cities includes 30 semantic classes, only 19 of which are given high-quality pixel-level annotations. There are three important subsets: training set composed of 2975 images, validation set made up of 500 images and testing set including 525 images. The PASCAL VOC 2012 dataset has three main subsets: training set consisting of 1464 images, validation set composed by 1449 images and testing set made up of 1456 images. The validation set are utilized in our ablation studies and the test set are used for comparison experiments.

4.1.2. Implementation details. All experiments are performed on the public platform PyTorch. And the “poly” learning rate strategy is applied, which can be obtained by equation (4):

$$
\kappa_r = \kappa_{base} \times \left(1 - \frac{\text{epoch}}{\text{max epoch}}\right)^{\text{power}}
$$

where $\kappa_{base}$ denotes the base learning rate, $\text{power}$ denotes decayed index, and $\text{max epoch}$ denotes the number of total epochs. In detail, $\kappa_{base}=0.01$, $\text{power}=0.9$, and weight decay is set 0.0001. And mini-batch stochastic gradient descent with batch size 6 and momentum 0.99 is utilized in training process. For data augmentation, random rotation between 10 and -10 degrees and random scaling between 0.5 and 2 are applied for the training process. The mean pixel intersection-over-union (mIoU) and pixel accuracy (Acc) are chosen as the performance metrics. The cross-entropy loss function is selected as optimization target to apply on each pixel over the categories in all experiments.

4.2 Ablation studies
4.2.1. Ablation Studies for TACM. Several ablation studies have been performed to explore the appropriate atrous rate in ACMs and number of ACMs. In these experiments, only TACM is installed on the top of backbone of Resnet 101 with atrous principle, and then the predicted results are obtained directly by upsample operation with bilinear interpolation. As illustrated in Table 1, the best segment results are obtained when atrous rate in ACM-1, ACM-2 and ACM-3 are set as 6, 12 and 18 respectively. Obviously, the application of TACM can achieve 2.04% improvement over baseline. Note that the baseline in Table 1 means that the outputs of pretrained ResNet 101 with atrous principle are upsampled directly to obtain the final prediction.

| Methods       | Rate   | mIoU (%) | Acc (%) |
|---------------|--------|----------|---------|
| baseline      |        | 75.75    | 94.77   |
| Baseline+ TACM| (6,12) | 76.95    | 95.29   |
| Baseline+ TACM| (6,12,18) | 77.79  | 95.33   |
| Baseline+ TACM| (6,12,18,24) | 77.61 | 95.27   |

4.2.2. Studies for ECAM. Several ablation studies are performed to find optimal number of ECAMs and where ECAMs start from. In these experiments, TACM is appended on the top of backbone and ECAMs are introduced from the middle stage of the backbone. From Table 2, we can see that, the introduction
of ECAM can improve segmentation results to some extent. Specifically, when we introduce ECAMs into ‘Res-3’, ‘Res-4’ and ‘Res-5’, the performance of the new model improves 0.92% points.

Table 2. Performance of ablation studies about ECAM on PASCAL VOC 2012 validation set.

| Methods                      | Starting from | mIoU (%) | Acc (%) |
|------------------------------|---------------|----------|---------|
| Baseline+TACM+ ECAMs         | Res-3, Res-5  | 78.16    | 95.50   |
| Baseline+TACM+ ECAMs         | Res-4, Res-5  | 78.44    | 95.49   |
| Baseline+TACM+ ECAMs         | Res-3, Res-4, Res-5 | 78.71  | 95.68   |
| Baseline+TACM+ ECAMs         | Res-2, Res-3, Res-4 | 78.53  | 95.53   |

4.2.3. Ablation Studies for GSAM. To evaluate the effectiveness of GSAM, we conduct several ablation studies, in which TACM is appended on the top of backbone, ECAMs are introduced from ‘Res-3’, ‘Res-4’ and ‘Res-5’ of backbone, and GSAM starts from shallow stage of backbone. From Table 3, we can see that the best segmentation result can be obtained when GSAM is introduced from ‘Res-2’. This is due to that the low-level feature maps are rich in spatial information, which is helpful for recovering objects details. Obviously, the introduction of GSAM helps the model improve 1.19% points.

Table 3. Performance of ablation studies about GSAM on PASCAL VOC 2012 validation set.

| Methods                      | Starting from | mIoU (%) | Acc (%) |
|------------------------------|---------------|----------|---------|
| Baseline+TACM+ ECAMs         | Res-2, Res-5  | 77.96    | 95.47   |
| Baseline+TACM+ ECAMs         | Res-3, Res-4  | 77.85    | 95.36   |
| Baseline+TACM+ ECAMs         | Res-3, Res-5  | 78.16    | 95.50   |
| Baseline+TACM+ ECAMs         | Res-4, Res-5  | 78.44    | 95.49   |
| Baseline+TACM+ ECAMs         | Res-3, Res-4, Res-5 | 78.71  | 95.68   |
| Baseline+TACM+ ECAMs         | Res-2, Res-3, Res-4 | 78.53  | 95.53   |

4.3 Comparison experiments

4.3.1. Results of contrast studies on Cityscapes dataset. Several comparison experiments with other advanced approaches have been performed to demonstrate the effectiveness of our model. Table 4 shows the results of comparison experiments and our model achieves 73.62% mIoU, which is superior to other state-of-the-art methods. To illustrate the effectiveness of our model intuitively, some visible results of experiments of our model performed on Cityscapes dataset are presented in Figure 2.

Table 4. Results of comparison experiments on Cityscapes test set.

| Methods     | mIoU (%) |
|-------------|----------|
| DPN [15]    | 66.8     |
| LRR [16]    | 69.7     |
| DeepLabv2-CRF [1] | 70.4   |
| RefineNet [6] | 73.6   |
| Our model   | 73.62    |

4.3.2. Results of contrast studies on PASCAL VOC 2012 dataset. After performing contrast studies on Cityscapes dataset, several contrast studies are conducted on PASCAL VOC 2012 dataset to further prove the superiority of our model. The results of contrast experiments can be seen in Table 5, and our model achieves 79.85% mean IoU without pre-training on extra datasets, which is better than above state-of-the-art methods. To illustrate the effectiveness of our model intuitively, some visible results of experiments of our model performed on PASCAL VOC 2012 dataset are presented in Figure 3.
Table 5. Performance of contrast experiments on PASCAL VOC 2012 test set.

| Method          | mIoU (%) | aero | bird | bus | car | cat | cow | dog | sheep | train | person | horse |
|-----------------|----------|------|------|-----|-----|-----|-----|-----|-------|-------|--------|-------|
| DeconvNet [17]  | 72.5     | 89.9 | 79.7 | 87.4| 81.2| 86.1| 77.0| 79.0| 83.4  | 80.7  | 80.2   | 80.3  |
| GCRF [18]       | 73.2     | 85.2 | 83.3 | 89.0| 82.7| 85.3| 79.5| 80.5| 85.5  | 77.3  | 81.0   | 79.3  |
| DPN [15]        | 74.1     | 87.7 | 78.4 | 89.3| 83.5| 86.1| 79.9| 81.9| 83.2  | 77.9  | 82.3   | 80.0  |
| Piecewise [19]  | 75.3     | 90.6 | 80.0 | 92.0| 85.2| 86.2| 81.2| 83.8| 83.2  | 80.8  | 84.8   | 83.9  |
| Our model       | 79.88    | 90.9 | 84.1 | 91.1| 84.9| 87.1| 80.9| 83.9| 85.4  | 81.2  | 84.4   | 82.4  |

Figure 2. Some examples of visualization results of our model conducted on Cityscapes dataset. From left to right: input image, ground-truth, the prediction of our model.

Figure 3. Some examples of visualization results of our model conducted on PASCAL VOC 2012 dataset. From left to right: input image, ground-truth, the prediction of our model.

5. Conclusion
To solve the two main problems in semantic segmentation tasks, a new model is designed starting from capturing effective information from different level of the pre-trained Resnet 101 backbone. In detail, to capture rich multi-scale information, TACM is designed and inserted in deep stage. To extract discriminative information, ECAM is proposed, which can utilize channel attention mechanism to reweight feature maps. To prevent the loss of spatial resolution, atrous principle is applied in Resnet 101 backbone, and GSAM is introduced in shallow stage to explore affluent spatial information and to enhance relationships between widely separated spatial regions. Lots of experiments on the PASCAL VOC 2012 and the Cityscapes datasets have proved the effectiveness and efficiency of our model.
References

[1] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2018) Deeplab: Semantic image segmentation with deep convolutional nets, dilated convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834–848.

[2] Chen, L., Papandreou, G., Schroff, F., Adam, H. (2017) Rethinking dilated convolution for semantic image segmentation. https://arxiv.org/abs/1706.05587.

[3] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018) Encoder-decoder with dilated separable convolution for semantic image segmentation. In: European Conference on Computer Vision. Munich, Germany. pp.833-851.

[4] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017) Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. pp. 6230-6239.

[5] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J. (2017) Large kernel matters–improve semantic segmentation by global convolutional network. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. pp.1743-1751.

[6] Lin, G., Milan, A., Shen, C., Reid, I. (2016) Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. pp. 1-5.

[7] Hu, J., Shen, L., Sun, G. (2018) Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. pp.7132-7141.

[8] Yuan, Y., Wang, J. (2018) OCNet: object context network for scene parsing. https://arxiv.org/abs/1809.00916.

[9] Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., Jia, J. (2018) PSANet: Point-wise spatial attention network for scene parsing. In: European Conference on Computer Vision. Munich, Germany. pp. 270–286.

[10] Fu, J., Liu, J., Tian, H., Fang, Z., & Lu, H. (2018) Dual attention network for scene segmentation. https://arxiv.org/abs/1809.02983

[11] Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018) CBAM: convolutional block attention module. https://arxiv.org/abs/1807.06521.

[12] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2018) Self-attention generative adversarial networks. https://arxiv.org/abs/1805.08318.

[13] Long, J., Shelhamer, E., Darrell, T. (2014) Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 39(4): 640-651.

[14] Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K. (2018) DenseASPP for semantic segmentation in street scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. pp. 3684–3692.

[15] Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X. (2015) Semantic image segmentation via deep parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. pp.1377–1385.

[16] Chollet, François. (2017) Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. pp.1800–1807.

[17] Noh, H., Hong, S., Han, B. (2015) Learning deconvolution net-work for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. pp. 1520–1528.

[18] Vemulapalli, R., Tuzel, O., Liu, M.Y., Chellapa, R. (2016) Gaussian conditional random field network for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. pp. 3224–3233.

[19] Lin, G., Shen, C., Hengel, A. V. D., Reid, I. (2016) Efficient piecewise training of deep structured models for semantic segmentation. In: IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA. pp.3194-3203.