Multiple Skip Connections and Dilated Convolutions for Semantic Segmentation

Takayoshi Yamashita
Computer Science Department
Chubu University
Aichi 487–8501, Japan
yamashita@cs.chubu.ac.jp

Hinonori Furukawa
Computer Science Department
Chubu University
Aichi 487–8501, Japan
hiro_f@vision.cs.chubu.ac.jp

Yuji Yamauchi
Robot Department
Chubu University
Aichi 487–8501, Japan
yuu@vision.cs.chubu.ac.jp

Hironobu Fujiyoshi
Robot Department
Chubu University
Aichi 487–8501, Japan
hf@cs.chubu.ac.jp

***Abstract*—We propose a scale-aware semantic segmentation method specifically for small objects. The contributions of this method are 1) to feed the features of a small region by using multiple skip connections, and 2) to extract context from multiple receptive fields by using multiple dilated convolution blocks. The proposed method has achieved high accuracy in the Cityscapes dataset. In comparison with state-of-the-art methods, it has achieved a comparative performance in category IoU and iIoU metrics.**

I. Introduction

Convolution neural networks (ConvNets) [1] achieve very high accuracy in object recognition [2] [8] [16]. In addition, these types networks can also be employed for object detection [5] and semantic segmentation [9] [10] [11] [12] [13]. Semantic segmentation is the task of recognizing object classes in an image pixel-by-pixel. In an earlier study, a bottom-up method was proposed in which clustering is performed using hand-crafted features, such as color or gradient histograms, and regions that have similar features are concatenated [3]. Advanced methods such as the Fully Convolutional Neural Network (FCN) [9] and encoder–decoder architectures such as Segnet [11] have improved the performance of semantic segmentation. Various encoder-coder-based methods have been proposed [12][18] [20] [21].

Semantic segmentation can be applied to pedestrian or vehicle region extraction for autonomous driving systems, including the self-driving robot system. One critical problem is the variation of object size. Objects of the same class can have different sizes and appearances depending on the position of the object relative to the camera. To address this problem, we have proposed using multiple dilated convolution blocks to deal with various object sizes. Dilated convolution convolves elements separated by a certain distance in the convolution process. We also propose using multiple skip connections to obtain robustness against appearance changes. The use of multiple skip connections is inspired by those used in the Residual Network (ResNet) [16], which achieves high accuracy in object recognition tasks. One method that has skip connections is U-Net, which connects the correspond layers from the encoder to decoder. Our multiple skip connections consist of 1) a skip connection in the encoder like ResNet, 2) a skip and concatenate output of multiple dilated convolution blocks, and 3) a skip connection like U-Net. The proposed method, which employs multiple dilated convolution blocks and multiple skip connections, is able to extract fine segmentation results of small objects.

II. Related works

Highly accurate network structures have been proposed for object recognition by deep learning through ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet, which is composed of 5 convolution layers and 3 fully connected layers, is a pioneering network [2]. It employs a rectified linear unit (ReLU) as an activation function and Dropout to improve generalization. In addition, the network uses GPUs to learn deep network structures in practical time. VGG16 is a deeper network composed of 13 convolution layers and 3 fully connected layers [8]. In this network, the filter size of each convolution layer is set to 3×3, and the pooling layer is placed after convolution layers. By fixing the small filter size, it is possible to reduce the number of parameters despite having a deeper structure than AlexNet. GoogleNet has 22 layers in which there are 9 inception modules that perform convolution processing with different filter sizes in parallel[6]. In an Inception module, by combining feature maps obtained by convolving filters of 1×1, 3×3, or 5×5, different features of the region of interest can be captured at the same time. ResNet has 152 layers and a very deep network structure [16]. For very deep network structures, there is a risk of error disappearance and explosion and a lack of accuracy improvement. In ResNet, a skip connection that bypasses multiple layers can propagate errors back to near the input layer. Also, at the time of inference, fine information near the input layer can be propagated forward by skip connections. Semantic segmentation learns network that recognizes class labels for each pixel end-to-end. A network used for this task is based on these object recognition network structures. A fully convolutional network (FCN) used a fine-tuned pre-trained VGG16, which learned object recognition using Imagenet [9]. To accommodate different input data sizes, the fully connected layers are replaced with convolution layers which have a 1×1 filter. To capture global information of the entire image and local information of each class, skip connection is adopted. Layers close to the input layer of the network structure capture fine information of the image. However, by repeating the pooling, the feature maps become small in size and lack fine information. The skip connection concatenates output feature maps and intermediate feature maps to obtain fine class labels. Segnet has a encoder–decoder network structure [11]. The encoder has convolution and pooling layers based

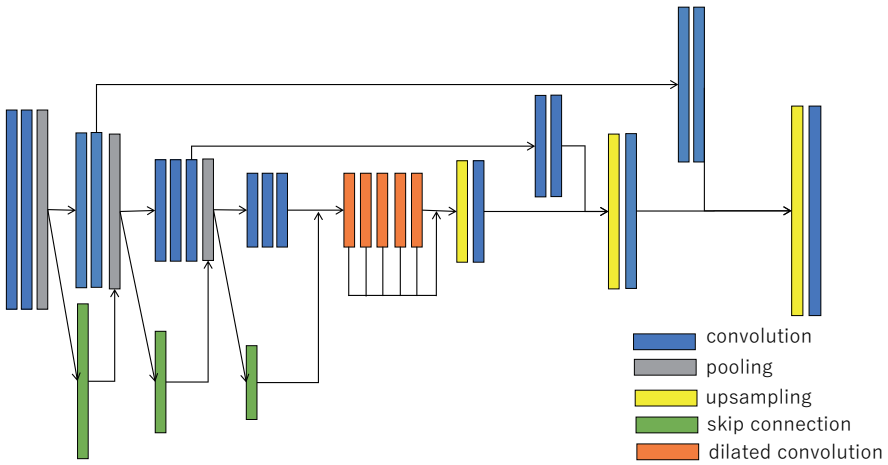

Fig. 1. Network architecture of our semantic segmentation. The colored boxes denote convolution (blue), pooling (gray), upsampling (yellow), dilated convolution (orange), and skip connection (green) layers. The network is based on encoder - decoder structure. The multiple dilated convolution layers are arranged between encoder and decoder. skip connection are introduced at them.

on the VGG16 structure and extracts features from the input image. The decoder has a structure paired with the encoder that consists of deconvolution and upsampling layers. In addition, the pooling indices are recorded, and the feature value is substituted to the recorded position in the upsampling layer, with 0 being substituted for the other positions. Thus, detailed class labels can be recognized from the decoded features. CRF-RNN is a post-process network that inputs the probability map of each class obtained by FCN or SegNet and performs more detailed segmentation[10]. It repeatedly performs error correction of local segmentation considering the probability distribution of each class between neighboring pixels. The CRF-RNN can be learned end-to-end with the network that outputs the probability map of each class. While CRF-RNN focuses on neighboring pixels, dilated convolutions take account of global information [15]. Dilated Convolutions perform the convoluting position sparsely. Consequently, it is possible to perform convolutions over a wide region.

TABLE I. STRUCTURES OF EACH LAYER. THE NETWORK HAS 10 LAYERS IN THE ENCODER, 5 LAYERS IN MULTIPLE DILATED CONVOLUTION BLOCKS AND 3 LAYERS IN THE DECODER.

Layer	Filter size	# of filters	Activation function	Pooling
1	3×3	32	ReLU	–
2	3×3	32	ReLU	max pooling
3	3×3	64	ReLU	–
4	3×3	64	ReLU	max pooling
5	3×3	128	ReLU	–
6	3×3	128	ReLU	–
7	3×3	128	ReLU	max pooling
8	3×3	256	ReLU	–
9	3×3	256	ReLU	–
10	3×3	256	ReLU	–
11	3×3, s=2	512	ReLU	–
12	3×3, s=4	512	ReLU	–
13	3×3, s=8	512	ReLU	–
14	3×3, s=16	512	ReLU	–
15	3×3, s=32	512	ReLU	–
16	3×3	256	ReLU	upsampling
17	3×3	128	ReLU	upsampling
18	3×3	# of classes	ReLU	upsampling

III. PROPOSED METHOD

We propose a network that can segment the details of small objects. The proposed network structure is shown in Figure I. The basis network is an encoder–decoder structure. Multiple skip connections are introduced at the encoder, decoder, and between them to keep local information. To perceive global information, multiple dilated convolution blocks are arranged between the encoder and decoder. The skip connection of the encoder is inspired by that of the ResNet. The skip and input feature maps are input with the feature maps of the following layers into the multiple dilated convolution blocks. These blocks also have skip connections that merge each block ' s feature map. Furthermore, so that the detailed information of the object is propagated to the decoder, the feature maps of the encoder are connected to the layer on the decoder which is paired. Each skip connection performs convolution of 1×1. The multiple dilated convolution blocks consist of multiple dilated convolution layers. In addition, each convolution layer performs batch normalization [14], which reduces variations in data between batches in mini-batch learning. It accelerates the convergence of learning and becomes robust to variations such as brightness.

A. Basis network

As shown in Figure I, the encoder side consists of 10 layers of convolution layers and the decoder side consists of 3 deconvolution layers. Filter sizes and number of filters for each layer are shown in Table I. The filter size in each layer is 3×3, and the number of filters is doubled after 2×2 max pooling. The decoder performs deconvolution for the number of times of pooling. For each deconvolution, the feature map is upsampled by a factor of 2 and convolution is performed. The filter size of the decoder is 3×3. Segnet has the same number of convolution layers as the encoding side at each deconvolution. Unlike the structure of Segnet, our network has only one convolution layer at each deconvolution. ReLU is applied as the activation function of each layer in both the encoder and decoder.

B. Multiple dilated convolution blocks

Multiple dilated convolution blocks that capture the global features are arranged between the encoder and the decoder. Dilated convolution is a process that separates elements to be convoluted with stride, as shown in Figure 2. When 3×3

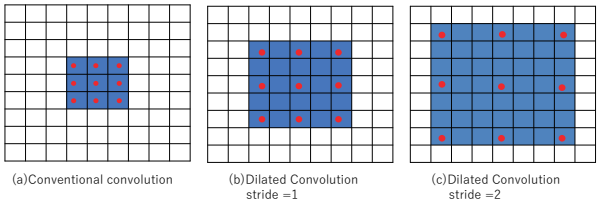

Fig. 2. Conventional convolution and dilated convolution. Filters are applied densely to elements in conventional convolution and applied sparsely with stride s to those in dilated convolution.

filter convolves to input maps, conventional convolution is performed on a dense 3×3 region as shown in Figure 2(a). The input value and the filter value are multiplied for each element to obtain the corresponding value. On the other hand, dilated convolution has stride, which is an interval between convolved elements. When it is set to 1 as shown in Figure 2(b), 3×3 filter is applied to a 5×5 region. When stride is set to 2, the convolution is performed on a 7×7 region as shown in Figure 2(c). The dilated convolution is a sparse connection to a wider region than that of conventional convolution. We stack 5 dilated convolution layers with different strides. Although the dilated convolution has the same filter size as the conventional convolution, it is able to perceive a wider range to capture global context by stacking them,

C. Multiple skip connections

In the ResNet, errors can be propagated close to the input layer even in a deep network structure by introducing skip connections. In FCN and U-Net, high-resolution segmentation results are obtained using the feature map of the middle layer. This idea can also be regarded as a kind of skip connection. In our network, the skip connection of the ResNet is introduced to the encoder. The skip connection of the FCN, which is connected between the encoder and decoder, is also introduced to our network. In the skip connection of the encoder, the value of each element of the feature map is added. When the number of channels of the feature map is inequivalent, the convolution with 1×1 is performed a number of times based on the number of channels of the upper layer to adjust the number of channels. In the case of adding the 32-channel feature map in the 2nd layer to the 64-channel feature map in the 4th layer, 64 of the 1×1 filters are applied to the channel feature maps in the 2nd layer to obtain a feature map of 64 channels. In the skip connection of FCN, the feature map of the encoder is added to the feature map of the decoder for each element. While skip connections such as those in the ResNet concatenate feature maps, it is known that concatenation is equivalent to addition are [19]. By performing a skip connection by adding, it is possible to suppress the amount of memory usage without increasing the number of feature maps. Through these skip connections, detailed information on the object can be propagated through two paths. We also introduced skip connections that merge the feature maps of each dilated convolution layers. It is possible to input feature maps that capture information in various respective regions.

D. Learning of our network

The proposed network is able to learn using the end-to-end approach. Unlike conventional ConvNet-based semantic segmentation methods, it does not use a pre-trained network. This makes it possible to flexibly change the network structure. The mini-batch size is 16. Adam [7] is used for the learning optimization method. A cropped region with a fixed size from a random image is input during the learning process. It is possible to consider various scenes and augment variation of the learning data. The input size is 720×720, which is cropped from 0.75 times to 1.25 times of the size and is resized. Our method is implemented by the Chainer framework and learned with the NVIDIA DGX-1. Because the memory size of the Tesla P100 in the DGX-1 is 16 GB, the mini-batch size that can be processed with one GPU is 2. Therefore, mini-batch learning is performed in data-parallel by using 8 GPUs.

IV. Experiments

We evaluated the proposed network using the Cityscapes dataset [17]. This dataset is composed of images taken in 50 cities in Europe during the day in fine weather and the 30 classes. Some classes frequently occur, therefore, 19 classes were used for evaluation. The Annotation includes fine and coarse annotations. While fine annotations are annotated in detail in 5000 images, the coarse ones are rough annotations that surround the area in 20000 images. In this experiment, we used fine-annotation data. It includes 2975 images for learning, 500 images for validation, and 1525 images for testing. The annotation data for the image used for testing is not published. Testing results can be obtained by upload the result. Therefore, in this experiment, comparison experiments are performed using a validation dataset.

V. Comparison of network structures

TABLE II. Comparison results on validation dataset.

Multiple Dilation Convolution Blocks	Multiple Skip Connections	Class[%] IoU	Class[%] iIOU	Category[%] IoU	Category[%] iIoU
none	none	54.9	37.8	83.6	73.2
use	use	56.1	40.2	84.3	76.1
use no skip	none	67.3	45.8	87.8	74.1
use no skip	use	72.5	52.5	89.2	78.2
use with skip	use	73.0	55.6	89.2	81.9

The evaluation result is shown in Table II. By introducing skip connections in the encoder and between the encoder and decoder, accuracy was improved from a plain encoder–decoder structure from 2% to 3%. By introducing multiple dilated convolution blocks without skip connection, the average class IoU increased from 54.9% to 67.3%, and the average class iIoU improved significantly from 37.8% to 45.8%. When both multiple skip connections and multiple dilated convolution blocks were introduced, the average category IoU was 89.2% and the average category iIoU was 81.9%. These two processes can greatly contribute to accuracy improvement.

VI. Comparison on test dataset

The comparison result is shown in Table III and Figure.3. As a result, the proposed method can obtain better results compared with common segmentation methods, such as Segnet and FCN, and with the method using dilated convolution. When compared with the state-of-the-art methods (SegModel, ResNet - 38) recorded in the benchmark result

Fig. 3. Result images from test dataset of Cityscapes. First and third columns are input images and second and forth columns are our results.

TABLE III. COMPARISON RESULT ON TEST DATASET OF CITYSCAPES

Method	Class[%]		Category[%]	
	IoU	iIoU	IoU	iIoU
SegModel	78.5	56.1	89.8	75.9
ResNet-38 [20]	78.4	59.1	90.9	81.1
Dilation10 [15]	67.1	42.0	86.5	71.1
FCN-8s [9]	65.3	41.7	85.7	70.1
Segnet basic [11]	57.0	32.0	79.1	61.9
proposed method	71.6	49.4	89.3	78.3

of Cityscapes, these methods are better on class IoU and class iIoU. Our method achieves an equivalent accuracy to the category IoU and category iIoU. Since the proposed method can classify at the category level, the multiple dilated convolution blocks and multiple skip connections can improve the accuracy of semantic segmentation. The processing time of our method is 600 ms using a GPU Pascal Titan X. The movie of our semantic segmentation result are published on https://vimeo.com/194006277.

VII. CONCLUSION

We proposed a semantic segmentation method with multiple skip connections and multiple dilated convolution blocks. The skip connections include one in the encoder like that in ResNet one between the encoder and decoder like FCN, and these connections merge the feature maps of each dilated convolution layer. We achieved segmentation with ahigher performance than FCN and Segnet, which are common segmentation methods, on the Cityscapes dataset. Moreover, compared to the state-of-the-art methods, our network achieved an equivalent accuracy on average category IoU and average category. However, if a class with a low frequency of occurrence appears large, erroneous segmentation may be performed.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition" , Proceedings of the IEEE, 1998.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems (NIPS2012), 2012.

[3] C. Farabet, C. Couprie, L. Najman, Y. LeCun, "Learning hierarchical features for scene labeling", IEEE transactions on pattern analysis and machine intelligence (PAMI), 2013.

[4] M. Everingham, A. S. M. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes challenge: A retrospective", International Journal of Computer Vision (IJCV), 2014.

[5] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", IEEE conference on computer vision and pattern recognition (CVPR2014), 2014.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2014), 2014.

[7] D. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

[8] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", International Conference on Learning Representation (ICLR2015), 2015.

[9] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), 2015.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du,P. H. Torr, "Conditional random fields as recurrent neural networks", IEEE International Conference on Computer Vision (CVPR2015), 2015.

[11] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", arXiv preprint arXiv:1511.00561, 2015.

[12] A. Kendall, V. Badrinarayanan, R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding", arXiv preprint arXiv:1511.02680, 2015.

[13] H. Noh, S. Hong, B.Han, "Learning deconvolution network for semantic segmentation", IEEE International Conference on Computer Vision (ICCV2015), 2015.

[14] S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", arXiv preprint arXiv:1502.03167, 2015.

[15] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions", International Conference on Learning Representation (ICLR2016), 2016.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele, "The cityscapes dataset for semantic urban scene understanding", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[18] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", arXiv preprint arXiv:1606.00915, 2016.

[19] P. O. Pinheiro, T. Y. Lin, R. Collobert, P. Dollar, "Learning to refine object segments", European Conference on Computer Vision (ECCV), 2016.

[20] Z. Wu, C. Shen, A. van den Hengel, "Wider or Deeper: Revisiting the ResNet Model for Visual Recognition", arXiv preprint arXiv:1611.10080, 2016.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid Scene Parsing Network", arXiv preprint arXiv:1612.01105, 2016.