Multimed Tools Appl (2018) 77:18689–18707
https://doi.org/10.1007/s11042-018-5653-x

Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions

Duc My Vo[1] **· Sang-Woong Lee**[1]

Received: 5 August 2017 / Revised: 18 December 2017 / Accepted: 14 January 2018 /
Published online: 22 February 2018

Abstract In this work, we investigate the effects of the cascade architecture of dilated convolutions and the deep network architecture of multi-resolution input images on the accuracy of semantic segmentation. We show that a cascade of dilated convolutions is not only able to efficiently capture larger context without increasing computational costs, but can also improve the localization performance. In addition, the deep network architecture for multi-resolution input images increases the accuracy of semantic segmentation by aggregating multi-scale contextual information. Furthermore, our fully convolutional neural network is coupled with a model of fully connected conditional random fields to further remove isolated false positives and improve the prediction along object boundaries. We present several experiments on two challenging image segmentation datasets, showing substantial improvements over strong baselines.

Keywords Semantic image segmentation · Fully convolutional neural networks · Fully connected conditional random fields · Multi-scale dilated convolutions

1 Introduction

The field of semantic segmentation has recently made remarkable contributions to the development of scene understanding and object recognition. Before the development of deep learning, most semantic segmentation algorithms relied heavily on different versions of the conditional random fields method [15, 26, 34], which is used to label pixels in an image with one of several predetermined object classes. As a result, the conditional random fields-based methods can recognize and segment a variety of different objects simultaneously.

✉ Sang-Woong Lee
slee@gachon.ac.kr

[1] Pattern Recognition and Machine Learning Lab, Gachon University, 1342 Seongnamdaero, Sujeonggu, Seongnam 13120, Korea

Most algorithms using conditional random field deal with the semantic segmentation problem by maximizing label agreement between neighboring pixels and developing a model of context information to classify different object classes. In general, a typical model of conditional random fields is computed by unary potentials on each pixel and pair-wise potentials on neighboring pixels. Jakob Verbeek [34] demonstrated that the conditional random fields model plays a key role in significantly improving the performance of pixel-level segmentation methods.

Over the last few years, deep convolutional neural networks [7, 16, 17, 24, 28, 33] have resulted in dramatic developments in the field of object detection and image recognition owing to the fact that they are able to automatically generate meaningful and rich hierarchies of features. Some researchers have successfully applied convolutional neural networks to semantic segmentation [6, 14, 32, 36] in order to recognize and understand the content of an image at the pixel level. Among the methods of deep convolutional neural networks, fully convolutional neural networks [21] have represented the dominant research direction for improving semantic segmentation, because of their computational efficiency for dense prediction. Many recent methods have been developed from fully convolutional neural networks including DeepLab [2], Boxsup [4], deep parsing [20], deconvolution [23], and recurrent neural networks using conditional random fields [36]. Among these methods, fully connected conditional random fields method is one of the key components that can make segmentation performance more successful, because it is able to obtain sharper object boundaries.

Current approaches to image classification also include multi-scale deep features selected from different layers of pooling and subsampling in a deep convolutional neural network, where the receptive field in an original image can be expanded to better cover global features [11, 22]. Unfortunately, these methods lead to a reduction in the resolution and a loss of detail and local features in an image. To avoid the consequences of losing resolution and rescaling images, Noal [23] and Ronneberger [27] employed up-convolution layers, which are useful for recovering the information lost through down-sampling processes at pooling layers in a convolution network. Apparently, this technique is only able to recover part of lost information. Therefore, the accuracy of this image recognition technique remains limited. To address this limitation, Chen [2] adopted dilated convolutions to extract denser feature maps without using downsampling operations at the last several layers of a pre-trained network.

The challenge of designing multi-scale context information without losing resolution motivated a new approach [6, 19] using a pyramid of different rescaled versions of an original image as input to an improved convolution neural network. These algorithms require extremely high computational costs, because of a huge amount of input parameters. Furthermore, combining deep features from different scaled images remains challenging. However, this approach inspired us to find a better solution for combining the advantages of multi-resolution images and multi-scale feature descriptors to extract both global and local information in an image without losing resolution.

In this work, we aim to design an efficient architecture for pixel-wise semantic segmentation by investigating the effects of a cascade architecture of dilated convolutions and a deep network architecture of multi-resolution image inputs on the accuracy of semantic segmentation. First, the cascade architecture of dilated convolutions is used at the end of our network to extract multi-scale features in local regions without increasing the number of training parameters. Second, because the same object might have different sizes from different images, we apply multi-scale input images to the same deep convolution network for searching multi-scale features in multi-scale image inputs. However, both techniques,

the cascade architecture of dilated convolutions and the deep network architecture of multi-resolution image inputs, have some disadvantages. On the one hand, enlarging receptive field size using dilated convolutions with big rates r results in a high possibility of losing context information at the locations where the dilated convolution introduces zeros between consecutive filter values. On the other hand, feeding multi-scale images in a single deep convolutional neural network are normally very expensive in a training process. These are the reasons why we combine the strengths of multi-scale image input and large field-of-view features for object semantic segmentation, and we can reduce their disadvantages. In addition, to sharpen and smooth object boundaries, we build a fully connected conditional random fields model and integrate it into the output of our network. Finally, we use the maxout layer as a strategy of searching the best features to fuse into the final score map.

We tested our proposed algorithms and its competitors to evaluate the accuracy of object semantic segmentation on challenging datasets. Based on extensive experiments, our algorithms are shown to significantly outperform state-of-the-art algorithms. In particular, our contributions are summarized as follows:

- We combined the idea of a cascade architecture of dilated convolution with the idea of a deep network architecture of multi-resolution image inputs, so that our network can extract multi-scale features, recover the spatial resolution, and restrict the increase of network parameters.
- We employed a fully connected conditional random fields model that is integrated into the output of our network to further remove isolated false positives, and improve the prediction along object boundaries.
- We used a maxout layer as a strategy for determining competitive and dominant features to fuse into the final score map. Thus, our network can be trained more efficiently.

The remainder of this paper is organized as follows. In Section 2, we briefly review some related state-of-the-art algorithms for semantic segmentation, which motivated our research. Section 3 will describe our method of fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions. In Section 4, the experimental results obtained from the challenging PASCAL VOC 2012 database, and the challenging dataset of Skin Lesion Analysis Toward Melanoma Detection, are presented. We conclude this paper, mentioning our intentions for our future work, in Section 5.

2 Related work

In the last decade, researchers have adopted hand-crafted features and some traditional classification methods such as random decision forests [30], Markov random fields [35], and conditional random fields [12, 29] to address challenging problems of semantic segmentation. It has been attempted to utilize context information to enhance segmentation performance. Although these methods successfully increase the efficiency of object segmentation in an image, the accuracy of these methods is strongly dependent on the quality of hand-crafted features, which normally does not generalize well. To remove this barrier, researchers have replaced handcrafted features by automatically learning informative features, especially deep learning features. Recently, semantic segmentation has mainly been developed using the theories of deep convolutional neural networks, conditional random fields, multi-scale features.

Most state-of-the-art methods for semantic image segmentation using deep learning [2, 21] basically constitute improvements to fully convolutional neural networks, which are based on the idea of adding convolutional layers at the end of networks instead of using any fully-connected layers. Long et al. [21] used convolutional layers for the end-to-end training of a state-of-the-art model for semantic segmentation. Unlike traditional deep networks with fully-connected layers, which cannot process image inputs of different sizes, fully convolutional neural networks can handle image inputs of any particular size owing to the fact that they only have convolutional layers that obey translation invariance, and their output is only dependent on the local area of input. Chen et al. [2] then improved the fully convolutional neural network by combining its output response with a fully connected conditional random field. In this approach, the fully convolutional neural network provides unary terms, and the pixels in the input image are treated as nodes for a local pairwise conditional random field. Conditional random field inference is then applied to directly minimize an energy function that employs two sets of potential functions, unary potentials, and pairwise potentials. The method of pixel-level conditional random fields is also used to generate a set of segmentation proposals, which are presented in detail in [3]. These segmentation proposals can be ranked again using a model of a deep convolutional neural network. Zheng et al. [36] even integrated a model of fully connected conditional random fields into a model of convolutional neural network to construct an advanced network in which these two models can be trained together end-to-end using the usual back-propagation algorithm. Zheng also aimed to combine the strengths of convolutional neural networks and the conditional random fields model. Similar to conditional random fields models, Liu et al. [20] has demonstrated that Markov random field models are also able to enrich context information for semantic segmentation tasks. That author also adopted a convolutional neural network to model unary terms and approximate the mean field algorithm for pairwise terms. This approach is able to achieve a high performance by training the Markov random field model and the convolutional neural network model together end-to-end.

One of the disadvantages of fully convolutional neural networks is the low-resolution output responses. To obtain higher-resolution predictions, Noh et al. [23] trained deconvolution layers to upsample the low resolution predictions. However, Lin et al. [19] proved that a network that can combine low-resolution and high-resolution predictions results in a better performance. He proposed a network architecture that merges multi-scale image inputs into the same feature map for semantic segmentation, to capture multi-scale information from background regions and increase the field-of-view for the network. In this network, features from the small scale image inputs provide the long-range context information, while the large scale image inputs encode detailed information of objects. Once again, pairwise potential functions are extracted by a convolutional neural network with a multi-scale image input, to optimize the conditional random fields model. Thus, object boundaries are effectively sharpened and smoothed. Farabet et al. [6] also combined multiple image resolutions in a deep convolutional neural network, and improved the performance by using a segmentation tree. Chen et al. [19] introduced an alternative method of effectively enlarging the field-of-view for feature maps without increasing the number of training parameters. Instead of using multi-scale inputs to capture objects at multiple scales, he used dilated convolution layers plugged into the end of a deep convolutional neural network. Dilated convolutional layers are applied for an input representation by dilating the filter before computing the usual convolution. These dilated convolution layers significantly increase the spatial resolution of the final feature maps, without increasing the number of training parameters and the computational cost of the network. Mostajabi et al. [22] enriched feature representations by extracting multi-scale local features from each superpixel in an image. He developed a

purely feed-forward architecture to exploit statistical structures in images and avoid complex and expensive inferences. For this reason, his approach achieved a high performance for semantic segmentation.

Another disadvantage of fully convolutional neural networks is that they generally have high computational costs. To alleviate this problem, Badrinarayanan [1] employed a convolutional encoder-decoder network that can remove unnecessary layers and reduce the number of training parameters. Paszke [25] construct the ENet network that employ a bottleneck module to reduce convolution computation. This method can make semantic segmentation run fast, but the accuracy significantly drops.

3 Proposed approach

3.1 Dilated convolutions

In conventional deep neural networks, max-pooling layers and stride operators are repeatedly used to down-sample an input representation, and to reduce its dimensionality. These techniques are also helpful for decreasing the computational cost by reducing the number of training parameters. However, they also lead to a significant reduction in the spatial resolution, as shown in Fig. 1. Dilated convolutional layers have recently been adopted to recover the spatial resolution without increasing the number of training parameters.

Unlike max-pooling layers and stride operators, a two-dimensional dilated convolution layer, also called a "convolution with dilated filter" layer, is applied for an input representation by dilating the filter before computing the usual convolution. The size of the filter is expanded, and the empty positions are filled completely with zeros. As a result, the weights are matched to distant elements in the input matrix. The distance is determined by the rate r. If the kernel center is aligned to an arbitrary location in an image, then the kernel elements are matched to input elements as shown in Fig. 2.

For a simple example of dilated convolutions, we can apply one-dimensional dilated convolutions to one-dimensional signals. The output $y[i]$ of the dilated convolution of an input $x[i]$ with a filter $w[k]$ is computed as follows:

$$y[i] = \sum_{k=1}^{m} x[i + r \cdot k]w[k] \tag{1}$$

where m is the length of the filter $w[k]$.

The main advantage of dilated convolutions is to expand the receptive field of filters at convolution layers, while the resolution of the input matrix is not reduced. By applying

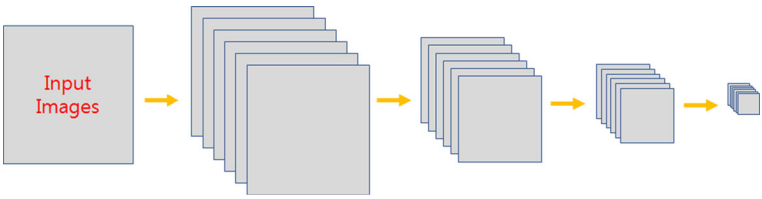

Fig. 1 Deep convolutional neural network without dilated convolutions

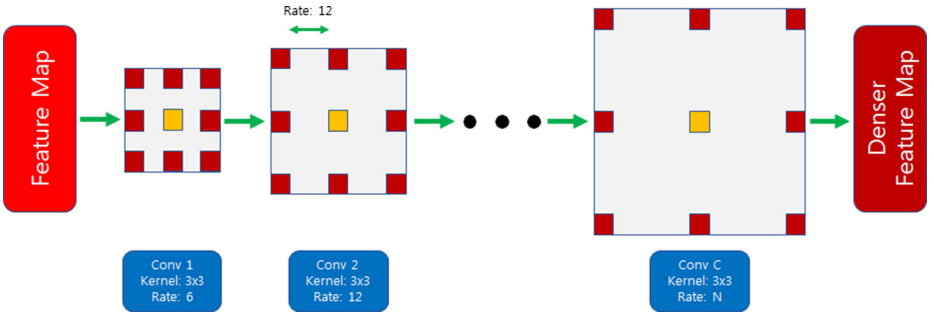

Fig. 2 Dense feature extraction with a cascade of dilated convolutions with different rates

dilated convolutions with a rate r, a filter with the kernel size k×k can be expanded up to $k' = k + (k-1)(r-1)$. This expansion offers some advantages, but also results in some disadvantages. On the one hand, we can apply dilated convolutions with large rates r to capture larger contexts without increasing computational costs. However, dilated convolutions with large rates r also introduce more zeros between filter values, and lose more local context information in smaller regions. On the other hand, dilated convolutions with small rates r can be used to improve the localization performance. Nevertheless, they also produce feature maps with narrow receptive fields. Therefore, a combination of dilated convolutions with different rates r is necessary to extract denser feature maps. Hence, we propose developing a cascade architecture of multi-scale dilated convolution layers for extracting contextual information at multiple scales. This cascade architecture is comprised of consecutive convolutional layers, each of which uses only dilated convolution kernels with the same rate r to produce dense feature maps with the same receptive field. The output matrix of the previous dilated convolution layer is connected to the input of the current dilated convolution layer. In addition, the previous layer uses dilated convolutions with a smaller rate than the current layer, so that it can extract local features, and improve the localization accuracy. In contrast, the current layer uses dilated convolutions with a larger rate, in order to increase the context assimilation. Therefore, the feature maps of the current dilated convolution layer are aggregated, and become denser than those of the previous one. The cascade of dilated convolutions can be added into a deep convolution neural network to compute the final feature map at a high resolution, as shown in Fig. 3.

3.2 Fully deep convolutional neural networks with multi-scale image input

Context information can be effectively captured by combining features from dilated convolutions with different rates r. However, enlarging receptive field sizes using large rates r

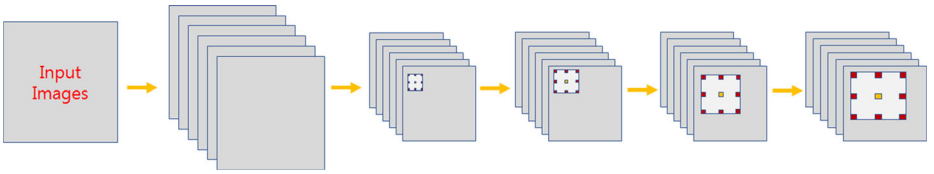

Fig. 3 Alternative architecture: a deep convolutional neural network with dilated convolutions

results in a high possibility of losing context information at the locations where the dilated convolution introduces zeros between consecutive filter values. The larger the rate r is, the more zeros are added.

Recently, multi-scale image inputs have been widely explored for capturing features, and have exhibited good performance in some recent segmentation methods. Because the same object may have different sizes in different images, searching multi-scale features in a multi-scale image input is reasonable. Hence, we can apply multi-scale input images to the same deep convolution network, where each scale is passed through one branch of this network, and the features are fused from all scales. Nevertheless, feeding multi-scale images into a single deep convolutional neural network is normally very expensive in a training process, because a huge number of training parameters are required to build such a network. Therefore, combining the strengths of multi-scale image inputs and large field-of-view features for segmentation has been our focus in this study. We developed a deep convolutional neural network using multi-scale images as input matrices, and integrating a cascade of dilated convolutions at the end of each sub-network, as illustrated in Fig. 4.

In this network, we build a model of a deep convolutional neural network in which a feature map is trained for each scale. In particular, we resize the input image to several scales by bilinear interpolation. Each scaled image is passed through one branch of this network. The output of each branch is a feature map that is then rescaled to have the same resolution with other feature maps. Finally, all feature maps are fused into a shared feature map. Because each object in an image is prominent in different feature maps, we adopt a maxout layer to obtain competitive and dominant features from all feature maps, and fuse these into the shared feature map. Generally speaking, a maxout layer constitutes an improvement of the maxout network [8]. Unlike the maxout network, the maxout layer is considered as the layer of maximal feature maps, as shown in Fig. 5. In particular, the convolution layer includes groups of feature maps, and feature values at the same coordinates from these groups are compared to select the maximal value, which is then assigned to the feature value at the same coordinates in the maxout layer.

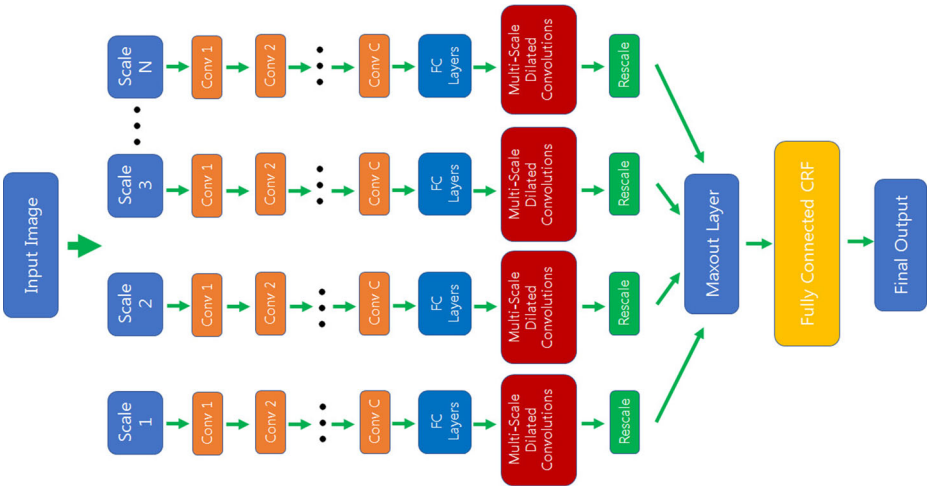

Fig. 4 Alternative architecture: a deep convolutional neural network with multi-scale images and multi-scale dilated convolutions

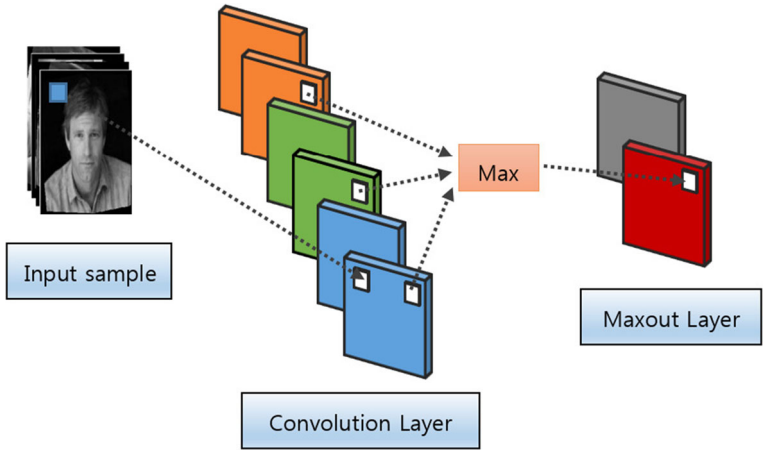

Fig. 5 An illustration of a maxout layer. The convolution layer includes groups of feature maps, and feature values at the same coordinates from these groups are compared to select the maximal value, which is assigned to the feature value at the same coordinates in the maxout layer

3.3 Fully connected conditional random fields

Although deep convolutional neural networks can give us a reliable final score map in which objects are identified, and their locations are roughly located, their boundaries are not extracted accurately and sharply. This problem stems from the fact that these networks must complete the two challenging tasks of object identification and pixel-level object localization, which have a trade-off in accuracy. We apply fully connected conditional random fields to address the challenge of pixel-level object localization, and recover object boundaries. The fully connected conditional random fields model is integrated into the output of our network to improve the segmentation performance, and fix object boundaries. In particular, this model will minimize an energy function that employs two sets of potential functions, unary potentials, and pairwise potentials, as presented below:

$$E(x) = \sum_i \alpha_i(x_i) + \sum_i \alpha_{ij}(x_i, x_j) \tag{2}$$

where $\alpha_i(x_i)$ is the unary potential measuring the inverse likelihood of the label x_i at the pixel i, and $\alpha_{ij}(x_i, x_j)$ is the pairwise potential estimating the cost of label assignments at the pixels i and j with the labels x_i and x_j, respectively. The unary potential is normally the output of a pixel-wise classifier, computed as follows:

$$\alpha_i(x_i) = -logP(x_i) \tag{3}$$

where $P(x_i)$ is the probability of assigning the label x_i to the pixel i. In our method, $P(x_i)$ is computed by our deep convolutional neural network, and the pairwise potential is computed based on the image gradients between the pixel and its neighbors. In particular, a pixel and its neighbor are classified into the same label if the computed gradient between them

is small. Thus, the pairwise potential encourages consistency in the appearances of the segmented objects, and improves object delineation. The pairwise potential is computed by the following formula:

$$\alpha_{ij}(x_i, x_j) = \delta(x_i, x_j) \sum_{m=1}^{K} \omega_m f^m(\tau_i, \tau_j) \tag{4}$$

where f^m is a Gaussian kernel weighed by the parameter ω_m. The kernel f^m is computed based on the features τ_i and τ_j collected for the pixels i and j, respectively. The features τ_i and τ_j represent pixel color intensities denoted as I and pixel positions denoted as p. Hence, the formula for the kernel f^m is

$$\omega_1 exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_1^2} - \frac{\|I_i - I_j\|^2}{2\sigma_2^2}\right) + \omega_2 exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_3^2}\right) \tag{5}$$

Finally, the energy $E(x)$ is minimized to find the best label assignment for the input image. However, this minimization problem is originally an intractable problem. Thus, to efficiently approximate the probabilistic inference, we apply a mean-field approximation to the distribution of conditional random fields. The distribution of conditional random fields $P(x)$ is approximated by a distribution $Q(x)$, which can be expressed by a product of independent distributions $Q(x) = \prod Q_i(x_i)$, as presented in detail in [15].

3.4 Deep residual networks

Network depth is very important in improving the accuracy of neural networks. However, training a deeper network is a difficult challenge. A deep residual network [13] is a deep convolutional neural network that can be trained at a consistently deeper level than a conventional deep neural network, because it adopts a residual learning framework that makes the training process easier, and achieves a better performance by increasing the depth of the network. Thus, we are even able to train a deep residual network with 100 or 1000 layers. Because deep residual networks have reached the state-of-the-art performance in image classification, we aim to learn a ResNet-based model with multi-scale input images, where we can pass each scaled image through a deep residual network. We employ ResNet-101 with five blocks and more than 100 layers for each scaled image. Such a network consists of a huge number of parameters, and its training process easily reaches the maximum RAM capacity of our GPU device. The computational cost for an increase in accuracy is extremely high. For this reason, we only employ two scaled image inputs, with scales of 1.0 and 0.75.

Each deep residual network can summarize the features of a scaled image input in the feature map taken from the output of the last block. This feature map is then passed through a cascade of dilated convolutional layers to capture multi-scale context information, and generate a score map. We apply a method of simple bilinear interpolation to increase the resolutions of the score maps, so that these maps have the same resolution. Then, the max-out layer plays an important role in merging these score maps into the final score map which can retain competitive features. Finally, the fully connected conditional random fields model is integrated into the output of the network to encode pixel-level pairwise similarities, and sharpen object boundaries. The entire architecture of the ResNet-101-based network is illustrated in Fig. 6.

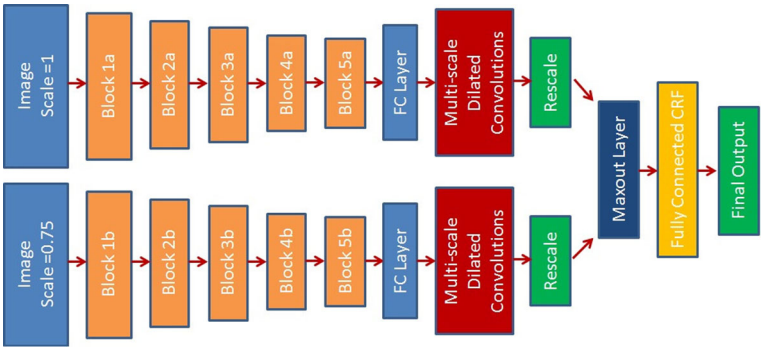

Fig. 6 The architecture of our ResNet-101-based network

3.5 Very deep convolutional neural networks

Similar to deep residual network, very deep convolutional neural networks [31] explore the advantages of network depth on the accuracy in image recognition tasks. In fact, a very deep convolutional neural network only employs a very small number of convolution filters and 19 weighted layers. Therefore, such a network requires a much smaller number of training parameters than a deep residual network. However, this network still achieves state-of-the-art results. Thus, we aim to utilize very deep convolutional neural networks to extract more multi-scale features. We build a deep model with three scaled image inputs with scales of 1.0, 0.75, and 0.5. We employ the VGG-16-based model which has 16 weighted layers, for each scaled image. The entire architecture of the VGG-16-based network is illustrated in Fig. 7.

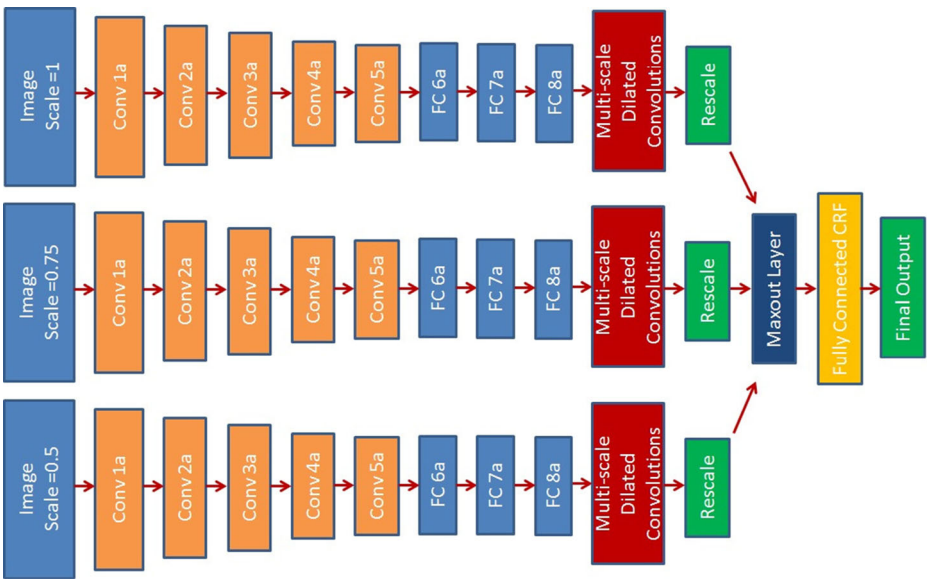

Fig. 7 The architecture of our VGG-16-based network

4 Experimental results and analysis

4.1 Dataset

We used the challenging PASCAL VOC 2012 database [5] to evaluate the accuracy of our proposed methods and the state-of-the-art algorithms which are the fully convolutional neural network (FCN-8s) [21], the DeepLab network [2], the BoxSup network [4], the Zoom-out network [22], the CRF-RNN network [36], the DPN network [20], and the DeconvNet network [23]. First, we pretrained our deep models on the MS-COCO dataset [18]. We then used the challenging PASCAL VOC 2012 database to train 20 classes, including person, bottle, car, train, and one background class. Because the original training dataset only consists of 1464 training images, which are not sufficient for training our deep network, we used its extended training dataset augmented by the extra annotations provided by Hariharan et al. [10]. In total, we employed 10,582 augmented images in the training process. We can easily conduct experiments for comparison of semantic segmentation methods on our computer, which is a PC with 3.6 GHz Intel Core i7 CPU and GeForce GTX 1070 GPU.

To evaluate the accuracy of our proposed methods, we also used the challenge dataset of Skin Lesion Analysis Toward Melanoma Detection [9], which is the largest collection of quality skin lesions images. Recognizing melanoma in dermoscopy images is a very challenging problem, because we often have to deal with the low contrast of skin lesions. We aim to prove that our network can improve the diagnostic performance of melanoma. This dataset consists of 900 training images and 350 testing images. Because all training and testing images are high resolution images with size 1024×768, we randomly crop sub-images with the same size on each training image to increase the training samples. In the testing process, the whole image is segmented entirely by combining the prediction results of overlapped sub-images spreading in the whole image.

The performance of our method was evaluated by applying pixel intersection-over-union (IOU) scores to a set of ground-truth and predicted bounding boxes. The IOU measurement, also called the Jaccard index, is a method for measuring the accuracy of an object segmentation algorithm. If an object segmentation algorithm can provide predicted bounding boxes, we can use the IOU measurement for evaluating its accuracy. The IoU measurement is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(P, G) = \frac{P \cap G}{P \cup G} \tag{6}$$

where $P \cap G$ is the area of overlap between the predicted bounding box P and the ground-truth bounding box G, and $P \cup G$ is the area covered by both the predicted bounding box P and the ground-truth bounding box G.

4.2 Training

We aimed to learn a ResNet-based model by adopting the ResNet-101 deep residual network, which is one of state-of-the-art classification networks, and has been pre-trained for classification tasks in [13]. The ResNet-101 deep residual networks were applied for each branch in our network architecture, as mentioned in the previous section. Each branched

network was modified by replacing the last fully-connected layers of the original ResNet-101 network with a cascade of dilated convolutions as mentioned above, so that the network becomes fully convolutional. Each cascade consists of dilated convolution layers with corresponding rates r of 6, 12, 18, and 24. The layer of fully connected conditional random fields was then decoupled with the network in the training stages. We fine-tuned the network by changing the number of training object classes at the last layers, and applying the loss function, which is the sum of cross-entropy terms for each spatial position in the final dense feature map. We used an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.0005. The learning rate is multiplied by 0.1 after 20,000 iterations.

Our second network was trained by adopting the VGG-16 convolutional neural network, which is pre-trained in ImageNet [31]. The architecture of this network is similar to the architecture of the above ResNet-based model, except for the number of multi-scale inputs. This network adopts three scales s of 1, 0.75, and 0.5 instead of the two scales in the ResNet-based model. For training this model we adopted stochastic gradient descent (SGD) to minimize the loss function, which is the sum of cross-entropy terms for each spatial position. The initial learning rate was set to 0.001 and the mini-batch size was 20 images with an initial learning rate of 0.001. The learning rate is multiplied by 0.1 after 2000 iterations. We set the momentum to 0.9 and weight decay to 0.0005.

4.3 Evaluation

To evaluate the semantic segmentation accuracy, the PASCAL dataset is employed for comparison. The results for IoU scores are shown in Table 1. Our ResNet-based model achieves an IoU score of 78.5, which is the best result among all methods, while the VGG-16-based model achieves an IoU score of 74.8, which still outperforms the competing classifiers, except for the DPN and BoxSup networks. This can be explained by the fact that multi-scale feature extraction plays an important role in recognizing objects in different contexts and scales. Thus, our network using multi-scale image inputs and multi-scale dilated convolutions, significantly outperforms the competitors that only employ a single scale input.

Table 1 Comparison of our proposed methods with other state-of-the-art methods on the PASCAL VOC 2012 dataset

Method	mean IOU (%)
FCN-8s	62.2
Zoom-out	69.6
DeepLab	72.7
CRF-RNN	74.7
DPN	77.5
DeconvNet	74.1
BoxSup	75.2
VGG-16-based Net	74.8
ResNet-101-based Net	**78.5**

Table 2 Processing time of proposed networks

Method	Run time (ms)
VGG-16-based Net	146
ResNet-101-based Net	475

Table 3 Performance of our ResNet-101-based network on the PASCAL VOC 2012 dataset

Method	mean IOU (%)
VGG-16+Dilated	74.0
VGG-16+Dilated+CRFs	74.8
ResNet-101+Dilated	77.8
ResNet-101+Dilated+CRFs	78.5

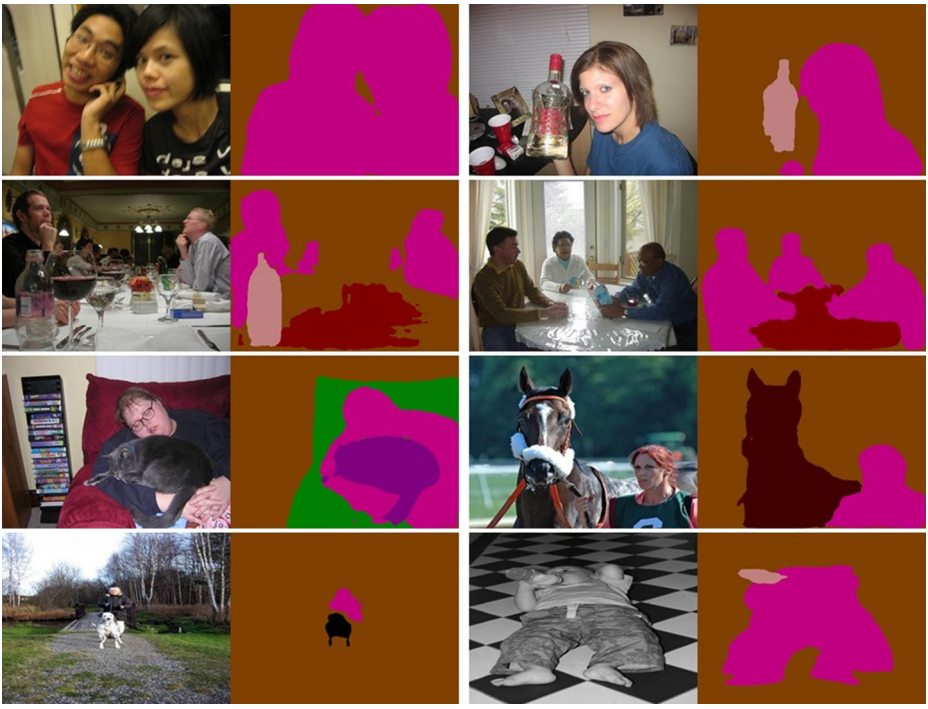

Fig. 8 Qualitative results for the proposed ResNet-based network on the Pascal VOC 2012 dataset

We have also seen that the ResNet-based model achieves a better segmentation performance than the VGG-based model. However, the run time of the VGG-based network is much faster than that of the ResNet-based network, as shown in Table 2, because it consists of far fewer network parameters. Our VGG-based network can be developed for real-time applications in future work.

As shown in Table 3, the layer of fully connected conditional random fields effectively improves the performance of our network. In particular, the component of fully connected conditional random fields introduces an extra 0.7% improvement to the ResNet-based model and an extra 0.8% improvement to the VGG-based model. These results clearly indicate the benefits of the fully connected conditional random fields model in significantly addressing the challenge of pixel-level object localization, and recovering object boundaries.

Figure 8 presents example segmentations of people and objects, which emphasize activities involving human-object interactions and human-human interactions. Typical examples of corrected segmentations of animals and vehicles are shown in Figs. 9 and 10, respectively. All of the example images were collected from the PASCAL VOC 2012 dataset.

We then evaluated the semantic segmentation accuracy of our methods and their competitors on the challenge dataset of Skin Lesion Analysis Toward Melanoma Detection. Table 4 shows the semantic segmentation results of our proposed networks. We have seen that that our ResNet-based model achieves much better results than the VGG-based model.

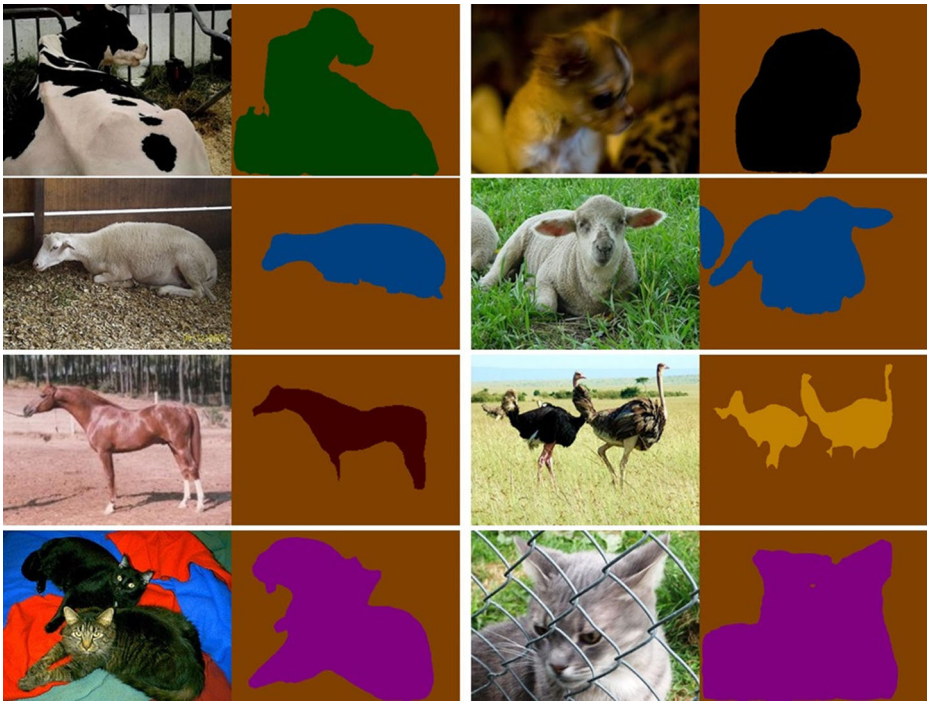

Fig. 9 Qualitative results for the proposed ResNet-based network on the Pascal VOC 2012 subset of animals

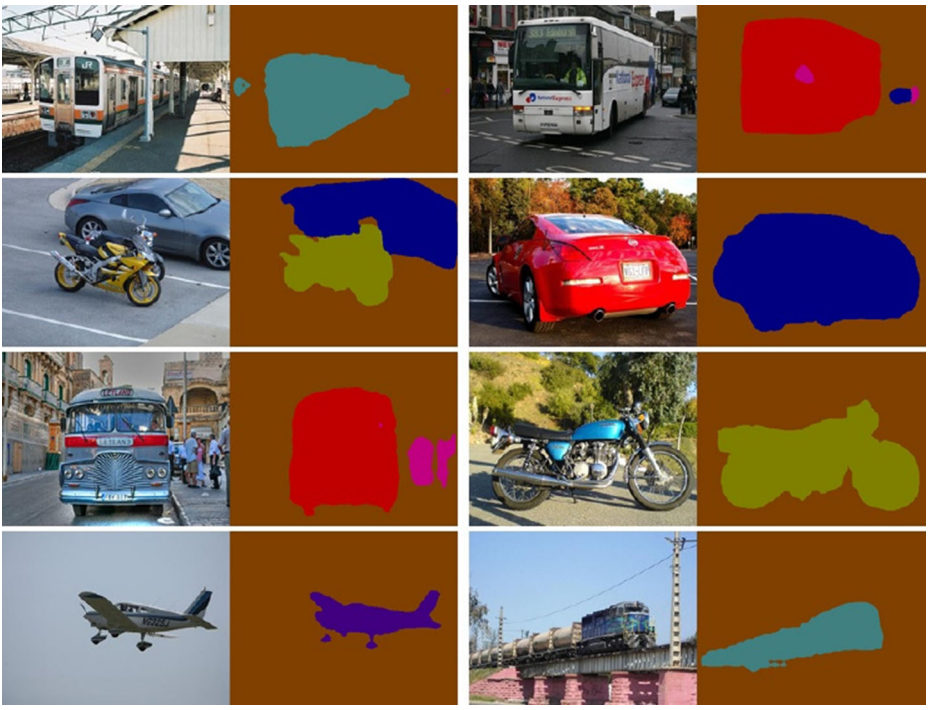

Fig. 10 Qualitative results for the proposed ResNet-based network on the Pascal VOC 2012 subset of vehicles

Our ResNet-based model achieves an IoU score of 83.5, which is the best result among all the methods. This also means that our networks are effective in detecting melanomas which often change in size and shape. This is because our networks can exploit multi-scale features from multi-scale image inputs and multi-scale dilated convolutions. In addition, the component of fully connected conditional random fields introduces an extra 0.4% improvement to the ResNet-based model. Figure 11 presents example segmentations of melanomas. These examples show the challenging tasks of melanoma recognition, including the complicated variation of melanomas, the existence of artifacts. However, our ResNet-based model still efficiently obtains a high segmentation performance.

Table 4 Performance of our proposed methods on the Skin Lesion Analysis Toward Melanoma Detection dataset

Method	mean IOU (%)
DeepLab	79.8
VGG-16+Dilated	81.3
VGG-16+Dilated+CRFs	81.9
ResNet-101+Dilated	83.1
ResNet-101+Dilated+CRFs	83.5

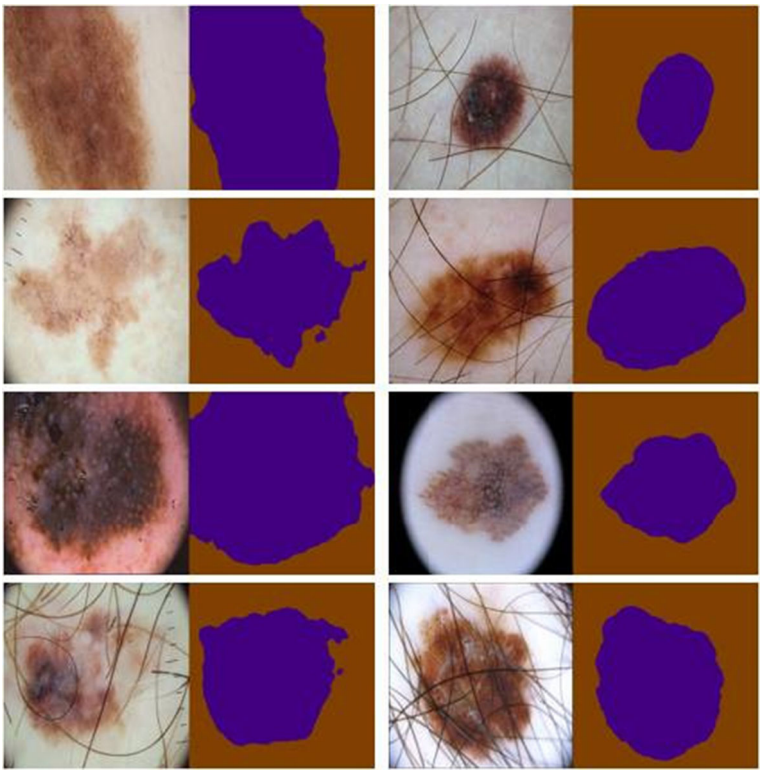

Fig. 11 Qualitative results for the proposed ResNet-based network on the Skin Lesion Analysis Toward Melanoma Detection dataset.

5 Conclusion

We combined the ideas of a cascade of dilated convolutions and a deep convolutional neural network using multi-scale input images to construct a novel method that can predict objects accurately, and produce detailed semantic segmentation maps. Furthermore, we employed a fully connected conditional random fields model, which is integrated into the output of our network to further remove isolated false positives, and improve predictions along object boundaries. Our experimental results show that the proposed method is consistently superior to other state-of-the-art methods for semantic segmentation. This is because our deep convolutional neural network model is not only able to efficiently capture larger contexts without increasing computational costs, but can also improve the localization performance. We also demonstrated that our deep network can achieve a high performance when dealing with medical image segmentation tasks.

For future developments, we intend to improve the performance of our deep convolutional neural network model by further exploring the key role of probabilistic graphical models in advancing the performance of convolutional neural networks. We also intend to develop our method in the field of video-based semantic segmentation. Thus, we will focus on constructing a real time semantic segmentation system based on a cascade convolutional neural network. This network outputs the cascade feature fusion to quickly achieve a high semantic segmentation performance. Unlike our ResNet-based network, the idea of

the cascade convolutional neural network is to build a coarse prediction map by passing the low-resolution image through one branch of the network, and improve this prediction map gradually by passing higher resolution images though the next branches of this network.

Acknowledgements This work was supported by the GRRC program of Gyeonggi province [GRRC-Gachon2017(B01), Analysis of behavior based on senior life log].

References

1. Badrinarayanan V, Kendall A, Cipolla R (2015) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561
2. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2015) Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR
3. Cogswell M, Lin X, Purushwalkam S, Batra D (2014) Combining the best of graphical models and ConvNets for semantic segmentation. In: Arxiv preprint arXiv:1412.4313
4. Dai J, He K, Sun J (2015) Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV
5. Everingham M, Eslami SMA, Gool LV, Williams CKI, Winn J, Zisserma A (2014) The pascal visual object classes challenge a retrospective. In: IJCV
6. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE PAMI 35(8):1915–1929
7. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR
8. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y (2013) Maxout networks. arXiv:1302.4389
9. Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A (2016) Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv:1605.01397
10. Hariharan B, Arbelaez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: International conference on computer vision (ICCV)
11. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hyper-columns for object segmentation and fine-grained localization. In: CVPR
12. He X, Zemel R, Carreira-Perpindn M (2004) Multiscale conditional random fields for image labeling. In: CVPR 2004, vol 2, pp II–695–II–702
13. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv:1512.03385
14. Hft N, Schulz H, Behnke S (2014) Fast semantic segmentation of rgb-d scenes with gpu-accelerated deep neural networks. In: KI 2014: advances in artificial intelligence, vol 8736 of lecture notes in computer science. Springer International Publishing, pp 80–85
15. Kraenbuehl P, Koltun V (2007) Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Proceedings of the 20th international conference on neural information processing systems. Vancouver, British Columbia
16. Krizhevsky A, Sutskever I, Hinton GE (2013) Imagenet classification with deep convolutional neural networks. In: NIPS
17. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. In: Proceedings of the IEEE
18. Lin TY (2014) Microsoft COCO: common objects in context. In: ECCV
19. Lin G, Shen C, Reid I (2015) Efficient piecewise training of deep structured models for semantic segmentation. arXiv:1504.01013
20. Liu Z, Li X, Luo P, Loy CC, Tang X (2015) Semantic image segmentation via deep parsing network. In: ICCV
21. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
22. Mostajabi M, Yadollahpour P, Shakhnarovich G (2015) Feed forward semantic segmentation with zoom-out features. In: CVPR
23. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. arXiv:1505.04366

24. Papandreou G, Kokkinos I, Savalle PA (2014) Untangling local and global deformations in deep convolutional networks for image classification and sliding window detection. arXiv:1412.0296
25. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: a deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147
26. Plath N, Toussaint M, Nakajima S (2009) Multi-class image segmentation using conditional random fields and global classification. In: Proceedings of the 26th annual international conference on machine learning, Montreal, Quebec, Canada
27. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: MICCAI
28. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229
29. Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV 2006. Springer, pp 1–15
30. Shotton J, Johnson M, Cipolla R (2008) Semantic texton forests for image categorization and segmentation. In: IEEE conference on computer vision and pattern recognition, pp 1–8
31. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
32. Socher R, Lin CC, Manning C, Ng AY (2011) Parsing natural scenes and natural language with recursive neural networks. In: ICML, pp 129–136
33. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. arXiv:1409.4842
34. Verbeek J, Triggs B Scene segmentation with conditional random fields learned from partially labeled images, Vancouver, British Columbia
35. Zhang Y, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. IEEE Trans Med Imaging 20(1):45–57
36. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr PH (2015) Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision, pp 1529–1537

Duc My Vo received his bachelor degree of engineering in automation from the University of Transport and Communication, Vietnam in 2006. He then moved to the Asian Institute of Technology, Thailand and finished his master degree in 2009. Afterwards, he continued his PhD study in the laboratory of Prof. Dr. Andreas Zells at the University of Tuebingen, Germany. His thesis projects were focused on addressing the use of RGB-D images for six important tasks of mobile robots: face detection, face tracking, face pose estimation, face recognition, person detection and person tracking. These topics have widely been researched in recent years because they provide mobile robots with abilities necessary to communicate with humans in natural ways. He was awarded his doctorate degree in 2015. After one year of research in the Vietnam National Satellite Center, he joined the Computer Vision and Multimedia Lab, Chosun University, South Korea in 2016. In 2017 he moved to the Pattern Recognition and Machine Learning Lab, Gachon University, South Korea. In his current project, he focuses on the development of action recognition, face recognition, semantic segmentation and biomedical imaging.

Sang-Woong Lee received his BS degree in Electronics and Computer Engineering from Korea University, Seoul, Korea, in 1996 and his MS and Ph.D. degrees in Computer Science and Engineering from Korea University, Seoul, Korea, in 2001 and 2006, respectively. From June 2006 to May 2007, he was a visiting scholar in Robotics Institute, Carnegie Mellon University. From September 2007 to February 2017, he worked as a professor in Department of Computer Engineering at Chosun University, Gwangju, Korea. Currently he is an associate professor in Department of Software at Gachon University. His present research interests include face recognition, computational aesthetics, machine learning, bioinformatics, and medical imaging analysis.