Attention Deeplabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation

Anonymous ECCV submission

Paper ID 100

Abstract. Skin lesion segmentation is a challenging task due to the large variation of anatomy across different cases. In the last few years, deep learning frameworks have shown high performance in image segmentation. In this paper, we propose Attention Deeplabv3+, an extended version of Deeplabv3+ for skin lesion segmentation by employing the idea of attention mechanism in two stages. We first capture the relationship between the channels of a set of feature maps by assigning a weight for each channel (i.e., channels attention). Channel attention allows the network to emphasize more on the informative and meaningful channels by a context gating mechanism. We also exploit the second level attention strategy to integrate different layers of the atrous convolution. It helps the network to focus on the more relevant field of view to the target. The proposed model is evaluated on three datasets ISIC 2017, ISIC 2018, and PH^2, achieving state-of-the-art performance [1].

Keywords: Medical Image Segmentation, Deeplabv3+ , Attention Mechanism

1 Introduction

With the development of computer vision, medical image segmentation has become an important part of computer-aided diagnosis. These years the computer-aided diagnosis (CAD) systems are required to assist the experts by providing accurate interpretation of medical images. Among many medical image processing tasks, automatic image segmentation is an important and effective step toward the analysis phase. Medical image segmentation is included in a large number of application domains like skin cancer segmentation. Skin cancer is one of the most widespread and deadly forms of cancer.

The human skin includes three types of tissues, i.e., dermis, epidermis, and hypodermis. The epidermis has melanocytes and under some conditions (like the strong ultraviolet radiation from sunshine), it produces melanin at a greatly unusual rate. A lethal type of skin cancer is melanoma which is the result of unusual growth of melanocytes [14]. With a mortality rate of 1.62%, melanoma is reported as the most lethal skin cancer [32]. In 2019, the American Cancer Society reported there are approximately 96,480 new cases of melanoma and about 7230 will die from this cancer [30]. The non-melanoma cancers are also the reason for a large number of deaths. The World Health Organization reported that between 2 and 3 million non-melanoma skin cancers and 132,000 melanoma skin cancers are recorded every year in the world [1].

[1] Code and trained models will be released in a public github page after publication of the paper

Although the dermatologists detect melanoma in medical images, their detection may be inaccurate. On the other hand, early diagnosis of the melanoma is critical in terms of treatment. The early detection and diagnosis of melanoma helps to have the proper treatment and ensure a complete recovery. It is reported that early detection increases the five-year relative survival rate to 92% [29]. As a result, skin lesion segmentation (Figure 1) has a critical role in the early and accurate diagnosis of skin cancer by computerized systems. During the last decade, automatic detection of skin cancer has been significantly taken into account. It is applied for different kinds of skin problems like three main types of abnormal skin cells are noticed i.e. Basic cell carcinoma, Squamous cell carcinoma and Melanoma .

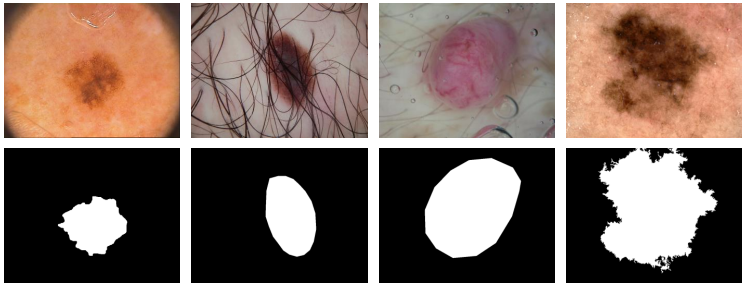


Fig. 1. Some samples of skin lesion segmentation.

Skin cancer segmentation is a challenging task due to several factors like low contrast of the images, differences in texture, position, color, and size of skin lesions in medical images. Moreover, the existence of air bubbles, hair, ebony frames, ruler marks, blood vessels, and color illumination make the lesion segmentation task extremely difficult. Various approaches have been proposed for the skin lesion segmentation. Different types of deep convolutional neural networks have been proposed for image segmentation, and like other fields of research in computer vision, deep learning approaches have achieved outstanding results in this field. Fully convolutional neural network (FCN) [21] was one of the first deep networks proposed for segmentation. This network was then extended to U-Net [27], consisting of an encoding and a decoding path. U-Net achieved good segmentation performance alongside leveraging the need of a large amount of training data.

CNN allows to learn increasingly abstract data representation which yields the network robust to local image transformation. Abstraction of spatial information may be undesirable for semantic segmentation. To solve that, Deeplab [7] was proposed by utilizing “atrous spatial pyramid pooling” (ASPP). Deeplabv3 [8] was then proposed to capture contextual information at multiple scales by employing several parallel ASPPs. Chen et al. [9] proposed DeepLabv3+ by combing the idea of both U-Net and Deeplabv3, as an extension of Deeplabv3 by adding a decoder module to recover the object boundaries. Recently, attention-based networks have been widely utilized in different tasks of computer vision [31]. The attention strategy helps the network by avoiding the use of multiple similar feature maps and focusing on the most salient and

informative features for a given task without additional supervision. It has been proved that attention enhances the result of semantic segmentation networks [35, 25, 31].

In this paper, we propose Att-Deeplabv3+ *(attention Deeplabv3+)* for skin lesion segmentation. In particular, we improve Deeplabv3+ by inserting two attention modules in the atrous convolution. The first attention module is a kind of channel wise attention and it is employed to recalibrate the feature map in each layer of the atrous convolution to pay more attention to more informative channels. That is to say, the network concentrates more on channel features with more useful information (based on their relationship) by assigning different weights to various channels of feature maps.

The second attention module is exploited to aggregate the features extracted by different layers of the atrous convolution through a multi-scale attention mechanism. Different layers of an atrous convolution extract features from the input features map at different sizes of field of view. Utilizing a small field of view for extracting features results in lower-level resolution which encodes local information. The local representation is important to discriminate the input image details. On the other hand, by enlarging the field of view, a larger image context is taken into account for extracting features, and therefore, more global information is taken to account. Especially, with large scales of receptive fields, the layer extracts long range features. The utilized scale attention module yields the network to emphasize the extracted features of the layers which are more relevant to the scale of the targets. We evaluate the proposed network on three datasets ISIC 2017, ISIC 2018, and Ph^2. The experimental results demonstrate that the proposed network achieves superior performance than existing alternatives.

The main contributions of the paper are as follows:

- We proposed an extended version of Deeplabv3+, Att-Deeplabv3+ for skin lesion segmentation.
- A channel wise attention module is utilized in each layer of the atrous convolution of Deeplabv3+ to focus on the more informative channel features.
- A multi-scale attention module is employed to aggregate the information of all layers of the atrous convolution.
- The proposed network achieves superior performance than existing alternatives on three datasets ISIC 2017, ISIC 2018, and PH^2.

The rest of the paper is organized as follows. Section 2 reviews related work. The proposed network is presented in Section 3. The experimental results are described in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

Semantic segmentation is one of the most important tasks in medical imaging. Before the revolution of deep learning in computer vision, traditional handcrafted features have been utilized in semantic segmentation. During the last few years, deep learning-based approaches have achieved outstanding results in image segmentation. These networks can be divided into two main groups, i.e., encoder-decoder structures and the models using spatial pyramid pooling [9].

2.1 Encoder-Decoder

The encoder-decoder networks have been successfully utilized for semantic segmentation. These networks include encoding and decoding paths. The encoder generates a large number of feature maps with reduced dimensionality, and the decoder produces segmentation maps by recovering the spatial information. U-Net [27] is one of the most popular encoder-decoder networks. A main advantage of this network is that the network works well with few training samples by employing the global location and context information at the same time. Many methods [20, 5, 4, 26] demonstrate the effectiveness of encoder-decoder structure for image segmentation in different applications.

Different extensions of the U-Net have been proposed for image segmentation. V-Net [23] is proposed to predict segmentation of a given volume. For segmentation of 3D volumes (e.g., MRI volumes), Çiçek et al. propose 3D U-Net [11]. BCDU-Net [5] improves the performance of the U-Net by utilizing ConvLSTM to combine the feature maps extracted from the corresponding encoding path and the previous decoding up-convolutional layer in a non-linear way. In that network, densely connected convolutions are also inserted in the last convolutional layer of the encoder to strengthen feature propagation. Inspired by the effectiveness of the recently proposed squeeze and excitation modules, Asadi et al. propose MCGU-Net [4] by inserting these modules in decoder for medical image segmentation.

Alom et al. [3] improve U-Net by using recurrent convolution network and recurrent residual convolutional network, and propose RU-Net and R2U-Net for medical image segmentation. While the U-Net uses skip connections to concatenate features from the encoder, the SegNet [6] passes pooling indices to the decoder to reduce computational complexities of weight concatenation. SegNet uses VGG-16-like network structure in the encoder and it works better than U-Net when the image content becomes complicated. Ghiasi et al. [15] propose a multi-resolution reconstruction architecture based on a Laplacian pyramid. In that network, the segmented boundaries of the low-resolution maps are reconstructed by the skip connections from the higher resolution maps. U-Net has been also improved by utilizing different attention-based modules in both encoder and decoder [25, 28].

2.2 Spatial pyramid pooling

The usual deep convolutional networks (DCNNs) have some drawbacks for semantic image segmentation. The spatial feature resolution is considerably decreased due to a consecutive max-pooling and down-sampling functions in the network. Moreover, objects (e.g., skin lesions) can have different scales in the images [7]. To mitigate this problem, spatial pyramid pooling models capture rich contextual information by pooling features at different resolutions.

PSPNet [34] is a pyramid scene parsing network to embed difficult scenery context features. In an FCN based pixel prediction framework, global context information is captured by using different region-based context aggregation through a pyramid pooling module. Atrous spatial pyramid pooling (ASPP) at several grid scales is performed in deeplab [7, 8]. In that network, parallel atrous convolution layers with different rates capture multi-scale information. Multi-scale information helps these models to improve

the performance of the network on several segmentation benchmarks. Hesamian et al. [16] utilize the Atrous convolution which increases the field of view of the filters and helps to improve the performance of the network for segmentation of Lung Nodule. Recurrent neural networks with LSTM have been also exploited in several methods [19, 33] to aggregate global context information for semantic segmentation.

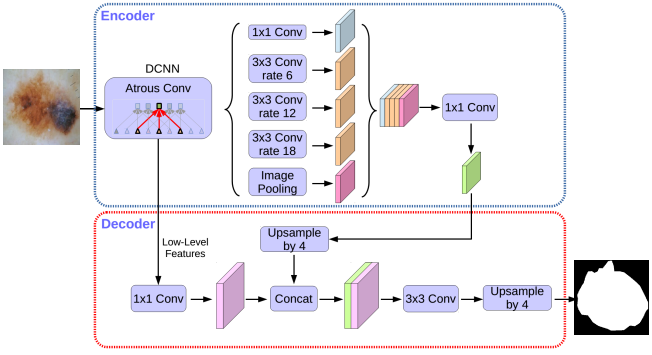


Fig. 2. Deeplabv3+ [9].

Chen et al. [9] take into account the advantages of both encoder-decoder networks and pyramid spatial pooling and introduce Deeplabv3+ shown in Figure 2. In that network, Deeplabv3 is extended by adding a decoder module to recover the object boundaries. Atrous convolution extracts rich semantic information in the encoding path and controls the density of the decoder features. The decoder module helps the network to recover the object boundaries. Many attention mechanisms have been utilized to improve the performance of the U-Net [25, 28]. In this paper, we improve the performance of the Deeplabv3+ by inserting two kinds of attentions, i.e, channel-wise attention and multi-scale attention, in the atrous convolutions.

3 Proposed Method

We propose Att-Deeplabv3+, an attention-based Deeplabv3+ extension, (Figure 3) for skin lesion segmentation. The network utilizes the strengths of attention mechanism to focus on the features with more informative data, and avoid the use of multiple similar feature maps. We highlight different parts of the proposed network in details in the following subsections.

3.1 Encoder

Xception model [10] works well in object classification with fast computation. This model has been adopted for the task of image segmentation [9], and achieved promising results. Therefore, we utilize the modified version of Xception model [9] as the encoder of the proposed network. In that method, a deeper Xception model is utilized, and extra

batch normalization and ReLU functions have been used after the 3×3 depthwise convolution. Moreover, all the max pooling layers are replaced by depthwise separable convolution with striding. By considering this idea, atrous separable convolution can be applied at an arbitrary resolution to extract features. This model consists of three parts, i.e., entry flow, middle flow, and exit flow. The entry flow includes 2 convolutional and 9 depthwise separable convolutional layers in four blocks. The middle flow contains one block consisting of three depthwise separable convolutional layers which is repeated 16 times. The exit flow includes 6 depthwise separable convolutional layers in 2 blocks. We refer the readers to [9] for more details.

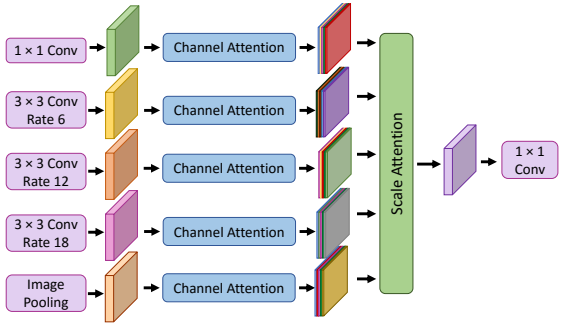


Fig. 3. Multi-level attention Deeplabv3+.

3.2 Attention-based Atrous Convolution

In usual CNNs, the spatial feature resolution is considerably decreased due to a set of consecutive max-pooling and down-sampling functions in the network. Moreover, objects can have different scales in the images. To mitigate this problem, atrous Convolutions with multiple rates [7] have been introduced. In atrous Convolutions, from the last few max pooling layers, the down-sampling operations have been removed while the filters have been up-sampled in the subsequent convolutional layers. To up-sample the filters, the full resolution image is convolved with filters with holes, i.e., zeros have been inserted between filter's values. Since non-zero filters values are only considered in the calculations, the number of parameters stay constant. The atrous convolution yields the method to control the spatial resolution of the feature responses. Moreover, we can enlarge the field of view of the filters to compute feature responses at any layer which results in incorporating larger context information. For one-dimensional signal, the atrous convolution [7] is calculated as

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

where y is the output feature map, i is a spatial location on y, x is the input feature map, and w is a convolution filter. Moreover, the atrous rate r determines the stride with which we sample the input signal.

As it can be seen in Figure 2, the atrous convolution consists of five layers. Each feature map includes a number of channels. In Deeplabv3+, the output of these layers are concatenated and passed to the next block of the network. In other words, all channels of a set of feature map are processed in the network with the same attention while some of these channels may be informative. Inspired by the squeeze and excitation network [17], we employ a channel-based attention on the output of each layer of the atrous convolution to capture explicit relationship between channels of the convolutional layers (shown in Figure 4). This strategy helps the network to selectively empathize informative features and suppress less useful ones by utilizing the global information of the input data. That is to say, it encodes feature maps by assigning a weight for each channel (i.e. channel attention).

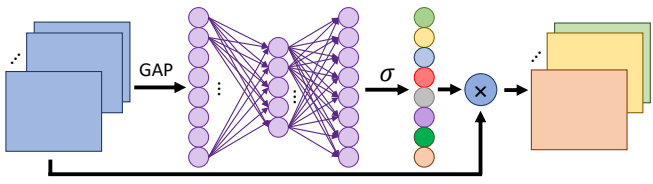


Fig. 4. Channel-wise attention for each layer of the atrous convolution.

The model exploits the global context information of the input features to produce the weight for each input channel. To do that, the global average pooling is calculated for each channel as

$$z_f = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} x_f(i,j) \tag{2}$$

where x_f is the fth channel, $H \times W$ is the size of the channel, and z_f is the output of the global average pooling. In the next step, we learn non-mutually-exclusive relationship and nonlinear interaction between channels. Inspired by [17], two fully connected layers are then employed to capture the channel-wise dependencies (Figure 4). The output of these layers is calculated as

$$s_f = \sigma\left(\mathbf{W}_2 \delta(\mathbf{W}_1 z_f)\right) \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the parameters of the two fc layers, δ is ReLU, and the σ refers to the sigmoid activation, and s_f is the learnt scale factor for the f_{th} channel. The final output of each layer of the attention-based atrous convolution is $X = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_F]$ where $\tilde{x}_f = s_f.x_f$ is a channel-wise multiplication between the fth channel x_f, and its corresponding scale factor s_f.

3.3 Multi-scale Attention-based Depthwise Aggregation in Atrous Convolution

In Deeplabv3+, the output of all layers of atrous convolution are concatenated and then passed to the depthwise separable convolution. Depthwise separable convolution factorizes the standard convolution into depthwise convolutions followed by a pointwise

convolution. In particular, a spatial convolution is independently employed for each input channel, and a pointwise convolution is then performed to combine the result of depthwise convolution for all of the input channels.

Atrous convolution produces multi-scale features with different resolutions containing different semantic information. Lower-level features encode more information about local representation while higher-level features focus on global representation. In particular, different layers of a multi-scale strategy contain different semantic information, i.e., the layers with lower resolution are responsible for smaller objects and layers with higher resolution are responsible for objects with a larger scale. In the original Deeplabv3+, these features are simply concatenated with the same weight. Since we have objects with different sizes, it is necessary to process the features of these layers with a different attention scale. To mitigate this problem, instead of a simple concatenation of multi-scale features, we integrate multi-scale features by employing a multi-scale depthwise attention strategy (Figure 5). Especially, the depthwise aggregation performs the integration independently for the corresponding channels of all five layers of atrous convolution.

Assume that X_s^f is the fth ($f \in \{1, 2, ..., 256\}$) channel of the sth ($s \in \{1, 2, ..., 5\}$) layer (scale) of the atrous convolution. The output of the original Deeplabv3+ is the concatenation of feature maps from all layers, i.e, $Y = [X_1, X_2, ...X_5]$. In Att-Deeplabv3+, these feature maps are integrated with a multi-scale attention depthwise method. We learn a weight for each scale, and apply a non-linear depthwise aggregation to combine these features. The output of the atrous convolution is calculated as

$$Y_f = \sigma\left(\sum_{s=1}^{5} w_s X_s^f\right) \tag{4}$$

where the X_s^f is the fth channel of the sth scale, w_s is the learnt weight for the sth scale, σ is the sigmoid function, and Y_f is the fth output channel of the atrous convolution.

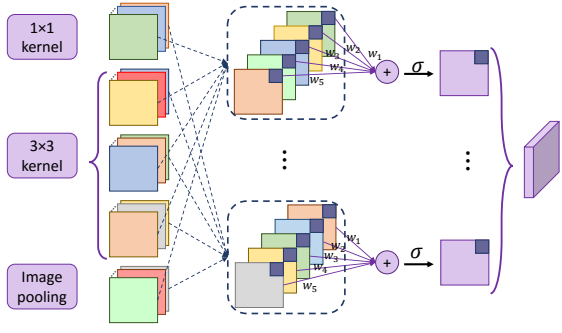


Fig. 5. Attention-based depthwise aggregation in atrous convolution.

3.4 Decoder

We utilize the same decoder as Deeplabv3+. To reduce the number of channels, a 1×1 convolution is applied on both encoder and decoder features. The encoder features are bilinearly upsampled and are then concatenated with the corresponding features (with the same spatial resolution) from the network backbone. To refine the features, the concatenated features are then passed to a few 3×3 convolutions. At the end, another bilinear upsampling is performed on the features.

4 Experimental Results

The proposed method is evaluated on three datasets ISIC 2017, ISIC 2018, and PH^2. Several performance metrics have been employed for the experimental comparative, including accuracy (AC), sensitivity (SE), specificity (SP), F1-Score, Jaccard similarity (JS), and area under the curve (AUC). We stop the training of the network when the validation loss remains the same in 10 consecutive epochs. For all three datasets, we finetune the network from a pre-trained model which is learnt on PASCAL VOC dataset.

4.1 ISIC 2017

The ISIC 2017 dataset [13] is designed for skin cancer segmentation published in 2017. Att-Deeplabv3+ is evaluated on the provided data for skin lesion segmentation. The dataset consists of 2000 skin lesion images with masks (including cancer or non-cancer lesions). Like other approaches [3], we use the standard evaluation setting for this dataset, i.e, 1250 samples for training, 150 samples for validation, and the other 600 samples for test. The original size of each sample is 576×767. We resize images to 256×256.

The segmentation outputs of the proposed network for some samples from this dataset are depicted in Figure 6. In Table 1, the results of the proposed network are compared with other approaches. By comparing the result of the proposed attention-based network with the original Deeplabv3+, we can see that the multi-level attention mechanisms enhance the performance of the network. The ROC diagrams of Att-Deeplabv3+ and the original Deeplabv3+ are depicted in Figure 7(a). By comparing the AUC scores shown on the diagrams, it is demonstrated that Att-Deeplabv3+ improves the performance of the original Deeplabv3+.

4.2 ISIC 2018

The International Skin Imaging Collaboration (ISIC) published the ISIC 2018 dataset [12] as a large-scale dataset of dermoscopy images in 2018. This dataset contains 2594 dermoscopy images. For each sample the original image and corresponding ground truth annotation (containing cancer or non-cancer lesions) are available. Like other approaches, we use the standard evaluation setting, and utilize 1815 images for training, 259 for validation and 520 for testing. We resize images to 256×256.

Figure 8 shows some segmentation outputs of the proposed network on ISIC 2018 dataset. In Table 2, the performance of the proposed method is compared with other

Table 1. Performance comparison of the proposed network and other methods on ISIC 2017 dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	Jaccard similarity
U-net [27]	0.8682	0.9479	0.9263	0.9314	0.9314
Melanoma det. [13]	-	-	-	o.9340	-
Lesion Analysis [18]	-	0.8250	0.9750	0.9340	-
R2U-net [3]	0.8920	**0.9414**	0.9425	0.9424	0.9421
BCDU-Net [5]	0.8810	0.8647	0.9751	0.9528	0.9528
MCGU-Net [4]	0.8927	0.8502	0.9855	0.9570	0.9570
Deeplabv3+ [9]	0.9162	0.8733	**0.9921**	0.9691	0.9691
Att-Deeplabv3+	**0.9190**	0.8851	0.9901	**0.9698**	**0.9698**

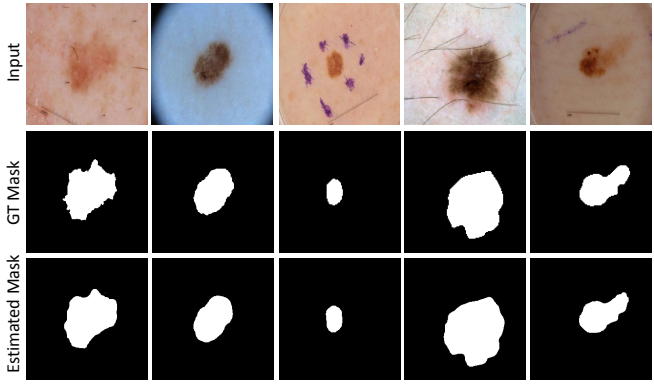
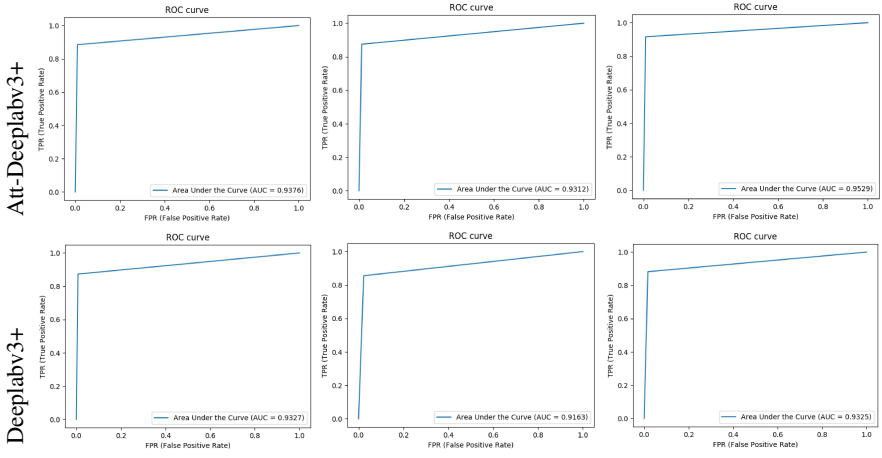


Fig. 6. Segmentation results of Att-Deeplabv3+ on ISIC 2017 dataset.

Table 2. Performance comparison of the proposed network and other methods on ISIC 2018 dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	Jaccard similarity
U-net [27]	0.647	0.708	0.964	0.890	0.549
Att U-net [25]	0.665	0.717	0.967	0.897	0.566
R2U-net [3]	0.679	0.792	0.928	0.880	0.581
Att R2U-Net [3]	0.691	0.726	0.971	0.904	0.592
BCDU-Net [5]	0.851	0.785	0.982	0.937	0.937
MCGU-Net [4]	0.895	0.848	0.986	0.955	0.955
Deeplab v3+ [9]	0.882	0.856	0.977	0.951	0.951
Att-Deeplab v3+	**0.912**	**0.875**	**0.988**	**0.964**	**0.964**

approaches. The proposed Att-Deeplabv3+ outperforms state-of-the-art methods. It can be seen there is a high gap between the result of the proposed method and the original Deeplabv3+. The ROC diagrams of the Att-Deeplabv3+ and the original Deeplabv3+ for ISIC 2018 dataset are shown in Figure 7(b). The AUC scores show that Att-deeplvabv3+ improves the performance of the original Deeplabv3+. We have also compute the threshold Jaccard index for comparing the performance of two networks of Att-Deeplabv3+ and the original Deeplabv3+. For Deeplabv3+, the threshold Jaccard index is 0.9404, and for the proposed network is 0.9599.



(a) ISIC 2017 dataset, (b) ISIC 2018 Dataset, and (c) PH^2

Fig. 7. The ROC curves of two networks of Att-Deeplabv3+ and and Deeplabv3+ for ISIC 2017, ISIC 2018, and PH^2 .

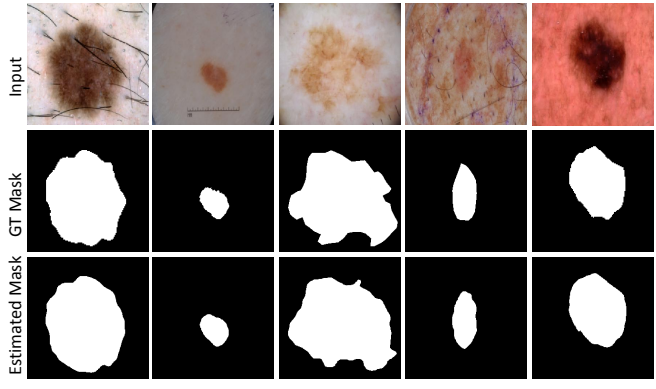


Fig. 8. Segmentation results of Att-Deeplabv3+ on ISIC 2018 dataset.

4.3 PH^2

The PH^2 dataset is a dermoscopic image database proposed for segmentation and classification [22]. The total number of samples for this dataset is 200 melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. The manual segmentations of the skin lesions are available as the ground truth. Each input image is a 8-bit RGB color images with the resolution of 768×560 pixels. There is not a standard evaluation setting (test and train sets) for this dataset. We use the same setting as [4] and randomly split the dataset into two sets of 100 images, and then use one set as the test data, 80% of the other set for the train, and the remained data for the validation.

Table 3. Performance comparison of the proposed network and other methods on PH^2 dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	Jaccard similarity
FCN [24]	0.8903	0.9030	0.9402	0.9282	0.8022
U-net [27]	0.8761	0.8163	0.9776	0.9255	0.7795
SegNet [6]	0.8936	0.8653	0.9661	0.9336	0.8077
FrCN [2]	0.9177	**0.9372**	0.9565	0.9508	0.8479
Deeplab v3+ [9]	0.9202	0.8818	0.9832	0.9503	0.9503
Att-Deeplab v3+	**0.9456**	0.9161	**0.9896**	**0.9657**	**0.9657**

Some precise and promising segmentation results of the proposed network for this dataset are shown in Figure 9. Table 3 lists the quantitative results obtained by other methods and the proposed network on PH^2 dataset. It is shown that the Att-Deeplabv3+ outperforms state-of-the-art approaches. Moreover, the proposed method surpasses the original Deeplabv3+. The ROC curves of the Att-Deeplabv3+ and the original Deeplabv3+ for PH^2 dataset are shown in Figure 7(c). By comparing these two curves, we find out that the multi-level attention idea is successful to enhance the performance of the network.

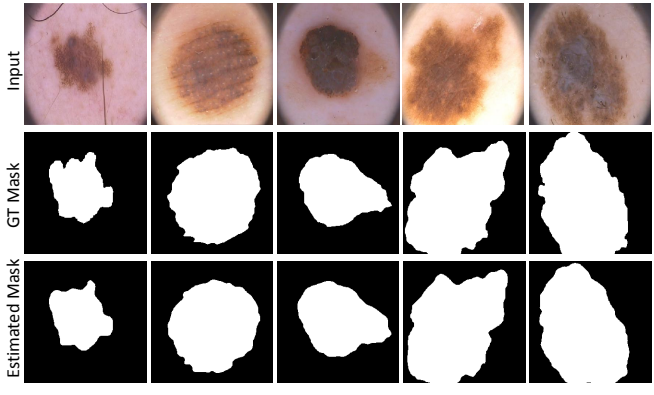


Fig. 9. Segmentation results of Att-Deeplabv3+ on PH^2 dataset

4.4 Discussion

In this paper, we employ multi-level attention mechanism to improve the performance of the Deeplabv3+. In Figure 10, the output segmentation results of some samples for both Att-deeplabv3+ and deeplabv3 networks are compared. It can be seen that Att-Deeplabv3+ performs better than the original network and its output results are more precious. The boundary of the segmented lesion of the Att-deeplabv3+ is finer, i.e., the output of the Att-Deeplabv3+ includes more local details in boundary of the object. Att-Deeplabv3+ is able to extract more useful and informative features from the input by employing the attention mechanism. In other words, the attention mechanism helps the network to focus more on feature maps with higher resolution for the segmentation.

In the proposed network, a channel wise attention is first applied to all the layers of the atrous convolution. This attention mechanism yields the network to focus on the more informative channels. The second level of the attention is used as a multi-scale attention method for aggregating all the layers of the atrous convolution. The multi-scale attention allows the network to focus on more scale relevant feature to the target.

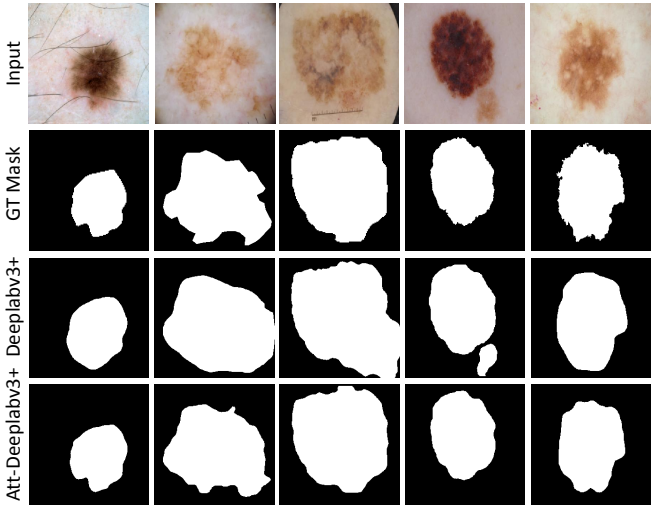


Fig. 10. Visual effect of the multi-level attention mechanism in the proposed network.

5 Conclusion

We proposed Att-Deeplabv3+ for skin lesion segmentation. It has been shown that by including two level attention based mechanism in Deeplabv3+, the network is able to learn more discriminative information. Compared to the original deeplabv3+, the proposed network has more precise segmentation results. The experimental results on three public skin cancer benchmark datasets showed high gain in semantic segmentation in relation to state-of-the-art alternatives.

References

1. who.int/uv/faq/skincancer/en/index1.html
2. Al-Masni, M.A., Al-antari, M.A., Choi, M.T., Han, S.M., Kim, T.S.: Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Computer methods and programs in biomedicine **162**, 221–231 (2018)
3. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955 (2018)
4. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020)
5. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
6. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
10. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
11. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
12. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
13. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th ISBI 2018. pp. 168–172. IEEE (2018)
14. Feng, J., Isern, N.G., Burton, S.D., Hu, J.Z.: Studies of secondary melanoma on c57bl/6j mouse liver using 1h nmr metabolomics. Metabolites **3**(4), 1011–1035 (2013)
15. Ghiasi, G., Fowlkes, C.C.: Laplacian pyramid reconstruction and refinement for semantic segmentation. In: European Conference on Computer Vision. pp. 519–534. Springer (2016)
16. Hesamian, M.H., Jia, W., He, X., Kennedy, P.J.: Atrous convolution for binary semantic segmentation of lung nodule. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1015–1019 (May 2019)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the CVPR. pp. 7132–7141 (2018)

18. Li, Y., Shen, L.: Skin lesion analysis towards melanoma detection using deep learning network. Sensors **18**(2), 556 (2018)
19. Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., Yan, S.: Semantic object parsing with local-global long short-term memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3185–3193 (2016)
20. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
22. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: Ph 2-a dermoscopic image database for research and benchmarking. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 5437–5440. IEEE (2013)
23. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
24. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the CVPR. pp. 1520–1528 (2015)
25. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
26. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4353–4361 (2017)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
28. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis **53**, 197–207 (2019)
29. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2018. CA: A Cancer Journal for Clinicians **68**(1), 7–30 (2018)
30. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: a cancer journal for clinicians **69**(1), 7–34 (2019)
31. Sinha, A., Dolz, J.: Multi-scale guided attention for medical image segmentation. arXiv preprint arXiv:1906.02849 (2019)
32. Tarver, T.: American cancer society. cancer facts and figures 2014. J Consumer Health Internet **16**, 366–367 (2012)
33. Wang, G., Luo, P., Lin, L., Wang, X.: Learning object interactions and descriptions for semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5859–5867 (2017)
34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
35. Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 267–283 (2018)