Spatial and Channel Attention Modulated Network for Medical Image Segmentation

Wenhao Fang[1] and Xian-hua Han[1 [0000−0002−5003−3180]]

Graduate School of Science and Technology for Innovation,
Yamaguchi University, Japan
b501vb@yamaguchi-u.ac.jp
hanxhua@yamaguchi-u.ac.jp

Abstract. Medical image segmentation is a fundamental and challenge task in many computer-aided diagnosis and surgery systems, and attracts numerous research attention in computer vision and medical image processing fields. Recently, deep learning based medical image segmentation has been widely investigated and provided state-of-the-art performance for different modalities of medical data. Therein, U-Net consisting of the contracting path for context capturing and the symmetric expanding path for precise localization, has become a meta network architecture for medical image segmentation, and manifests acceptable results even with moderate scale of training data. This study proposes a novel attention modulated network based on the baseline U-Net, and explores embedded spatial and channel attention modules for adaptively highlighting interdependent channel maps and focusing on more discriminant regions via investigating relevant feature association. The proposed spatial and channel attention modules can be used in a plug and play manner and embedded after any learned feature map for adaptively emphasizing discriminant features and neglecting irrelevant information. Furthermore, we propose two aggregation approaches for integrating the learned spatial and channel attentions to the raw feature maps. Extensive experiments on two benchmark medical image datasets validate that our proposed network architecture manifests superior performance compared to the baseline U-Net and its several variants.

1 Introduction

In modern and clinic medicine, medical images have played an important role for conducting accurate disease diagnosis and effective treatment. A large number of medical images using different imaging technologies, such as X-ray, computed tomography (CT), ultrasound, and magnetic resonance imaging (MRI) and so on, have made great contributions to research evolution for developing computer-aided diagnosis (CAD) systems [1] using image processing and machine learning [2–5]. On the contrary, the developed CAD system is prospected to conduct rapid analysis and understanding of large amount of medical data to reduce the doctor's interpretation time, and further extends the wide use of medical images in clinic medicine. The CAD systems can not only conduct fast screening for

supporting doctors but also provide quantitative medical image evaluation to assist more accurate treatment. There proposed a lot of CAD systems in recent decades of years. Therein automatic medical image segmentation that extracts specific organs, lesion or regions of interest (ROI) [6–8] in medical images is a crucial step for the downstream tasks of medical image analysis systems. Traditional medical image segmentation methods mainly rely on hand engineered feature for classifying pixels independently. Due to large variation of uncontrollable and complex geometric structures in medical images, traditional methods usually lead to unsatisfied segmentation results. Inspired by the great success of deep convolutional neural network for image classification in recent years, deep learning [9] based methods has been widely investigated and provided impressive performance for different vision tasks including semantic image segmentation. In the last few years, numerous CNN [10] models have been proposed and validated that deeper networks generally result in better performance for different recognition and segmentation tasks. However, it is mandatory to prepare large scale of annotated samples for training very deep models, which is difficult in medical application scenario. Further the training procedure for a very deep model is usually unstable due to the vanishing or explosive gradient problems and needs rich experience for hyper-parameter turning.

In semantic medical image segmentation scenario, simple network architectures are most preferred due to small scale of annotated training samples. In 2015, a simple and easily implemented CNN architecture: U-Net [11] was proposed specifically for medical image segmentation, and has become a very popular meta architecture for different modalities of medical data segmentation. To boost segmentation performance, different variants of U-Net via integrating more advance modules such as recurrent unit, residual block or attention mechanism, have been widely investigated. Among them, the attention mechanism manifests promising performance for segmentation task on different CNN architectures including U-Net. Many existing attention approach usually investigate spatial attention via focusing on salient regions, which aid at better estimation of the under-studying target greatly while may neglect the possible different contribution in the learned feature maps. Thus it still deserves the further studying of exploring different attentions not only on spatial domain but also on channel direction.

This study proposes a novel attention modulated network based on the baseline U-Net, and explores embedded spatial and channel attention modules for adaptively highlighting interdependent channel maps and focusing on more discriminant regions via investigating relevant feature association. The two explored attention modules: spatial attention module (SAM) and channel attention module (CAM) can be employed to any feature map for emphasizing discriminant region and selecting important channel maps in a plug and play manner, and further can be combined as spatial and channel attention module (SCAM) for simultaneously conducting spatial and channel attention. In addition, we propose two aggregation approaches for integrating the learned spatial and channel attentions into the raw feature map. Extensive experiments on two benchmark

medical image datasets validate that our proposed network architecture manifests superior performance compared to the baseline U-Net and its several variants.

2 Related Work

In the past few years, semantic medical image segmentation has been actively researched in computer vision and medical image processing community, and substantial improvement have been witnessed. This work mainly concentrates on the medical image segmentation using deep learning methods. Here, we briefly survey the related work.

2.1 Medical image segmentation

Semantic segmentation of medical images is a crucial step in many downstream medical image analysis and understanding tasks, and has been extensively studied for decades of years. Traditional medical image segmentation approaches generally employ hand engineered features for classifying pixels independently into semantic regions, and lead to unexpected results for the images with large variation in intensities. In the last few years, with the rapid evolution of deep learning technique, many medical image segmentation models based on convolutional neural network (CNN) have been proposed. Via replacing the fully connected layers of standard classification CNNs with convolutional layers, fully CNN (FCN) [12] has been proposed to conduct dense pixel prediction at one forward step, and successfully applied for generic object segmentation. Further, FCN employs skip connection among network for reusing the intermediate feature maps to improve the prediction capabilities. Later many variants inspired from the FCN such as SegNet [13], DeepLab [14] have been investigated for boosting segmentation performance and made great progress for generic image segmentation in computer vision applications.

On the other hand, U-Net architecture was firstly proposed specifically for semantic medical image segmentation, and has become very popular due to its simple implementation and efficiency for network training. In this architecture, there have contractive and expansive paths, where contractive path is implemented using the combination of convolutional and pooling layers for learning different scales of contexts while expansive path employs the combination of convolutional and upsampling layers for mining semantic information. Then, similarly as in FCN [12], skip connections are used to concatenate the context and semantic information from two paths for accuracy prediction. To further improve segmentation results, different variants of U-Net models have been proposed. Kayalibay et al. [10] proposed to integrate multiple segmentation maps and forward feature maps from different paths, and then predict the final segmentation results from the integrated maps. Drozdzal et al. [15] explored and evaluated the importance of skip connections for biomedical image segmentation while Reza Azad et al. [16] proposed to employ convLSTM unit to integrate the feature maps from two paths instead of simple skip connection. Chen

et al. [17] proposed a deep contour-aware network (DCAN) to extract multi-level contextual features with a hierarchical architecture while McKinley et al. [18] designed a deep dig-like convolutional architecture, named as Nabla-Net for biomedical image segmentation. Further, several works [19, 20] embedded recurrent and residual structures into the baseline U-Net model, and showed impressive performance.

The other research direction is to extend the conventional U-Net to 3D counterpart for 3D medical image segmentation tasks. V-Net: a powerful end-to-end 3D medical image segmentation model [21], has firstly been proposed via combining FCN [12] and residual connections while a deeply supervised 3D model [22] was explored attempting to employ multi-block features for final segmentation prediction. To refine the segmentation results from CNN model, Kamnitsas et al. [23] integrated fully connected CRF into a multi-scale 3D CNN for brain tumor segmentation. Residual structure in the 3D CNN model was also extensively studied for medical image segmentation such as High-Res3DNet [24] and Voxresnet [25].

2.2 Deep Attention Network

Attention mechanisms are capable of emphasizing important and relevant element of the input or the under-studying target via learning strategy, and thus have become a very popular component in deep neural network. The integration of these attention modules have made great progress in many vision tasks such as image question-answering [26], image captioning [27] and classification [28]. Attention mechanisms have also been integrated into semantic image segmentation networks, and proven performance beneficial for this pixel-wise recognition tasks [29–34]. For instance, Zhao et al. [30] proposed a point-wise spatial attention network (PSANet), which allows a flexible and dynamic aggregation of different contextual information by connecting each position in the feature map with all the others through adaptive attention maps. Fu et al. [31] investigated a dual attention network for scene segmentation. Despite the growing interest on exploring attention mechanisms for image segmentation of natural scenes, there are still limited work for adopting to medical images segmentation. The existed study for integrating attention mechanisms into medical image segmentation networks [35–39], generally employ simple attention models, and the improvement are limited.

3 Spatial and channel modulate network

This study aims to explore a novel spatial and channel modulate network. We combine attention mechanism with U-Net to propose a attention modulate network for semantic segmentation of medical images. The schematic concept of the proposed SCAM-Net is given in Fig. 1. The mainstream of the proposed SCAM-Net follows the encoder-decoder architecture, and various feature maps, which

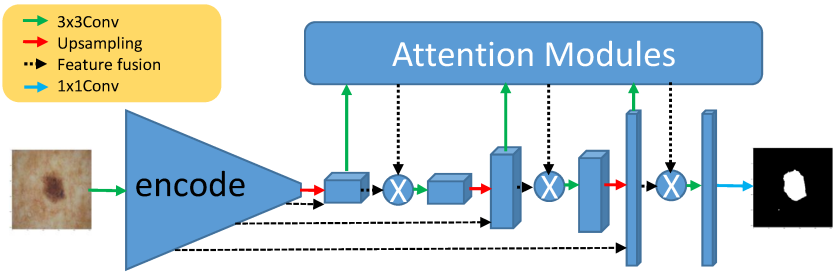


Fig. 1. The schematic concept of the proposed SCAM-Net

may contribute differences to final prediction, can be learned for representing different contexts. To adaptively learn more effective contexts for better predicting target in the model training procedure, we propose to leverage attention mechanism, which can emphasize important and relevant elements among the learned maps. We explore two attention modules: spatial attention module (SAM) and channel attention module (CAM), which can be employed to any feature maps of the main encoder-decoder architecture (U-Net), and then combine them as spatial and channel attention module (SCAM) for simultaneously conducting spatial and channel attention. Next, we would introduce the mainstream of the encoder-decoder architecture, and then describe the spatial, channel attention modules and their combinations.

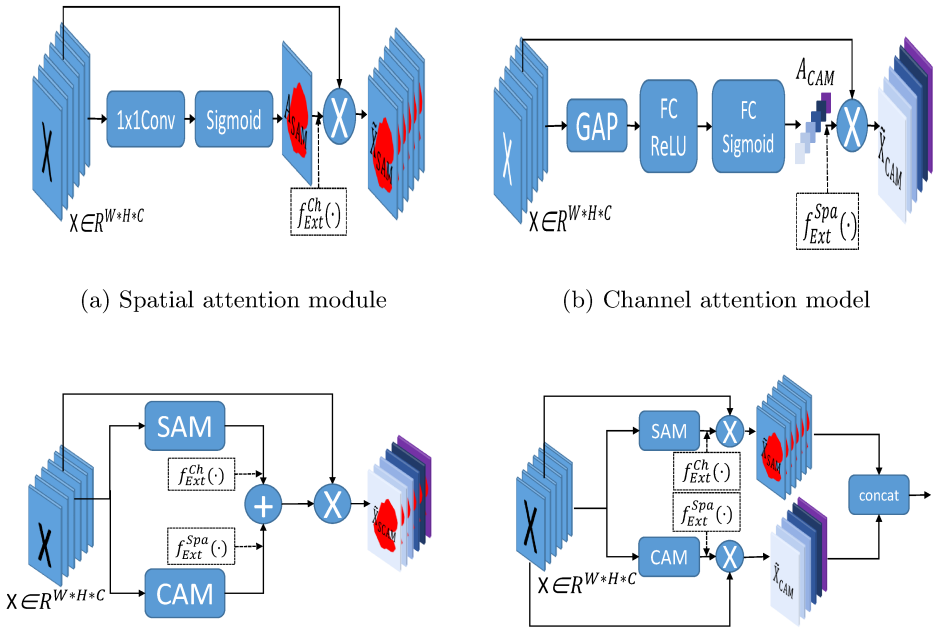
3.1 The mainstream of the encoder-decoder architecture

The used main backbone network follows U-Net architecture consisting of two paths: encoder and decoder. Both encoder and decoder paths are divided into four blocks, and each block is mainly implemented in 3 convolutional layers with kernel sizes 3*3 following RELU activation function after each convolutional layer. The channel numbers of feature maps are increased to double and feature maps sizes are decreased to half via employing MaxPooling layer with a 2*2 kernel in horizontal and vertical directions, respectively, between blocks of the encoder while decreasing to half of channel number and increasing to double of map sizes are inversely implemented with up-sampling layers between blocks of the decoder. It is known that the encoder generally extracts multi-scale features retaining detail structures of the input while the decoder learns multiple features with more semantic information for predicting the target. However not all semantic features learned by the decoder path aid to prediction of the specific under-studying target and may have different contributions to the final result. Thus we employ attention mechanism to emphasize discriminant region and select important channel maps for boosting performance.

Let us denote the learned feature map of a block of the decoder as $\mathbf{X} \in \Re^{W \times H \times C}$, we implement a series of transformations for extraction attention map of $\mathbf{X}$, and then add the explored attention map to the raw feature for emphasizing important context, which is formulated as the following:

$$\bar{\mathbf{X}} = \mathbf{X} + f_{AM}(\mathbf{X}) \tag{1}$$

where $f_{AM}(\cdot)$ denotes the transformation operators for extracting attention map. With the attention module in Eq. (1), it is expected that more discriminant features for prediction of the target can be adaptively and automatically emphasized in model training procedure. Further the attention module can be feasibly employed to any learned feature of the decoder in a plug and play manner. Next, we would describe the detail implementation of our proposed spatial and channel attention modules.



(a) Spatial attention module

(b) Channel attention model

(c) Attention fusion (SCAM-AF)

(d) Attention modulated feature fusion (SCAM-AMFF)

Fig. 2. Different attention modules

3.2 Spatial attention module (SAM)

The simple up-sampling process of the decoding path in the mainstream architecture may lead to un-expected spatial information and detail structure lost. To solve this problem, U-Net employs skip connections to combine (concatenate) the feature map with detail spatial information in the encoding path and the feature map of the decoding path. However, this simple concatenation brings many redundant low-level features. Therefore, we leverage a spatial attention module (SAM) in the decoding path to effectively suppress the activation regions with little discriminant information and thereby reduce the number of redundant features. The structure of the proposed SAM is shown in Fig. 2(a).

Given a feature map extracted by a block of the decoder as $\mathbf{X} \in \Re^{W \times H \times C}$, we implement the spatial attention mechanism via firstly employing a convolutional layer with 1*1 kernel and output channel 1, being formulated as:

$$\mathbf{X}_{SAM} = f_{Conv1*1}(\mathbf{X}) \tag{2}$$

where $\mathbf{X}_{SAM} \in \Re^{W \times H}$ has the same spatial size with $\mathbf{X}$. Then a non-linear transformation is conducted to generate the spatial attention map with magnitude range [0, 1] using an activation function, where a coefficient close to 1 indicates more relevant features. The activation operation is expressed as:

$$A_{SAM} = \sigma(\mathbf{X}_{SAM}) \tag{3}$$

where $\sigma(\cdot)$ is sigmoid activation function. Finally, the extracted spatial attention map is employed to the raw feature map $\mathbf{X}$ for emphasizing discriminant regions:

$$\bar{\mathbf{X}}_{SAM} = \mathbf{X} \otimes f_{SAM}(\mathbf{X}) = \mathbf{X} \otimes f_{Ext}^{Ch}(A_{SAM}) \tag{4}$$

where $f_{Ext}^{Ch}(\cdot)$ extends the spatial attention map in channel direction to the same size of $\mathbf{X}$ for being combined with the raw feature map. After that, it is passed normally into the mainstream.

3.3 Channel attention module (CAM)

Recently, the channel attention module has attracted a lot of interest and has shown great potential for improving the performance of deep CNN. The core idea is to automatically learn the indexed weights for each channel of feature map, so that the feature maps with more important information for final result prediction have larger weights while the feature maps with invalid or less discriminant information have small weights.

We implement the channel attention via exploring the correlations between different channels of features. The learned feature maps $\mathbf{X}$ in the decoder's block are aggregated to generate channel contribution index by employing global average pooling, formulated as:

$$m_k = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} x_k(w, h) \tag{5}$$

where $x_k(w,h)$ denotes the feature value on the spatial position (w,h) and the channel k of the feature map in $\mathbf{X}$, and m_k represents the global information of the $k-th$ channel of feature map. Then the channel-wise dependencies are investigated via using two fully connected (FC) layers. The first FC layer encodes the channel global vector $\mathbf{m} = [m_1, m_2, \cdots, m_K]^T$ to a dimension-reduced vector with reduction ratio while the second FC layer recovers it back again to the raw channel K as an the channel attention vector $\mathbf{X}_{CAM}$, which is expressed as the following:

$$\mathbf{X}_{CAM} = \mathbf{W}_2(\mathbf{W}_1\mathbf{m}) \tag{6}$$

where $\mathbf{W}_1 \in \Re^{\frac{K}{r}\times K}$ and $\mathbf{W}_2 \in \Re^{K\times\frac{K}{r}}$ represent the parameters of the two FC layers, respectively, and the r represents the ratio of scaling parameters. In our experiment, there is a compromise between accuracy and parameter amount(r=16).

Then, similar as in the SAM, a non-linear transformation is conducted to generate the attention map with magnitude range [0, 1] using a sigmoid activation function $\sigma(\cdot)$, which is expressed as:

$$A_{CAM} = \sigma(\mathbf{X}_{CAM}) \tag{7}$$

Finally, the channel attention modulated feature map is formulated as:

$$\bar{\mathbf{X}}_{CAM} = \mathbf{X} \otimes f_{CAM}(\mathbf{X}) = \mathbf{X} \otimes f_{Ext}^{Spa}(A_{CAM}) \tag{8}$$

where $f_{Ext}^{Spa}(\cdot)$ extends the channel attention map in spatial direction to the same size of $\mathbf{X}$. Similar as in SAM, it will be passed normally into the mainstream.

3.4 Spatial and channel attention module (SCAM)

In view of the above two attention modules, it naturally leads to the consideration of combining these two attention modules to generate a spatial and channel attention module for simultaneously emphasizing discriminant regions and selecting useful channel features. We explore two aggregation strategies, and the conceptual diagrams of the two methods are shown in Fig. 2(c) and Fig. 2(d), respectively.

The flowchart of the first aggregation method, called as attention fusion based SCAM (SCAM-AF), is shown in Fig. 2(c), which intuitively integrates the extended spatial and channel attention maps using element-wise addition, as expressed in the following:

$$\mathbf{A}_{SCAM} = f_{Ext}^{Ch}(A_{Spatial}) + f_{Ext}^{Spa}(A_{Spectral}) \tag{9}$$

Then, the attention map is added to the raw feature map for generating attention modulated feature map:

$$\bar{\mathbf{X}}_{SCAM} = \mathbf{X} \otimes \mathbf{A}_{SCAM} \tag{10}$$

The second combination method, called as attention modulated feature fusion (SCAM-AMFF), fuses the separately modulated feature maps by the SAM and CAM using a concat layer shown in Fig. 2(d). The approach also combines the feature maps of the two attention modules. It is also possible to give more weight to the space and the more effective areas in the channel at the same time, as expressed in the following:

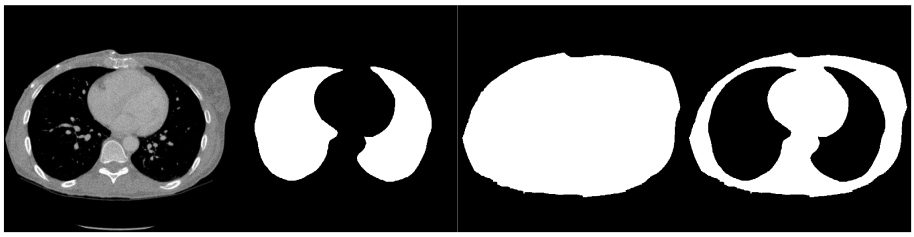
$$\bar{\mathbf{X}}_{SCAM} = Concat(\bar{\mathbf{X}}_{SAM}, \bar{\mathbf{X}}_{SAM}) \tag{11}$$

In general, the two aggregation methods can conduct both special and channel attention to adaptively learning discriminant information, and thus benefit for more efficient training of our medical image segmentation model. In addition, the attention modulated modules can be integrated with any feature map extracted in decoder 's blocks, and is expected to learn more effective features for predicting the under-studying target.

4 Experimental setup and Results

4.1 Database

Lung Segmentation dataset: The used lung segmentation dataset was presented in 2017 at the Kaggle Data Science Bowl in the Lung Nodule Analysis (LUNA) competition. This dataset consists of 2D and 3D CT images with respective label images for lung segmentation. In this paper, we used a total of 1021 slice images with size 512×512 extracted from the 3D images as the dataset. We use 70% of the dataset as the train subset and the remaining 30% as the test subset. Since the lung region in the CT image have almost the same Hausdorff value with non-interested object: air region but contains interference factors such as alveoli and blood vessels in small regions, it would be difficult to perfectly distinguish the lung from the region with the same Hausdorff value. It is known that the lung is surrounded by other body tissues anatomically and the regions outside body tissues are air. Thus it is prospected to achieve better performance via taking the lung and air regions as the same class, and other tissue region as the other class for segmentation learning. After obtaining the results of the



Lung image Lung mask All tissue mask The used mask

Fig. 3. The used ground-truth mask in our experiments for lung segmentation.

Table 1. Performance comparison of the proposed attention modulated networks and the state-of-the-art methods on LUNA dataset.

Models	F1-Score	Sensitivity	Specificity	Accuracy	AUC
U-Net	0.9658	0.9696	0.9872	0.9872	0.9784
RU-Net	0.9638	0.9734	0.9866	0.9836	0.9800
R2U-Net	0.9832	**0.9944**	0.9832	0.9918	0.9889
SAM	0.9736	0.9890	0.9955	0.9922	0.9918
CAM	0.9634	0.9936	0.9860	0.9873	0.9898
SCAM-AF	**0.9841**	0.9823	**0.9971**	**0.9946**	0.9897
SCAM-AMFF	0.9800	0.9902	0.9938	0.9932	**0.9920**

lung and air region segmentation, it is very easy to get the lung region due the complete separation by other body tissues between the two regions. The used ground-truth mask for network training in our experiment are shown in Fig. 3

Skin Segmentation dataset: the ISIC dataset is a large-scale dermoscopy image dataset, which was released by the International Dermatology Collaboration Organization (ISIC). This dataset is taken from a challenge on lesion segmentation, dermoscopic feature detection, and disease classification. It includes 2594 images, in which we used 1815 images for training, 259 for validation and 520 for testing. The training subset consists of the original images and corresponding ground truth annotations. The original size of each sample is 700 $\times$ 900, and was resized to 256 $\times$ 256 in our experiments.

Table 2. Performance comparison of the proposed attention modulated networks and the state-of-the-art methods on ISIC dataset.

Models	F1-Score	Sensitivity	Specificity	Accuracy	Precision
U-Net	0.647	0.708	0.964	0.890	0.779
Attention U-Net	0.665	0.717	0.967	0.897	0.787
RU-Net	0.679	0.792	0.928	0.880	0.741
R2U-Net	0.691	0.726	0.971	0.904	0.822
SAM	0.773	0.699	0.970	0.913	0.866
CAM	0.851	0.779	**0.986**	0.942	0.938
SCAM-AF	**0.870**	**0.817**	0.983	**0.948**	0.931
SCAM-AMFF	0.869	0.809	**0.986**	**0.948**	**0.940**

4.2 Evaluation Results

We evaluate the experimental results using several quantitative metrics including accuracy (AC), F1-Score, sensitivity (SE), specificity (SP), precision (PC) and

area under the curve (AUC). The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values are needed for calculating the evaluation metrics, which is expressed as:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$PC = \frac{TP}{TP + FP} \tag{13}$$

$$SE = \frac{TP}{TP + FN} \tag{14}$$

$$SP = \frac{TN}{TN + FP} \tag{15}$$

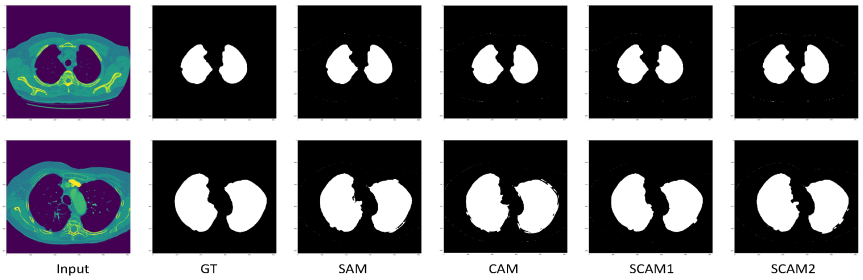
$$F1 - score = \frac{2SE * PC}{SE + PC} \tag{16}$$

To evaluate the effectiveness of the proposed SCAM-Net, we provide the compared results with several state-of-the-art methods including the baseline U-Net[11], Recurrent Residual U-Net[19], Attention U-Net[40], R2U-Net[20], and our proposed network with SAM or CAM for both skin lesion segmentation (ISIC) and lung segmentation dataset.

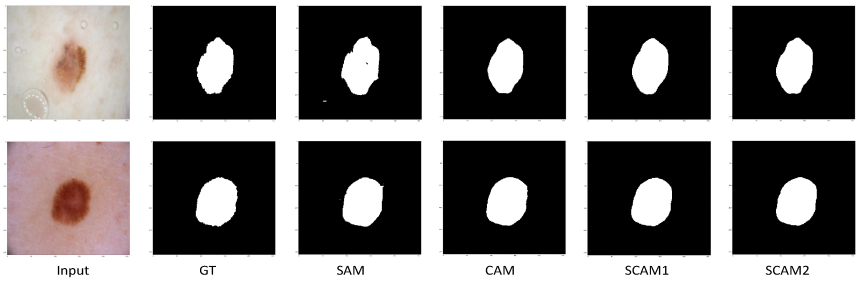
Table 2 and Table 1 provides the compared quantitative evaluations on two datasets, which demonstrates improved results compared with the baseline U-Net and its variants. At the same time, the proposed network with only one attention module can also achieve better performance than the baseline U-Net method, and better or comparable results with the extended version of U-Net. Meanwhile, it can be seen from Table 1 and 2 that CAM performs better than SAM in the ISIC dataset regard with the quantitative evaluation while SAM performs better than CAM in the LUNA lung segmentation dataset. Thus different attention models may be applicable to different datasets and deserved to be further investigated. Next, we conducted experiments on both datasets using the combined attention modules (SCAM-AF and SCAM-AMFF), and the compared results are also provided in Table 1 and 2, which manifests that the quantitative evaluation with the proposed SCAMs is better than not only the baseline U-Net but also the proposed networks with only one attention module (SAM or CAM). Finally, the visualization results of segmentation for two example images on both the LUNA and ISIC datasets, are shown in the Fig. 4, which manifests the segmentation results using the proposed networks with different attention modules are very similar to the ground-truth annotation.

5 Conclusion

This study proposed a novel spatial and channel attention modulated network for effective segmentation of medical images. module. To emphasize discriminate regions and adaptive select more important channel of feature maps, we explored both spatial and channel attention modules for integrating into the



(a) Results of two examples images from the LUNA datasets



(b) Results of two examples images from the ISIC datasets

Fig. 4. The segmentation results of two example images from the LUNA and ISIC datasets, respectively, using the backbone network with different attention modules. (a) Results of two examples images from the LUNA datasets. (b) Results of two examples images from the ISIC datasets.

main encoder-decoder architecture in a plug and play manner. Further, we proposed two aggregation strategies to combine the two attention modules into a unified unit in the mainstream network for boosting segmentation performance. Comprehensive experiments on two benchmark medical data sets showed that our proposed method not only obtains better performance than the baseline U-Net and its variants, but also manifested encouraging performance compared with a single space or single channel attention module.

6 ACKNOWLEDGE

This research was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20K11867.

References

1. Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested convolutional neural networks for auto-

mated pancreas localization and segmentation. Medical Image Analysis **45** (2018) 94–107
2. Cerrolaza, J.J., Summers, R.M., Linguraru, M.G.: Soft multi-organ shape models via generalized pca: A general framework. In: MICCAI. (2016)
3. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K.S., Davidson, B.R., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal ct with dense dilated networks. In: MICCAI. (2017)
4. Saito, A., Nawano, S., Shimizu, A.: Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. Medical image analysis **28** (2016) 46–65
5. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Carapella, V., Kim, Y.J., Suzuki, H., Kainz, B., Matthews, P.M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Glocker, B., Rueckert, D.: Human-level cmr image analysis with deep fully convolutional networks. ArXiv **abs/1710.09289** (2017)
6. Shih, F., Zhong, X.: High-capacity multiple regions of interest watermarking for medical images. Inf. Sci. **367-368** (2016) 648–659
7. Sanchez, V.: Joint source/channel coding for prioritized wireless transmission of multiple 3-d regions of interest in 3-d medical imaging data. IEEE Transactions on Biomedical Engineering **60** (2013) 397–405
8. Raja, J.A., Raja, G., Khan, A.: Selective compression of medical images using multiple regions of interest. (2013)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
10. Kayalibay, B., Jensen, G., van der Smagt, P.: Cnn-based segmentation of medical imaging data. CoRR **abs/1701.03056** (2017)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. (2015)
12. Khened, M., Varghese, A., Krishnamurthi, G.: Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. CoRR **abs/1801.05173** (2018)
13. 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** (2017) 2481–2495
14. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018) 834–848
15. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: LABELS/DLMIA@MICCAI. (2016)
16. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (2019) 406–415
17. Chen, H., Qi, X., Yu, L., Heng, P.: Dcan: Deep contour-aware networks for accurate gland segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2487–2496

18. McKinley, R., Wepfer, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., Reyes, M.: Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation. In: BrainLes@MICCAI. (2016)
19. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T., Asari, V.: Recurrent residual u-net for medical image segmentation. Journal of Medical Imaging **6** (2019) 014006 – 014006
20. Alom, M., Hasan, M., Yakopcic, C., Taha, T., Asari, V.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. ArXiv **abs/1802.06955** (2018)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV) (2016) 565–571
22. Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.: 3d deeply supervised network for automated segmentation of volumetric medical images. Medical Image Analysis **41** (2017) 40–54
23. Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A.D., Menon, D., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis **36** (2017) 61–78
24. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M., Vercauteren, T.K.M.: On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. ArXiv **abs/1707.01992** (2017)
25. Chen, H., Dou, Q., Yu, L., Heng, P.: Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. ArXiv **abs/1608.05895** (2016)
26. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. CoRR **abs/1511.02274** (2015)
27. Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. CoRR **abs/1612.01033** (2016)
28. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. CoRR **abs/1704.06904** (2017)
29. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. CoRR **abs/1511.03339** (2015)
30. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)
31. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. CoRR **abs/1809.02983** (2018)
32. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. CoRR **abs/1805.10180** (2018)
33. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. CoRR **abs/1808.00897** (2018)
34. Zhang, P., Liu, W., Wang, H., Lei, Y., Lu, H.: Deep gated attention networks for large-scale street-level scene segmentation. Pattern Recognit. **88** (2019) 702–714
35. Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., Heng, P., Ni, D.: Deep attentional features for prostate segmentation in ultrasound. In: MICCAI. (2018)
36. Li, C., Tong, Q., Liao, X., Si, W., Sun, Y., Wang, Q., Heng, P.: Attention based hierarchical aggregation network for 3d left atrial segmentation. In: STACOM@MICCAI. (2018)

37. 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** (2019) 197–207
38. Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: MICCAI. (2018)
39. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel squeeze & excitation in fully convolutional networks. CoRR **abs/1803.02579** (2018)
40. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.J., Heinrich, M., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. ArXiv **abs/1804.03999** (2018)