2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433899

# MSSA-NET: MULTI-SCALE SELF-ATTENTION NETWORK FOR BREAST ULTRASOUND IMAGE SEGMENTATION

*Meng Xu*, Kuan Huang, Qiuxiao Chen, Xiaojun Qi*

Department of Computer Science, Utah State University, Logan, UT 84322-4205

## ABSTRACT

Ultrasound imaging is one of the most commonly used diagnostic tools to detect and classify abnormalities of the women breast. Automatic ultrasound image segmentation provides radiologists a second opinion to increase diagnosis accuracy. Deep neural networks have recently been employed to achieve better image segmentation results than conventional approaches. In this paper, we propose a novel deep learning architecture, a Multi-Scale Self-Attention Network (MSSA-Net), which can be trained on small datasets to explore relationships between pixels to achieve better segmentation accuracy. Our MSSA-Net integrates rich local features and global contextual information at different scales and applies self-attention to multi-scale feature maps. We evaluate the proposed MSSA-Net on three public breast ultrasound datasets and compare its performance with six state-of-the-art deep neural network-based approaches in terms of five metrics. MSSA-Net achieves best overall segmentation results and improves the second best approach by 1.21% for Jaccard Index (JI) and 0.94% for Dice's Coefficient (DSC).

***Index Terms*—** breast ultrasound image segmentation, multi-scale self attention, MSSA-Net

## 1. INTRODUCTION

Breast cancer is one of the most common cancers among U.S. women [1]. Estimated 32,5010 new women cases and 42,170 women death cases are reported in U.S. in 2020 [2]. Ultrasound imaging is one of the most common diagnostic tools to detect and classify abnormalities of the women breast. Computer-Aided Diagnosis (CAD) systems using ultrasound images have recently been developed to aid radiologists to increase diagnosis accuracy at the early stage [3]. However, CAD is still challenging due to the lack of accessible data and various ultrasound artifacts. The Breast Ultrasound (BUS) image segmentation, an effective CAD method, has been studied for many years. Traditional BUS image segmentation techniques [4] include thresholding, clustering, watershed, graph method, Active Contour Model (ACM), and Markov Random Field (MRF) method. The first four methods are easy to implement, but are sensitive to initial parameters and employed similarity measures. ACM and MRF methods achieve better and robust BUS segmentation results. However, they are complex and time-consuming [4].

Deep neural networks have recently been widely used in image segmentation due to its superior performance. Here, we briefly review a few representative deep neural networks used for segmentation. U-Net [5] is trained on small datasets to achieve high segmentation accuracy for medical images. Res-UNet [6] uses ResNet [7] as the backbone of U-Net to generate multi-level feature maps in the U-shape network to achieve better segmentation results for medical images. FCN [8] is another segmentation network defining a "skip" architecture to combine shallow and deep features. FCN32s, FCN16s and FCN8s are three networks built using strides of 32, 16, and 8, where FCN8s achieves the best segmentation result for natural images. PSPNet [9] employs a pyramid pooling module to address scene parsing to improve segmentation performance. Deeplabv3+ [10] proposes atrous convolution and atrous spatial pyramid pooling to segment objects. All these deep networks perform well for image segmentation and can be directly deployed in medical image segmentation. However, they simply utilize learned feature maps to segment tumors without considering relationships between pixels. To address this shortcoming, researchers employ self-attention [11] to improve segmentation results by exploring the relationship between pixels and their context. However, it only computes the impact of a pixel on other pixels in one feature map, which is insufficient to represent contextual relationship.

To address the above issues, we propose a novel deep neural network named Multi-Scale Self-Attention Network (MSSA-Net) to achieve segmentation accuracy of 76.05%, 71.90%, and 90.76% on DatasetB [12], Dataset BUSI [13] and Dataset 3 [14], respectively. Our main contributions are: (1) Employing ResNet-101 as the backbone to build MSSA-Net to integrate rich spatial and high-level semantic information via multi-scale features. (2) Designing a novel MSSA mechanism to explore rich contextual relationship among pixels in multi-scale feature maps to boost segmentation performance. (3) Performing extensive experiments on three public datasets by comparing MMSA-Net with six recent deep neural network-based segmentation techniques.

[⋆] Corresponding author.

978-1-6654-1246-9/21/$31.00 ©2021 IEEE

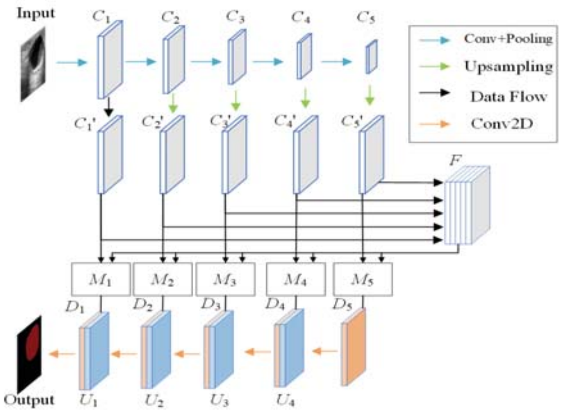

**Fig. 1**. An overview of the proposed MSSA-Net.

## 2. PROPOSED METHOD

The proposed method incorporates the MSSA module in a deep neural network, which uses ResNet-101 as a backbone, to achieve better segmentation results. This MSSA module combines multi-scale features learned by different convolutional blocks to represent the original image at several semantic levels. It integrates both low-level local spatial and high-level semantic contextual information captured in multi-scale features to compute contextual relationship. In this section, we first present the architecture of MSSA-Net and then present the MSSA module.

### 2.1. Overview

The architecture of the proposed MSSA-Net is illustrated in Fig. 1. MSSA-Net uses ResNet-101 as its backbone, which consists of five blocks. We use $C_i$ to denote the output of one of the five blocks of ResNet-101, where integer $i$ corresponds to a block number ranging from 1 to 5. It should be noted that $C_i$ contains feature maps of different scales at different depths, where scales decreases and depth increases with increasing $i$. To integrate both local spatial details and high-level semantics, we employ outputs from five blocks to form a multi-scale feature map $F$. To maintain local spatial details at the highest resolution, we resize each output (e.g., $C_2$, $C_3$, $C_4$, and $C_5$) to a high resolution output with the same dimension as $C_1$ by:

$$C_i' = upsample(C_i)\&\&|C_i'| = |C_1| \tag{1}$$

where $i$ = 2, 3, 4, and 5 and $|x|$ represents the dimension of a feature map $x$ without depth. We then concatenate all resized outputs to construct a multi-scale feature map $F$ by:

$$F = C_1' \oplus C_2' \oplus C_3' \oplus C_4' \oplus C_5' \tag{2}$$

where $\oplus$ represents the concatenation operation. Each high resolution $C_i'$ and the multi-scale feature map $F$ are individually fed into the proposed MSSA module, which will be explained in subsection 2.2, to calculate contextual relationships among pixels and obtain its weighted feature map $D_i$ by:

$$D_i = MSSA(C_i', F) \tag{3}$$

Starting with $D_5$, we convolve it with a 3 × 3 filter and concatenate the filtered result with $D_4$ to integrate spatial and semantic information obtained from blocks 5 and 4. We repeat the same operation to combine spatial and semantic information from blocks 4 and 3, blocks 3 and 2, and blocks 2 and 1. The algorithmic view of chained concatenation operations is as follows:

```
$U_5 = D_5$
for $i$ = 5 to 2 do
    $U_{i-1} = conv(U_i) \oplus D_{i-1}$
end for
```

where $i$ represents the block number and $U_{i-1}$ contains spatial and semantic information from $i^{th}$ and $i-1^{th}$ blocks. A $3 \times 3$ convolution is then applied to $U_1$, followed by bilinear interpolation and softmax to generate the segmentation result.

### 2.2. Multi-Scale Self-Attention

Self-attention methods [11, 15] have been widely used to compute contextual relationships to better represent features learned by convolutional layers. They take a feature map as the input and output a weighted feature map containing contextual relationships. However, this weighted feature map cannot provide sufficient contextual information. Specifically, a feature map learned by shallow layers contains rich local spatial details while missing high-level semantics. A feature map learned by deep layers contains rich high-level semantic information while missing local spatial details.

To address aforementioned shortcomings, we propose a MSSA module to integrate both local spatial and high-level semantic contextual information via multi-scale features learned by different convolutional blocks. The MSSA module takes a multi-scale feature map $F$ and a resized local feature map $C_i'$ as inputs and generates a weighted multi-scale feature map $D_i$ that contains contextual relationships among pixels from local spatial and high-level semantic perspectives.

Fig. 2 illustrates the proposed MSSA model. For the input of a feature map $C_i' \in \mathbb{R}^{H \times W \times Ch_1}$ with $H$, $W$, and $Ch_1$ respectively representing the height, width, and channel dimensions and $i$ representing the block number, we use a $1 \times 1$ convolution to transform $C_i'$ into a new feature map $Y \in \mathbb{R}^{H \times W \times Ch_1/8}$. We use a ratio of $1/8$ to reduce the channel number to its $1/8$ since this ratio has been empirically determined to be optimal [11]. Similarly, for the input of a multi-scale feature map $F \in \mathbb{R}^{H \times W \times Ch_2}$, we use a $1 \times 1$ convolution to generate a new feature map $Z \in \mathbb{R}^{H \times W \times Ch_1/8}$. Since $Ch_2$ is significantly larger than $Ch_1$, we reduce the channel number of $F$ to $Ch_1/8$ to conserve time and memory space and enable matrix computations in the next few steps. We then reshape $Y$ to $Y_r$ of size $(H \times W) \times Ch_1/8$ and reshape and transpose $Z$ to $Z_{rt}$ of size $Ch_1/8 \times (H \times W)$. A multiplication between $Y_r$ and $Z_{rt}$ generates a map of size

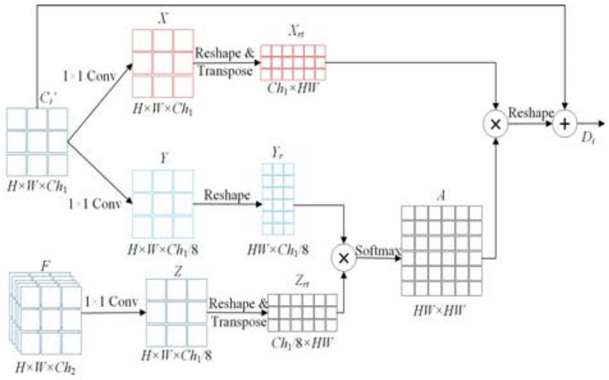

**Fig. 2**. Illustration of the proposed MSSA module.

$(H \times W) \times (H \times W)$. A softmax is performed on this map to generate a normalized map $A$, also called the attention map. In other words, the attention map $A$ is computed by:

$$A(m,n) = \frac{exp(Y_r(m,:) \cdot Z_{rt}(:,n))}{\sum_{n=1}^{H\times W} exp(Y_r(m,:) \cdot Z_{rt}(:,n))} \quad (4)$$

where : is an operator to get all values in a row or a column and $A(m,n)$ represents the impact of the $n^{th}$ column of $Z_{rt}$ on the $m^{th}$ row of $Y_r$. A large value in $A$ indicates a high correlation between $Y_r$ and $Z_{rt}$ (i.e., between $C_i'$ and $F$).

On a second branch, we use another $1 \times 1$ convolution to transform $C_i'$ into a new feature map $X \in \mathbb{R}^{H\times W\times Ch_1}$ and reshape and transpose $X$ to $X_{rt}$ of size $Ch_1 \times (H \times W)$. We then perform a matrix multiplication between $X_{rt}$ and $A$. This result is reshaped to the size $H \times W \times Ch_1$ and multiplied with a learnable parameter $\mu$ to gradually assign appropriate weights to A to generate a weighted attention map as in [11], which is further added to the input $C_i'$ to generate a weighted feature map $D_i \in \mathbb{R}^{H\times W\times Ch_1}$.

$$D_i(m,n) = \mu \times reshape((X_{rt}(m,:) \cdot A(:,n))) + C_i'(m,n) \quad (5)$$

where $D_i(m,n)$ contains the value of a weighted feature map at location $(m,n)$ and $\mu$ is initialized to 0 to allow the network to rely on cues of local neighborhood to maximize learning.

## 3. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed MSSA-Net method by conducting experiments on three public BUS image datasets: Dataset B [12], Dataset BUSI [13] and Dataset 3 [14]. Dataset B has 163 grayscale images of a mean size of $760 \times 570$, where most images contain small tumors. Dataset BUSI contains 780 grayscale images of an average size of $500 \times 500$ for women between 25 and 75 years old. It is the most challenging one among the three datasets since tumors come in different sizes and some tumors have irregular borders. Dataset 3 contains 320 grayscale images of a size of $128 \times 128$ for patients whose ages are in the range of $46.6 \pm 14.2$. In total, there are 1263 BUS images.

We further compare the performance of MSSA-Net with six state-of-the-art deep neural network-based segmentation methods on three aforementioned datasets. The six compared methods are U-Net [5] with ResNet-101 as a backbone [6] (denoted as U-ResNet), U-ResNet with self-attention [11] applied on five blocks (denoted as U-ResNet SA), ResNet-101 [7] by resizing the output of the $5^{th}$ block to the input size, FCN8s [8], PSPNet [9], and Deeplabv3+ [10]. We use five metrics, namely, True Positive Ratio (TPR), False Positive Ratio (FPR), Jaccard Index (JI), Dice's Coefficient (DSC), and Area Error Ratio (AER), to evaluate segmentation results. Specifically, TPR and FPR respectively compute the proportion of correct and incorrect predictions in positive class. JI computes the overlap percentage between the ground truth and the segmentation result. DSC or $F$-measure evaluates the similarity between the ground truth and the segmentation result. AER computes average test errors over all datasets.

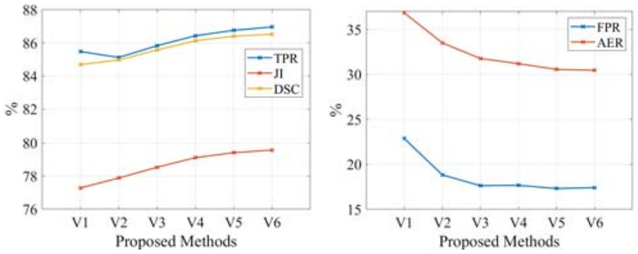

**Fig. 3**. Comparison of MSSA-Net and its variants in terms of five metrics: TPR, JI, and DSC (left); FPR and AER (right).

Fig. 3 compares performance of MSSA-Net and its five variants on three datasets in terms of five measures. MSSA-Net involves combined attention layers $U_5$ through $U_1$ while its variants involve some selected attention layers or no attentions. Five variants of MSSA-Net are as follows: V1 for variant 1 without involving attention layers; V2 for variant 2 involving one attention layer $U_1$; V3 for variant 3 involving combined attention layers $U_2$ and $U_1$; V4 for variant 4 involving combined attention layers $U_3$ through $U_1$; V5 for variant 5 involving combined attention layers $U_4$ through $U_1$; V6 for the proposed MSSA-Net. We compute average values of each metric for three datasets to compare segmentation performance. Specifically, we present TPR, JI, and DSC results in the left plot of Fig. 3 since larger values indicate better segmentation results and present FPR and AER results in the right plot of Fig. 3 since smaller values indicate better segmentation results. It clearly shows that MSSA-Net yields the largest TPR, JI, and DSC values and the smallest FPR and AER values. Variant 1 yields the smallest JI and DSC values and the largest FPR and AER values. With the exception for the TPR metric, JI and DSC values gradually increase and FPR and AER values gradually decrease as more attention layers are employed. In other words, segmentation results gradually improve as more attention layers are employed.

Table 1 summarizes segmentation results of MMSA-Net and six peer methods in terms of five measures on three datasets. MMSA-Net has the highest TPR, JI, and DSC val-

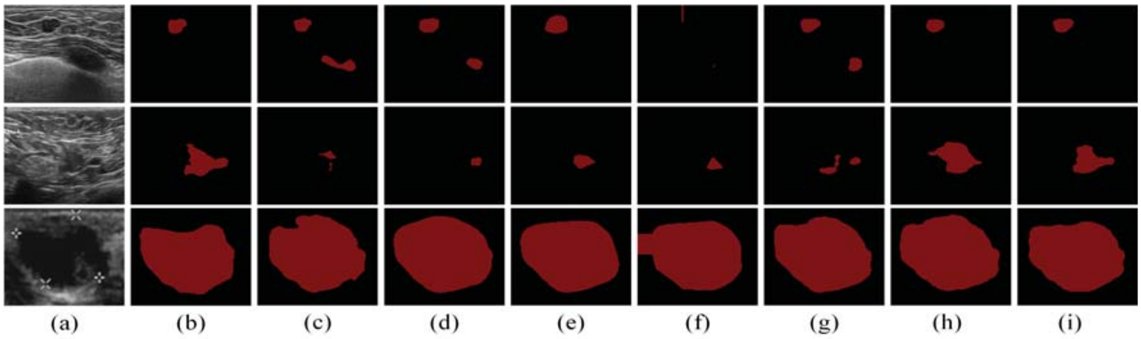

**Fig. 4**. Illustration of segmentation results. (a) BUS images; (b) Ground truth; Segmentation results obtained by (c) Deeplabv3+; (d) PSPNet; (e) FCN8s; (f) ResNet-101; (g) U-ResNet; (h) U-ResNet SA; (i) MSSA-Net.

ues and the lowest FPR and AER values on Dataset BUSI and therefore achieves the best performance. Specifically, it improves the second best method by 2.35%, 1.42%, 1.32%, 2.82%, and 5.67% for TPR, JI, DSC, FPR, and AER, respectively. MMSA-Net achieves the best performance in terms of JI, DSC, FPR, and AER and a comparable TPR for the other two datasets. MMSA-Net also achieves the smallest standard deviation for all five metrics (e.g., 0.21 for TPR, 1.30 for FPR, 0.21 for JI, 0.21 for DSC, and 1.36 for AER). Its variants 4 and 5 outperform six compared methods for Dataset BUSI in five metrics and outperform six compared methods in JI, DSC, FPR, and AER and achieve comparable TPRs for the other two datasets. MSSA-Net and U-ResNet SA respectively have 71,534,626 and 98,374,562 trainable parameters. On average, it takes 0.031 seconds for MSSA-Net and 0.035 seconds for U-ResNet SA to segment an image.

**Table 1**. Summary of tumor segmentation results (%)

| Datasets | Methods | TPR | FPR | JI | DSC | AER |
|---|---|---|---|---|---|---|
| Dataset B[12] | U-ResNet | **85.67** | 24.12 | 74.70 | 82.83 | 38.45 |
| | U-ResNet SA | 84.32 | 24.98 | 74.87 | 82.85 | 40.67 |
| | ResNet-101 | 46.52 | 26.70 | 37.35 | 46.77 | 80.18 |
| | FCN8s | 77.95 | 32.98 | 62.27 | 72.90 | 55.03 |
| | PSPNet | 81.06 | 23.77 | 70.65 | 79.73 | 42.71 |
| | Deeplabv3+ | 63.44 | 76.20 | 50.57 | 60.78 | 112.77 |
| | proposed | 85.63 | **19.48** | **76.05** | **83.78** | **33.85** |
| Dataset BUSI [13] | U-ResNet | 79.20 | 34.82 | 70.59 | 79.03 | 55.63 |
| | U-ResNet SA | 79.02 | 29.80 | 70.89 | 79.60 | 50.78 |
| | ResNet-101 | 53.30 | 36.94 | 44.37 | 54.20 | 83.64 |
| | FCN8s | 78.19 | 44.94 | 64.00 | 74.28 | 66.75 |
| | PSPNet | 78.76 | 33.96 | 69.79 | 78.56 | 55.20 |
| | Deeplabv3+ | 57.34 | 44.51 | 48.12 | 57.98 | 87.18 |
| | proposed | **81.06** | **28.96** | **71.90** | **80.65** | **47.90** |
| Dataset 3[14] | U-ResNet | 94.24 | 4.65 | 90.14 | 94.76 | 10.40 |
| | U-ResNet SA | 94.05 | 4.48 | 90.09 | 94.73 | 10.43 |
| | ResNet-101 | 90.31 | 6.88 | 84.53 | 91.57 | 16.57 |
| | FCN8s | **94.27** | 5.67 | 89.32 | 94.30 | 11.40 |
| | PSPNet | **94.27** | 4.63 | 90.18 | 94.77 | 10.37 |
| | Deeplabv3+ | 91.78 | 4.89 | 87.58 | 93.26 | 13.11 |
| | proposed | 94.18 | **3.80** | **90.76** | **95.14** | **9.62** |

Fig. 4 presents segmentation results of MSSA-Net and six compared methods for one representative BUS image in Dataset B (top row), Dataset BUSI (middle row), and Dataset 3 (bottom row). For the BUS image in Dataset B containing a small tumor and a large tumor-like region with a clear contour, Deeplabv3+, PSPNet, ResNet-101, and U-ResNet mistakenly segment the tumor-like region and ResNet-101 mistakenly segments the tumor region. FCN8s and U-ResNet SA segment a single tumor with a JI value of 63.28% and 72.46%, respectively. MSSA-Net gives a more accurate segmentation result with the highest JI value of 82.17%. For the BUS image in Dataset BUSI containing one irregular tumor without a clear contour, MSSA-Net achieves the highest JI and DSC values of 74.43% and 84.68%, and the lowest AER value of 28.79%. The other six methods fail to segment the tumor since their JI values are less than 55%, DSC values are less than 70%, and AER values are larger than 65%. For the BUS image in dataset 3 containing a big tumor at the center with a clear contour, all methods achieve good segmentation results. MSSA-Net outperforms the other six methods in all five measures and therefore achieves the best segmentation result.

MSSA-Net is implemented by Pytorch. All experiments are conducted on Ubuntu 18.04 system, Intel(R) Xeon(R) CPU E5-2620 2.00 GHz, and two NVIDA GeForce 1080 graphics cards. Input images and ground truths are resized to $128 \times 128$. The Stochastic Gradient Descent (SGD) optimizer utilizes a learning rate of 0.001, a momentum of 0.99, a batch size of 12, and epochs of 100. Cross-entropy is employed in the loss function. To ensure fair comparison, we set these parameters to be the same for all compared methods. We also employ a 10-fold cross-validation to evaluate the performance of all compared methods on three datasets.

## 4. CONCLUSIONS

We propose a novel MSSA-Net for BUS image segmentation. It integrates rich spatial and high-level semantic information via multi-scale feature maps and designs an MSSA mechanism to explore rich contextual relationship among pixels to boost segmentation performance. MSSA-Net outperforms six state-of-the-art deep neural network-based methods in terms of FPR, JI, DSC, and AER and achieves a compare performance in TPR on three public datasets. The training set containing 90% of images in each dataset (around 1146 images in total) is sufficient to train the network to achieve good segmentation results.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [12, 13, 14]. Ethical approval was not required as confirmed by the license attached with the open access data.

## 6. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

## 7. REFERENCES

[1] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 6, pp. 438–451, 2019.

[2] "U.s. breast cancer statistics," `https://www.breastcancer.org/symptoms/understand_bc/statistics`.

[3] E. L. Henriksen, J. F. Carlsen, I. M. Vejborg, M. B. Nielsen, and C. A. Lauridsen, "The efficacy of using computer-aided detection (cad) for detection of breast cancer in mammography screening: a systematic review," *Acta Radiologica*, vol. 60, no. 1, pp. 13–18, 2019.

[4] Q. Huang, Y. Luo, and Q. Zhang, "Breast ultrasound image segmentation: a survey," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 3, pp. 493–507, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[6] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 International Conference on Information Technology in Medicine and Education (ITME)*, 2018, pp. 327–331.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[10] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[11] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[12] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[13] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, pp. 104863, 2020.

[14] Q. Huang, Y. Huang, Y. Luo, F. Yuan, and X. Li, "Segmentation of breast ultrasound image with semantic classification of superpixels," *Medical Image Analysis*, vol. 61, pp. 101657, 2020.

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.