Channel Attention Module With Multiscale Grid Average Pooling for Breast Cancer Segmentation in an Ultrasound Image

Haeyun Lee, *Student Member, IEEE*, Jinhyoung Park, *Member, IEEE*, and Jae Youn Hwang, *Member, IEEE*,

***Abstract*— Breast cancer accounts for the second-largest number of deaths in women around the world, and more than 8% of women will suffer from the disease in their lifetime. Mortality due to breast cancer can be reduced by its early and precise diagnosis. Many studies have investigated methods for segmentation, and computer-aided diagnosis based on deep learning techniques, in particular, has recently gained attention. However, recently proposed methods such as fully convolutional network (FCN), SegNet, and U-Net still need to be further improved to provide better semantic segmentation when diagnosing breast cancer by ultrasound imaging, because of their low performance. In this article, we propose a channel attention module with multiscale grid average pooling (MSGRAP) for the precise segmentation of breast cancer regions in ultrasound images. We demonstrate the effectiveness of the channel attention module with MSGRAP for semantic segmentation and develop a novel semantic segmentation network with the proposed attention module for the precise segmentation of breast cancer regions in ultrasound images. While a conventional convolutional operation cannot use global spatial information on input images and only use the small local information in a kernel of a convolution filter, the proposed attention module allows using both global and local spatial information. In addition, through ablation studies, we come up with a network architecture for precise breast cancer segmentation in an ultrasound image. The proposed network was constructed with an open-source breast cancer ultrasound image data set, and its performance was compared with those of other state-of-the-art deep-learning models for the segmentation of breast cancer. The experimental results showed that our network outperformed other segmentation methods, and the proposed channel attention module improved the performance of the network for breast cancer segmentation in ultrasound images.**

***Index Terms*— Breast cancer, deep learning, semantic segmentation, ultrasound image.**

I. Introduction

BREAST cancer is one of the most commonly diagnosed cancers among women. It leads to the second-largest number of deaths in women around the world, and more than 8% of women will suffer from breast cancer in the course of their lifetime [1]. Unfortunately, what causes breast cancer is not yet known. To reduce the mortality of breast cancer, the early detection of breast cancer is needed [2]. Digital mammography is primarily used for the diagnosis of breast cancer. However, the technique has inherent limitations. It has a high misdiagnosis rate due to surrounding dense tissues, which have attenuation coefficients similar to breast tumors, and increases the risk of radiation exposure to patients. In many previous studies that related to the diagnosis of breast cancer, breast cancer detection rates have been significantly improved by the addition of mammary ultrasonography [3]. Ultrasound imaging methods have been also used as a safer alternative to digital mammography for the initial diagnosis of breast cancer [4].

However, diagnosing breast cancer using ultrasound images relies on the experience of the radiologist, who must interpret speckle noises and image complexities in the ultrasound image. Previous studies have shown that a computer-aided diagnosis (CAD) system with high sensitivity and specificity can be utilized as a diagnostic aid by radiologists and produce better clinical evaluations [5]. This reduces the dependence on an ultrasound image examiner for the diagnosis of breast cancer [6].

In the past several decades, a variety of ultrasound image analysis approaches have been proposed for diagnosing breast cancer, ranging from simple filtering to sophisticated learning-based approaches. Drukker *et al.* [7] proposed radial gradient index filtering approaches for locating and classifying breast tumor regions. Yap *et al.* [8] proposed a new approach for the detection of breast lesions using multifractal processing techniques and hybrid filtering. A level-set-based segmentation approach, which combined both global statistical and local

Manuscript received January 6, 2020; accepted February 5, 2020. Date of publication February 10, 2020; date of current version June 29, 2020. This work was supported in part by the National Research Foundation of Korea (fNRF) under Grant NRF-2017R1A2B4010726, in part by NRF through the Bio and Medical Technology Development Program under Grant NRF-2017M3A9G8084463, and in part by the Industrial Strategic Technology Development Program under Grant 10085624. The work of Jae Youn Hwang was supported by the Ministry of Trade, Industry, and Energy of Korea. *(Corresponding author: Jae Youn Hwang.)*

Haeyun Lee and Jae Youn Hwang are with the Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, South Korea (e-mail: haeyun@dgist.ac.kr; jyhwang@dgist.ac.kr).

Jinhyoung Park is with the Department of Biomedical Engineering, Sungkyunkwan University, Suwon 2066, South Korea (e-mail: jin.park@skku.edu).

Digital Object Identifier 10.1109/TUFFC.2020.2972573

edge information, was proposed [9]. A fully automated breast cancer segmentation method that considered both texture and spatial features was also proposed for the diagnosis of breast cancer [10]. Also, an active contour-based breast cancer segmentation [11] and a 3-D graph-based segmentation method [12] for ultrasound images were proposed. More recently, a semi-automatic technique based on a curvilinear blind-ended method was proposed for the 3-D segmentation of a left atrial appendage in an ultrasound image [13]. A segmentation method via semantic classification of superpixels was developed for the segmentation of breast cancer in an ultrasound image [14]. However, these traditional methods have shown fairly low performance in the segmentation of breast cancer in an ultrasound image [15].

Recently, the development of various deep learning-based approaches in the computer vision field has resulted in significant improvements over conventional techniques. There have also been numerous attempts to apply deep learning techniques to medical images. Traditional semantic segmentation networks such as fully convolutional network (FCN) [16], SegNet [17], and U-Net [18] were primarily used to detect breast cancer [19]. Also, the latest semantic segmentation network, Pyramid Scence Parsing Network (PSPNet) [20], which consists of Resnet and pyramid pooling modules, was applied to the diagnosis of breast cancer [21]. In addition, Xu *et al.* [22] proposed deep learning networks for the 3-D segmentation of breast tumors in ultrasound images. Although they have performed well in many medical imaging fields and have been widely used in other areas, they are still limited when it comes to the precise segmentation of breast cancer in an ultrasound image. Since these architectures only consist of simple convolution filters, only local information within receptive fields is used, and global information on an input image is lost during the process of architectures. Although the receptive fields of Resnet [23] and other Resnet-based networks are theoretically larger than an input image, Zhou *et al.* [24] demonstrated that the actual receptive field is much smaller than the theoretical one, especially on high-level feature levels.

In order to solve these issues, several methods for the utilization of global information have been studied. Channel-wise attention is computed from the global information obtained by global average pooling and is multiplied to each channel in order to recalibrate different channels in feature maps, so that informative features can be selectively used for more accurate image classifications [25]. Unlike image classification, semantic segmentation involves labeling each pixel in an image, and therefore, not only global information but also local content information is important. However, because existing channel attention modules only use global information, the information on different feature distributions over various image regions can be lost by the conventional modules. Note that local information as well as global information is needed to enhance the performance of convolutional neural networks (CNNs) for the semantic segmentation of breast cancer in ultrasound images.

In this article, we propose a novel channel attention module with multiscale grid average pooling (MSGRAP) that is capable of improving the performance of CNNs for breast cancer segmentation in an ultrasound image. Then, we demonstrate the capability of a visual geometry group network (VGGNet)-based deep learning architecture using the proposed channel attention module [26] for precise breast cancer segmentation. Note that precise breast cancer segmentation in an ultrasound image is highly important to determine treatment plans or lymph node metastasis. In Section II, we first apply the channel attention mechanism proposed by Hu *et al.* [25] to the proposed network. This mechanism has the advantage of being able to use global information on an input image, and therefore, it can resolve the issues of conventional breast cancer segmentation approaches. After that, we demonstrate that a channel attention module using grid average pooling (GRAP) instead of global average pooling (GAP) can improve the segmentation performance by using the local information of an input image. Finally, we show that the proposed channel attention module with MSGRAP outperforms the two channel attention modules mentioned above. To evaluate the proposed deep learning network, its performance was compared with those of other semantic segmentation techniques including FCN, U-Net, SegNet, PSPNet-18, and Context Encoding Network (ENCNet)-18 for segmenting breast cancer regions in terms of global accuracy, F1 score, sensitivity, specificity, a false positive rate (FPR), precision, and so on. The results demonstrate that the deep-learning structure based on the channel attention module with MSGRAP offers better performance than the classical methods [25] for breast cancer segmentation tasks.

II. Methods

In this section, we introduce a channel attention module proposed by Hu *et al.* [25]. After that, we propose a channel attention module with MSGRAP, which is revised from the channel attention module based on various grounds, and a new breast cancer segmentation network with the proposed module. In addition, we evaluate the capability of the new attention module for semantic segmentation using an ablation study.

A. Channel Attention Module With GAP

Many studies have shown that attention is one of the most important factors in human perception [27], [28]. Recently, attention mechanisms have become an integral part of powerful sequence models and transformation models for a variety of tasks [29]. In particular, channel attention modules have been used in various computer vision applications to improve the performance of CNNs [25], [30].

Recent deep learning approaches for image classification and semantic segmentation [16], [18], [23], [26], [31] consist of continuous simple convolution operations. Therefore, CNN-based networks merely treat the local parts of an image, thus resulting in the loss of global information. To resolve this issue, we applied the channel attention module proposed by Hu *et al.* [25] as a preliminary step to our proposed network.

As shown in Fig. 1, channel-wise statistical information can be obtained by using GAP. For each channel of an input feature map x_c, the statistical information of a channel z_c is obtained by the following equation:

$$z_c = \mathrm{GAP}(x_c) = 1/(H \times W) \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

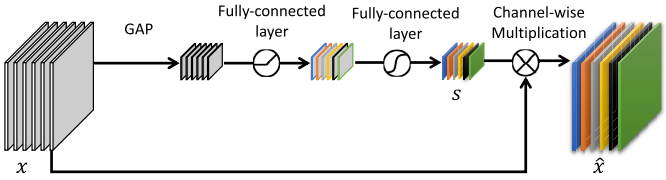

Fig. 1. Channel attention module with GAP.

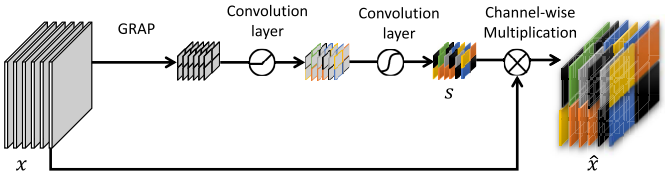

Fig. 2. Channel attention module with GRAP.

where H and W are the height and width of the input feature map, whereas i and j are the spatial coordinates of the input feature map, respectively. Note that the channel attention module with GAP is identical to the squeeze-and-excitation (SE) block shown in the previous study [25].

The obtained information z can be used to calculate dependencies between channels with two fully connected layers and two nonlinear activation functions, ReLU and Sigmoid

$$s = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where σ and δ are a sigmoid and an ReLU nonlinear activation function, respectively, and $W_1 \in \mathbb{R}^{C\times C/r}$ and $W_2 \in \mathbb{R}^{C/r\times C}$ are the weights of fully connected layers. We use a reduction ratio r to reduce the number of nodes in W_1 and increase the number of feature maps by C in W_2. In the previous study [25], when r was 16, there was a little drop in performance, but the number of parameters could be reduced. Thus, here, we use 16 as the channel reduction factor. The scaling factor s, obtained using the above equation, is multiplied by the input feature map for each channel

$$\hat{x_c} = s_c \times x_c \tag{3}$$

where s_c and $\hat{x_c}$ are the c-th channel of scaling factor s and the rescaled feature map. The channel attention module is utilized to maintain global information in our VGG-Net based model for semantic segmentation.

B. Channel Attention Module With GRAP

The channel attention module introduced in Section II-A has been shown to improve image classification and detection because it is determined using the entire image content [25]. However, although the channel attention module performs well for image classification and detection, it demonstrates limited performance in semantic segmentation, which predicts labels for all pixels and requires a complete understanding of the scene. In particular, when a single image has various types of contents in different local areas, the channel attention module cannot characterize the locally different natures of the image because of the GAP operation.

Inspired by this observation, GRAP was employed in the channel attention module instead of GAP. As shown in Fig. 2, channel-wise statistical information can be achieved by using $k \times k$ GRAP. In the GRAP, the input feature map is first divided into $k \times k$ grid cells with a pixel size of H/k and W/k, and the pixels for each partitioned grid cell are averaged. For each channel of an input feature map x_c, we can obtain statistical information for the channel z_c

$$z_c(i,j) = \mathrm{GRAP}(x_c) = \frac{(k\times k)}{(H\times W)} \sum_{a=(H/k)\times i}^{(H/k)\times(i+1)} \sum_{b=(W/k)\times j}^{(W/k)\times(j+1)} x_c(a,b) \tag{4}$$

where x_c is the c-th channel of an input feature volume x, k is the grid size of average pooling, a is the x-coordinate of an input feature map, b is the y-coordinate of the input feature map, i is the x-coordinate of a resultant image after the GRAP, j is the y-coordinate of the resultant image, H is the height of the feature map, W is the width of the feature map, and $H/k \times W/k$ is the pixel size of each grid cell. The obtained information can be used to calculate the dependencies between channels with two 1×1 convolution layers and two nonlinear activation functions, ReLU and sigmoid

$$s = \sigma(W_2\delta(W_1 z)) \tag{5}$$

where σ and δ are a sigmoid and an ReLU nonlinear activation function, respectively, and $W_1 \in \mathbb{R}^{C\times C/r}$ and $W_2 \in \mathbb{R}^{C/r\times C}$ are the weights of the 1×1 convolution layers. The fully connected layers are used in the channel attention module proposed by Hu *et al.* [25], However, here, 1×1 convolution layers are used in the proposed channel attention module because z_c can be obtained by using GRAP. The scaling factor s, obtained from the above equation, is multiplied by the input feature map for each channel ($r = 16$)

$$\hat{x_c} = U(s_c) \times x_c \tag{6}$$

where s_c is the c-th channel of a scaling factor s, $\hat{x_c}$ is the c-th channel of the rescaled feature map, $\times$ is a pixel-wise multiplication, and U is an upsampling function. Unlike the channel attention module, we used $k \times k$ GRAP and, thus, obtained an attention information map with a spatial size of $k \times k$. For the sake of convenient calculation, an upsampling function was used to align the spatial size of the attention information map with that of the input feature map. A bilinear function is here used for upsampling because grid patterns appear in the resultant images when the nearest neighborhood function is applied to the upsampling. This channel attention module in our model allows us to employ available local information. We empirically determined the grid size as 10×10, which offered the best performance compared with others by using the Grid Search, which is a conventional hyperparameter tuning method.

C. Channel Attention Module With MSGRAP

The network architecture of high-level vision tasks such as image classification and semantic segmentation mostly includes pooling layers such as max pooling or average

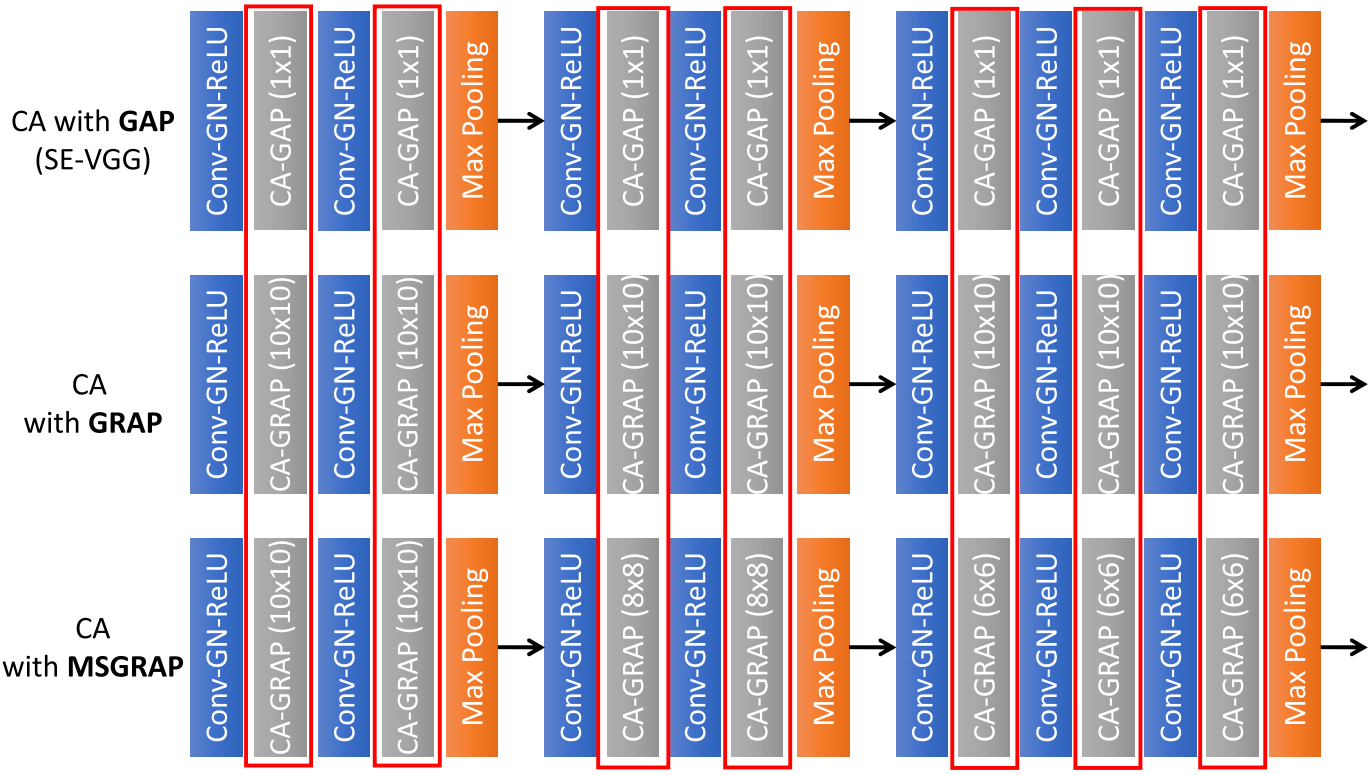

Fig. 3. Illustration of three encoder architectures based on visual geometry group (VGG) Net [26]. VGG-CA was proposed in [25], and VGG-CA with GRAP and VGG-CA with MSGRAP are proposed in this study.

pooling. After passing over a pooling layer, a feature map has a smaller height and width than the previous feature map. When splitting a high-level feature map into a grid using a low-level feature map, the size of the feature map becomes similar to or less than the size of the connection kernel. In addition, the feature map that has passed over several pooling layers includes high-level information such as contextual information. Splitting a feature map including high-level information can interfere with training and test tasks. For the above-mentioned reasons, a larger sized grid should be used to obtain channel attention weights for a feature map passing over pooling layers. The grid size of channel attention with GRAP should be determined individually, thus depending on the size of the feature map. Therefore, we devised a channel attention module using MSGRAP. In Section II-D, we demonstrate and compare the proposed network architecture with three channel attention modules such as a channel attention module, a channel attention module with GRAP, and a channel attention module with MSGRAP.

In the attention module, we utilize 10×10 GRAP before the first pooling layer in an encoder. Whenever the poling layer is passed, the grid size is reduced by two. Fig. 3 shows three encoders used in the networks. The first row in Fig. 3 is some parts of encoder architecture with the channel attention module mentioned in Section II-A. This architecture was first proposed in [25] and was called SE-VGG. The architectures in the second and third rows in Fig. 3 illustrate the networks, which have the channel attention modules with GRAP and multiscale average pooling, as mentioned in Sections II-B and II-C, respectively.

TABLE I
COMPARISON OF BATCH NORMALIZATION AND GROUP NORMALIZATION WITH U-NET IN TERMS OF F1 SCORE. THE BEST PERFORMANCE IS IN BOLD

Concatenation connection	U-Net with BN	U-Net with GN
F1 score	0.7883	**0.7969**

D. Architecture of the Proposed Network

In this section, we describe our proposed end-to-end network architecture. Fig. 4 shows the architecture of our breast cancer segmentation network. The network receives a breast ultrasound image as an input and predicts its semantic segmentation result.

As shown in Fig. 4, our network consists of two parts: an encoder and a decoder. The number above each block represents the number of feature maps. Here, we use all convolution layers with C' filters with a size of $3 \times 3 \times C$, in which C and C' are the previous and current number of feature maps, respectively, except for the final convolution layer. For the final convolution layer, we use a convolution layer with two filters with a size of $3 \times 3 \times 64$. To obtain the final segmentation results, we use an argmax function with a threshold value of 0.5.

The encoder in our network is based on the VGGNet architecture [26] except for the batch normalization [32] and channel attention modules. In the previous study [33], the batch normalization is highly influenced by a batch size:

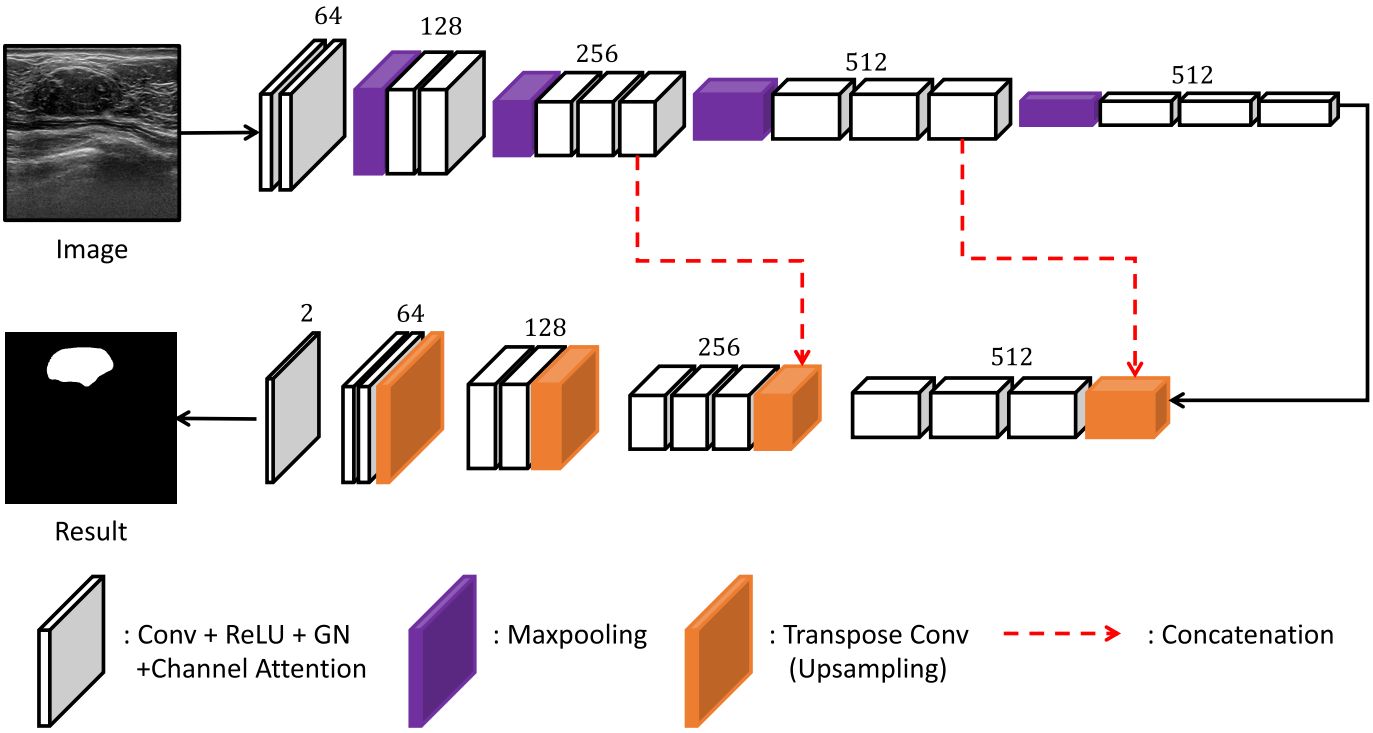

Fig. 4. Illustration of the proposed network architecture.

TABLE II
COMPARISON OF U-NET-LIKE AND OUR ARCHITECTURE IN TERMS OF F1 SCORE. THE BEST PERFORMANCE IS IN BOLD

Concatenation connection	Unet-like architecture	Ours architecture
F1 score	0.7845	**0.7958**

the smaller the batch size, the lower the performance is. In other words, a small batch size reduces the generalization ability of deep neural networks. Since the data set used in the experiments had an enough spatial size to use a smaller batch size, we, therefore, use group normalization which has little effects on the batch size. The group normalization is a technology that divides each channel into N groups and normalizes the features within each group regardless of the batch size [33]. Therefore, it does not depend on the batch size and can overcome the generalization issues caused by the small batch size when the network is trained with large input images.

We additionally examined a U-Net network to demonstrate the effect of group normalization with two of the tenfold data sets. As shown in Table I, the U-Net with group normalization shows a higher performance than the U-Net with batch normalization.

A symmetrical encoder architecture was built in the decoder. To upsample the feature maps, we utilize a 4×4 transpose convolution with a stride of 2. Unlike the U-Net architecture, the two feature maps from the encoder are only connected to the decoder. By performing an additional experiment, we demonstrated that the network architecture, with two feature maps connected between the encoder and decoder, performed better than the U-Net-like architectures. For this reason, it is better not to use low-level features in the network because most ultrasound images are noisy. We also conduct the ablation study for architecture with two of the tenfold data sets. Table II shows the additional experiment results, thus demonstrating that our architecture resulted in better performance than the U-Net-like architecture in the semantic segmentation using just two feature map connections. Since this task only segments breast cancer into binary classes, the final convolution layer has two filters with a size of $3 \times 3 \times 64$.

The three channel attention modules presented in Section II-C were applied to this model and compared with the performance of three networks. The network with the channel attention module proposed by Hu *et al.* [25] is named as Ours-GAP. On the other hand, the networks, which used channel attention modules with GRAP and MSGRAP, are denoted by Ours-GRAP and Ours-MSGRAP, respectively.

III. EXPERIMENTS

In this section, we describe the experimental setting and the data set used to evaluate the performance of our networks.

A. Data Set

To evaluate the performance of the proposed network, we used a data set of breast cancer ultrasound images obtained from the previous study [19]. The equipment used to obtain the data set was the Siemens' ACUSON Sequoia C512 system and a 17L5 HD linear array transducer at 8.5 MHz. These data [19] were obtained from 163 different women with the

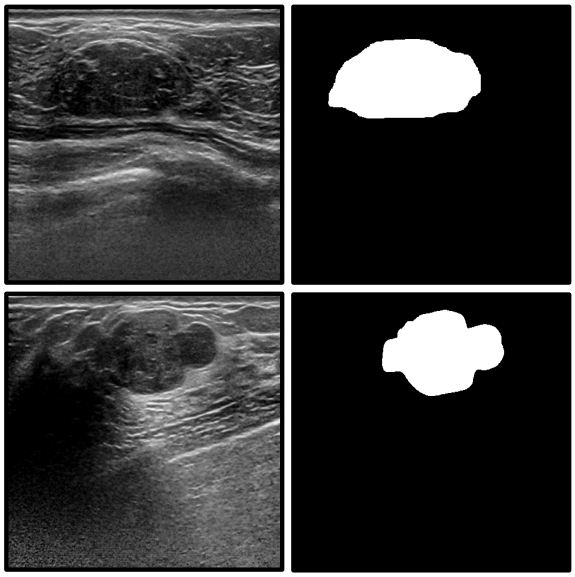

Fig. 5. Representative images of a breast ultrasound cancer data set [19]. Ultrasound image (left). Ground truth (right).

abovementioned equipment, consisting of 53 cancerous masses and 110 benigns, with an average image size of 454 × 537 pixels. There are 40 cases of invasive ductal carcinomas, two cases of lobular carcinomas, four cases of coronary carcinoma *in situ*, and seven cases of unspecified malignant lesions in the cancerous mass images. There are 39 cases of fibroadenomas, 65 cases of unspecified cysts, and six cases of other types in the benign images. For this data set, the lesions were delineated by experienced radiologists. Fig. 5 shows an example of this data set.

B. Performance Metric

For quantitative comparisons of our model with other methods, we used several performance metrics such as global accuracy, F1 score, sensitivity, and specificity.

Global accuracy is the most basic metric for several computer vision tasks, such as image detection and segmentation. The formula for global accuracy is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{7}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative, respectively.

F1 score is a good metric for the imbalanced data. Since there was a small portion of cancer among all the breast ultrasound images, this data set can also be considered as imbalanced data. Thus, we chose the F1 score as a performance metric. The formula for the F1 score is as follows:

$$F1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{8}$$

where TP, FP, and FN are the true positive, false positive, and false negative, respectively.

The data set, which used in this study, consists of 5% of cancer pixels and 95% of normal pixels. Since this data set was imbalanced, we used other metrics, such as FPR, precision, an intersection over union (IoU), and area under the curve (AUC) of the precision and recall (PR) for a fair evaluation. The AUC metrics used a sweep of the threshold from $p = 0$ to $p = 1$, as opposed to the $p = 0.5$ (argmax) used for the remaining non-AUC metrics. FPR represents the number of false positive over the one of the condition negatives. The formula for the IoU, which is one of the metrics commonly used in the semantic segmentation, is as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{9}$$

Sensitivity (recall) and specificity were also used as performance metrics since they have been widely used in medical fields.

C. Training Strategy

To evaluate the performance of the proposed networks, we trained different models, such as FCN, U-Net, SegNet, and PSPNet-18, and then compared their performance.

Specifically, given a training data set $D = \{\ldots, (I^{(i)}, \text{GT}^{(i)}), \ldots\}$, where $I^{(i)}$ and $\text{GT}^{(i)}$ are the ith ultrasound image and its corresponding ground-truth semantic label, respectively. The following loss function is minimized:

$$L(\Theta, k; D) = -\sum_{c=0}^{M-1} \text{GT}_c \log(f(I; \Theta)_c) \tag{10}$$

where Θ is a set of network parameters, c is an indicator of a class, M is the number of classes, $f(I; \Theta)$ is the result from networks with parameters Θ before applying the argmax, and $f(I; \Theta)_c$ and GT_c are the predicted probability and the binary indicator for the class, c, respectively. Since the breast cancer segmentation is a binary classification for each pixel, we use M as 2.

The number of breast ultrasound cancer data sets provided by Yap *et al.* [19] was somewhat limited, and therefore, we have configured the training and testing processes as tenfold cross-validation. When performing the tenfold cross-validation, we divided the data into 146 or 147 breast cancer ultrasound images for training and 16 or 17 breast cancer ultrasound images for testing in each validation step. In addition, we augmented each patch by random horizontal and vertical flips and random 90° rotations. For training and testing, we set all the images sizes to an average of 454×537 pixels.

For training, the weights of all convolution layers were initialized by the Kaiming initialization. An Adam optimization method with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ was used. The learning rate was 10^{-3} and was reduced by half every 30 epochs. The mini-batch size was eight. The models were trained for 120 epochs. We used PyTorch [34] to implement and train our networks. It took four days to train the models using an Intel Zeon E5-2620 at 2.0 GHz and an NVIDIA TITAN RTX (24 GB).

IV. Results and Discussion

In this section, we evaluate the performance of our networks, Ours-GAP, Ours-GRAP, and Ours-MSGRAP. For the

TABLE III
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS. THE FIRST AND THE SECOND BEST PERFORMANCE ARE IN **BOLD** AND UNDERLINED, RESPECTIVELY

AUMethods	Global Acc.	F1 score	Sensitivity(Recall)	Specificity	FPR	Precision	IoU	AUC(PR)	AUC(ROC)
FCN [16]	97.384	0.7123	0.7702	0.9834	0.0166	0.6907	0.5627	0.7812	0.9634
U-Net [18]	97.435	0.7132	0.7846	0.9827	0.0173	0.6696	0.5613	0.7579	0.9332
SegNet [17]	97.576	0.7225	0.8006	0.9802	0.0198	0.6877	0.6001	0.7952	0.9604
PSPNet-18 [20]	97.736	0.7520	**0.8088**	0.9847	0.0153	0.7058	0.6058	0.8047	0.9510
ENCNet-18 [35]	97.597	0.7266	0.7990	0.9834	0.0166	0.6859	0.5770	0.7511	**0.9647**
Ours-GAP	97.462	0.7205	0.7638	0.9840	0.0160	0.6983	0.5993	0.7818	0.9490
Ours-GRAP	97.632	0.7445	0.7769	0.9852	0.0148	0.7276	0.6185	0.8084	0.9551
Ours-MSGRAP	**97.794**	**0.7658**	0.8041	**0.9866**	**0.0134**	**0.7459**	**0.6226**	**0.8149**	0.9606

evaluation, we used the breast cancer ultrasound image data set mentioned in Section III-A. We compared our networks with several state-of-the-art semantic segmentation networks, such as FCN [16], U-Net [18], SegNet [17], and PSPNet [20], and ENCNet [35] with ResNet-18. Semantic segmentation networks based on ResNet-34, 51 [23] and DenseNet [31], which have more than 18 convolution layers, were excluded for fair comparisons because they used many more parameters than ours.

A. Quantitative Results

Table III shows quantitative comparisons in terms of global accuracy, F1 score, sensitivity, and specificity. Ours-MSGRAP exhibited a better F1 score than the other models. The F1 score for ours-MSGRAP was 0.7658, while those of FCN, U-Net, SegNet, PSPNet-18, and ENCNet-18 were 0.7123, 0.7132, 0.7225, 0.7520, and 0.7266, respectively. The F1 score for Ours-MSGRAP was 7.47% higher than that of FCN, whereas it was 1.80% higher than that of PSPNet-18. Also, Ours-MSGRAP showed higher performance than other models in terms of global accuracy, specificity, FPR, precision, and IoU.

In addition, we compared the PR and receiver operating characteristic (ROC) curves of the models for quantitative comparisons of the performance of the models (Fig. 6). In the previous study, a PR curve is more suited for an imbalanced data set than an ROC curve [36]. Therefore, we utilized the PR curves of the models because the distribution of the data set utilized for the models was imbalanced. Note that the ratio of the tumor and normal pixels was 0.05 : 0.95. In Table III, the AUC value of the PR curve of Ours-MSGRAP was 0.8147, which was the highest one among those of other models. In contrast, the AUC value of the ROC curve of Ours-MSGRAP was 0.9606, which was somewhat lower than those of FCN and ENCNet. Note that the PR curve is more informative than the ROC curve for such an imbalanced data set. Although FCN and ENCNet offered the higher AUC values of ROCs than Ours-MSGRAP, Ours-MSGRAP offered the highest AUC of the PR curve, global accuracy, F1 score, specificity, FPR, precision, and IoU value compared to other models.

It was here shown that Our-MSGRAP outperformed Ours-GAP and Ours-GRAP in terms of all performance metrics, validating the effectiveness of the channel attention module with MSGRAP. Also, Ours-GAP and Ours-GRAP outperformed FCN, SegNet, and U-Net in most cases. This showed that the channel attention module was effective for breast cancer segmentation. Specifically, the channel attention module with adaptive average pooling showed better performance than a conventional channel attention module.

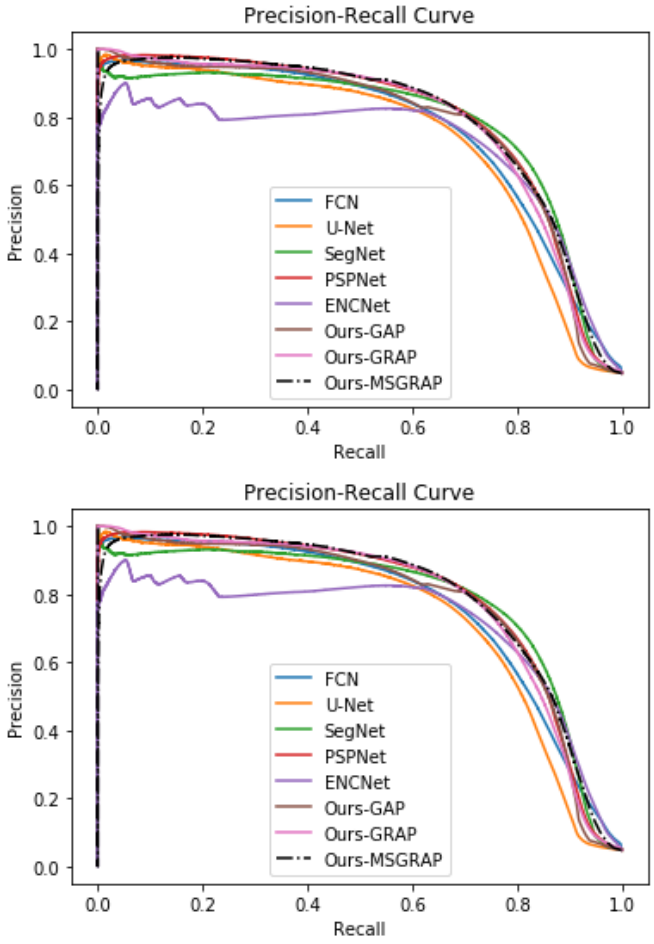

Fig. 6. PR and ROC curves of the models.

B. Qualitative Results

Fig. 7 shows a qualitative comparison of our models with state-of-the-art methods. Fig. 7 shows that Our-MSGRAP produces more accurate results than other methods. As shown in the first column to the fourth column and last column of Fig. 7, all methods except for Our-MSGRAP segmented normal regions as breast tumor regions. From these results, it is shown that Ours-MSGRAP provides better sensitivity.

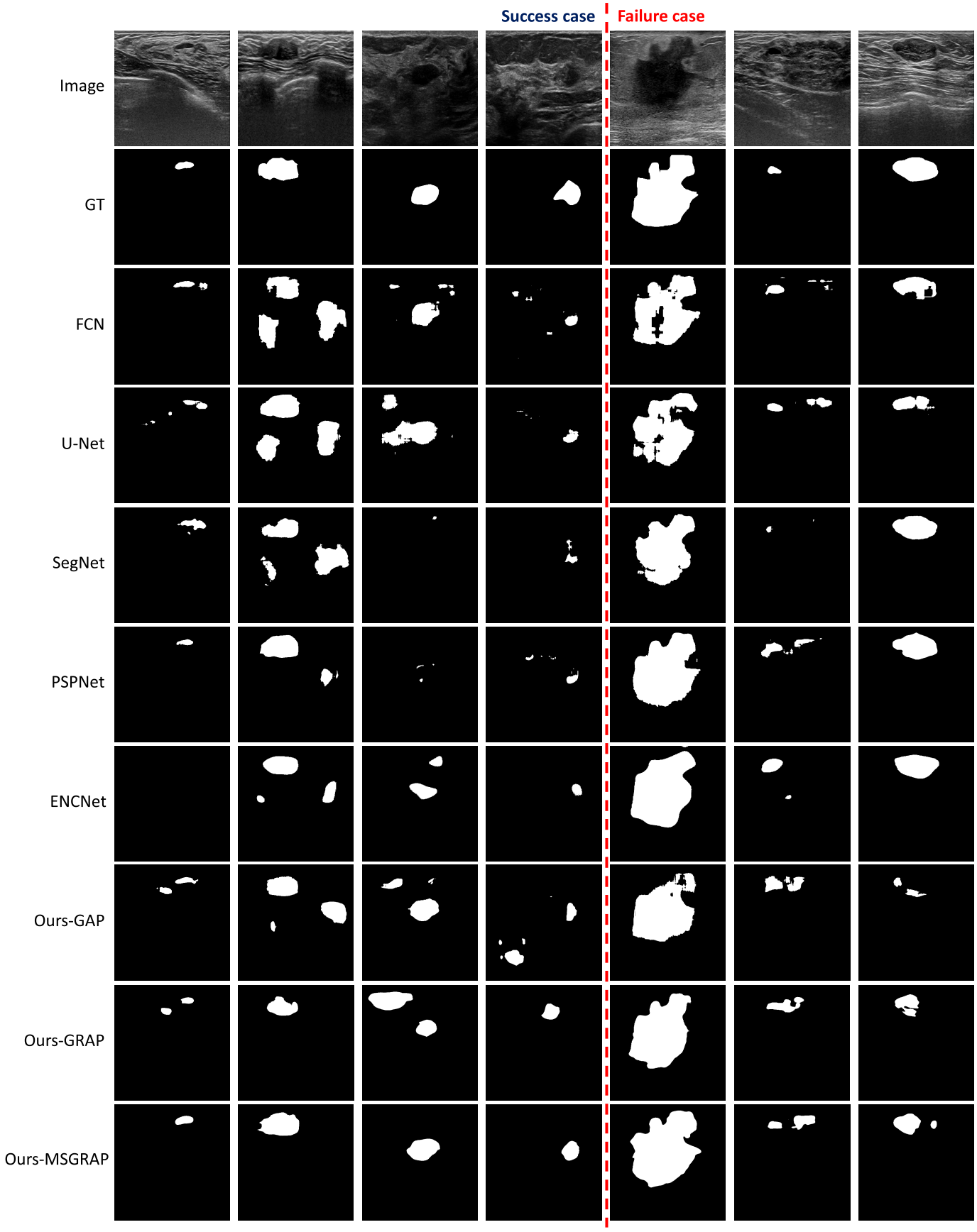

Fig. 7. Qualitative comparisons of the segmented results obtained by using different models such as FCN, U-Net, SegNet, PSPNet-18, ENCNet-18, and our models. GT: ground truth.

Especially, in the second and last columns, there are few false positives in the results of Ours-GRAP and Ours-MSGRAP compared to other methods. These two methods can make good use of local information by using GRAP. In addition, in the fifth column of Fig. 7., the failure cases of our model are shown. We determined the failure cases when our models exhibited similar or inferior outcomes compared with those of other models (Fig. 7). The rate of the failure case was found to be 14.1%.

Both quantitative and qualitative comparisons showed that Our-MSGRAP offered not only better outcomes in the segmentation of breast tumors in breast ultrasound images as shown in Fig. 7 but also higher F1-score than other methods. These results suggested that our method might have the potentials to be an alternative suited for a CAD system.

V. Conclusion

In this article, we proposed a novel architecture including a channel attention module with MSGRAP for the semantic segmentation of breast tumors in an ultrasound image. The results demonstrated that Ours-MSGRAP outperformed other methods such as FCN, U-Net, SegNet, PSPNet-18, and ENCNet-18 in terms of the global accuracy, F1-score, specificity, precision, IoU, and AUC value of a PR curve. The channel attention module with GAP could only maintain global information. In contrast, the channel attention module with MSGRAP allowed maintaining local and global information for the semantic segmentation of breast tumors in ultrasound images. Note that both the local and global information are crucial in the segmentation of ultrasound images due to speckle noises and low contrast of ultrasound images. MSGRAP here offered better performance than other channel attention modules as well as the channel attention module with GAP (SE block) in the breast cancer segmentation. In the previous study, the SE block was applied to VGGNet, denoted by SENet. Ours-MSGRAP was inspired by the SENet. However, the SENet is a classification model. Therefore, it could not be directly applied to semantic segmentation. To utilize the SENet architecture for the semantic segmentation in the proposed network, fully connected layers of the SENet were replaced with the decoders. Altogether, these results suggest that the proposed method has the potential to be implemented in a CAD system for segmenting breast cancer in ultrasound images.

Furthermore, the use of the proposed channel attention module with MSGRAP is not limited to the ultrasound image segmentation. It can also be integrated with other networks for semantic segmentation in different domains as the SE block. Also, the encoder architecture, VGGNet applying a channel attention module with MSGRAP, may be suited not only for semantic segmentation tasks but also for other high-level vision tasks. The associated investigation remains as a future study.

References

[1] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[2] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit.*, vol. 39, no. 4, pp. 646–668, Apr. 2006.

[3] W. A. Berg *et al.*, "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer," *Jama*, vol. 299, no. 18, pp. 2151–2163, 2008.

[4] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[5] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors," *Breast Cancer Res. Treat.*, vol. 89, no. 2, pp. 179–185, Jan. 2005.

[6] K. Drukker, N. P. Gruszauskas, C. A. Sennett, and M. L. Giger, "Breast US computer-aided diagnosis workstation: Performance with a large clinical diagnostic population," *Radiology*, vol. 248, no. 2, pp. 392–397, Aug. 2008.

[7] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Med. Phys.*, vol. 29, no. 7, pp. 1438–1446, Jun. 2002.

[8] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A novel algorithm for initial lesion detection in ultrasound breast images," *J. Appl. Clin. Med. Phys.*, vol. 9, no. 4, pp. 181–199, Sep. 2008.

[9] B. Liu, H. D. Cheng, J. Huang, J. Tian, X. Tang, and J. Liu, "Probability density difference-based active contour for ultrasound image segmentation," *Pattern Recognit.*, vol. 43, no. 6, pp. 2028–2042, Jun. 2010.

[10] J. Shan, H. D. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound Med. Biol.*, vol. 38, no. 2, pp. 262–275, Feb. 2012.

[11] R. Rodrigues, R. Braz, M. Pereira, J. Moutinho, and A. M. G. Pinheiro, "A two-step segmentation method for breast ultrasound masses based on multi-resolution analysis," *Ultrasound Med. Biol.*, vol. 41, no. 6, pp. 1737–1748, Jun. 2015.

[12] H. Chang, Z. Chen, Q. Huang, J. Shi, and X. Li, "Graph-based learning for segmentation of 3D ultrasound images," *Neurocomputing*, vol. 151, pp. 632–644, Mar. 2015.

[13] P. Morais *et al.*, "Fast segmentation of the left atrial appendage in 3-D transesophageal echocardiographic images," *IEEE Trans. Ultrason., Ferroelectr., Freq. Contr.*, vol. 65, no. 12, pp. 2332–2342, Dec. 2018.

[14] Q. Huang, Y. Huang, Y. Luo, F. Yuan, and X. Li, "Segmentation of breast ultrasound image with semantic classification of superpixels," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101657.

[15] Y. Hu *et al.*, "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," *Med. Phys.*, vol. 46, no. 1, pp. 215–228, Jan. 2019.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[19] M. H. Yap *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[21] A. Vakanski, M. Xian, and P. Freer, "Attention enriched deep learning model for breast tumor segmentation in ultrasound images," 2019, *arXiv:1910.08978*. [Online]. Available: http://arxiv.org/abs/1910.08978

[22] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, Jan. 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, *arXiv:1412.6856*. [Online]. Available: http://arxiv.org/abs/1412.6856

[25] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[28] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.
[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[30] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[31] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 448–456. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045167
[33] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[34] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS) Autodiff Workshop*, Dec. 2017.
[35] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.
[36] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118432.

Haeyun Lee (Student Member, IEEE) was born in Iksan, South Korea. He received the B.S. degree in mathematics from Chonbuk National University, Jeonju, South Korea, in 2016, and the M.S. degree in information and communication engineering from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, in 2018. He is currently pursuing the Ph.D. degree in information and communication engineering with DGIST.

His current research interests include the image restoration, medical image analysis based on deep learning.

Jinhyoung Park (Member, IEEE) was born in Busan, South Korea, in 1975. He received the B.S. degree in astronomy and the M.S. degree in biomedical engineering from Seoul National University, Seoul, South Korea, in 2002 and 2004, respectively, and the Ph.D. degree in biomedical engineering from the University of Southern California at Los Angeles, Los Angeles, CA, USA, in 2011.

From 2004 to 2008, he was a Principal Engineer with SIEMENS Ultrasound Group Korea, Seoul. He was a Postdoctoral Research Associate with the Department of Biomedical Engineering, University of Southern California at Los Angeles in 2012 and a Senior Engineer with Volcano Corporation, Rancho Cordova, CA, USA, (which was acquired by Philips in 2015) from 2013 to 2016. Since 2016, he has been an Assistant Professor with the Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea. His research interests include the fabrication of ultrasound systems, transducers applied to intravascular ultrasound imaging, and noninvasive neuro modulations. He also has expertise in algorithms for interventional flow measurement. He has authored 25 peer-reviewed articles.

Jae Youn Hwang (Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2001, the M.S. degree in biomedical engineering from Seoul National University, Seoul, in 2003, and the Ph.D. degree in biomedical engineering from the University of Southern California at Los Angeles, Los Angeles, CA, USA, in 2009.

He was a Faculty with the Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, where he is currently an Associate Professor. His current research interests include the development of multimodality imaging and novel mobile healthcare systems based on high-frequency ultrasound and optical imaging techniques for diagnosis of various diseases.