SWIN-SFTNET : SPATIAL FEATURE EXPANSION AND AGGREGATION USING SWIN TRANSFORMER FOR WHOLE BREAST MICRO-MASS SEGMENTATION

Sharif Amit Kamran[†⋆], *Khondker Fariha Hossain*[†⋆], *Alireza Tavakkoli*[†], *George Bebis*[†], *Sal Baker*[§]

[†] Department of Computer Science & Engineering, University of Nevada, Reno, NV, USA
[§] School of Medicine, University of Nevada, Reno, NV, USA

ABSTRACT

Incorporating various mass shapes and sizes in training deep learning architectures has made breast mass segmentation challenging. Moreover, manual segmentation of masses of irregular shapes is time-consuming and error-prone. Though Deep Neural Network has shown outstanding performance in breast mass segmentation, it fails in segmenting micro-masses. In this paper, we propose a novel U-net-shaped transformer-based architecture, called Swin-SFTNet, that outperforms state-of-the-art architectures in breast mammography-based micro-mass segmentation. Firstly to capture the global context, we designed a novel Spatial Feature Expansion and Aggregation Block(SFEA) that transforms sequential linear patches into a structured spatial feature. Next, we combine it with the local linear features extracted by the swin transformer block to improve overall accuracy. We also incorporate a novel embedding loss that calculates similarities between linear feature embeddings of the encoder and decoder blocks. With this approach, we achieve higher segmentation dice over the state-of-the-art by 3.10% on CBIS-DDSM, 3.81% on InBreast, and 3.13% on CBIS pre-trained model on the InBreast test data set.

Index Terms— Breast mass segmentation, Mammogram, Swin Transformer, Deep learning, Medical Imaging

1. INTRODUCTION

Breast cancer is one of the most dominant cancer types in the world, and Mammography has been acknowledged as a vital tool for the early detection of breast cancer. However, asymmetrical shapes, microcalcifications, and small masses complicate automated breast mass segmentation. Additionally, most computer-aided diagnosis (CAD) systems rely on traditional image-processing-based approaches, which are quite error-prone and require manual intervention. Recently, machine learning and deep learning approaches have outperformed these conventional methods [1] and have become a popular technique for such tasks. Nonetheless, most CAD tools are still plagued by manually extracting suspicious regions or segments from low-resolution images, which fail to segment micro masses with accurate contour and high probability.

Deep Neural Network has shown excellent performance in medical image segmentation. Popular networks like U-Net [2], FCN [3], AUNet [4], ARF-Net [5] demonstrated outstanding outcomes for breast mass segmentation from both mammography images. These networks implemented diverse methods like generating multi-scale feature maps, attention-guided dense upsampling, and additive channel attention to learn robust feature maps to segment tumors of different sizes with more than 85%+ dice scores. However, the dice score of these systems falls to 5-15% when applied to images with micro-masses.

One reason for the failure of CNN-based approaches on micro-masses is they overtly focus on global semantic information. And to eliminate similar problems, Vision-transformer (ViT) [6] was proposed to prioritize local patch-level information. Taking 2D image patches with positional embeddings as input, Vision Transformers has outperformed most medical imaging downstream tasks [7–9]. Recently, Swin-UNet has achieved phenomenal results in organ segmentation like Gallbladder, Spleen, Liver, etc. Although Swin-UNet can capture local information correctly for precise boundary segmentation of organs, the organ is unique in shape and does not contain similar-looking artifacts. One of the primary problems of segmenting micro-masses in the breast is that surrounding fatty tissues can throw off the segmentation boundary of the model and might raise the false-positive rate as well. To address the above issues, we propose a novel transformer network named Swin Spatial Feature Transformer Network (Swin-SFTNet) and a novel embedding similarity loss to achieve a segmentation dice improvement over the state-of-the-art by 3.10%, 3.81%, and 3.13% on CBIS-DDSM [10], InBreast [11], and CBIS pre-trained on InBreast dataset respectively. Our main contributions are: (1) Employing a Swin-Transformer as a basic building block to create Swin-SFTNet to incorporate spatial global and sequential local context information in a multi-scale feature fusion configuration. (2) Designing a novel Spatial Feature Expansion and Aggregation Block to convert sequential linear patches into structured spatial features for capturing global context information for better micro-mass

[⋆] Equal Contribution

segmentation and (3) Utilizing a novel embedding loss that calculates similarities between features of the encoder and decoder blocks.

2. METHODOLOGY

2.1. Overall Architecture

The overall architecture of our proposed Swin-SFTNet is illustrated in Fig. 1. Swin-SFTNet incorporates an encoder, a decoder, three skip connections between the encoder and decoder, and three parallel SFEA blocks followed by patch extract and patch embedding layer before concatenating with the output feature map. Our architecture is an enhanced version of Swin-UNet [7], a UNet-like auto-encoder that replaces Swin-Transformer blocks [12] with regular convolution layers. We first transform the breast mammography grayscale images into RGB, providing the model with learning essential features. We utilize a patch-embedding layer to transform the input into non-overlapping patches of size 4×4. So for three RGB channels, we get to $4 \times 4 \times 3$ = 48 depth dimension. Next, we utilize a dense layer to project feature dimension into C arbitrary dimension. Following this layer, we have our encoder blocks, each consisting of two successive swin-transformer blocks and a patch-merging layer. We explain the swin-transformer block in Subsection 2.2. We repeat the encoder blocks three times to downsample the feature dimensions from $\frac{H}{4} \times \frac{W}{4} \times C$ to $\frac{H}{8} \times \frac{W}{8} \times 2C$, $\frac{H}{16} \times \frac{W}{16} \times 4C$ and $\frac{H}{32} \times \frac{W}{32} \times 8C$ successively.To conclude the encoder, we utilize two swin-transformer blocks after the last patch-merging layer.

Similar to the encoder, we design a symmetric decoder composed of multiple Swin Transformer blocks and patch expanding layer. Each decoder black is concatenated with the skip-connection features from the encoder with the same spatial dimension. As a result, we avoid any loss of spatial information due to successive downsampling. In contrast to the patch merging layer, the patch expanding layer reshapes the feature maps with $2\times$ up-sampling of spatial dimension. Additionally, it utilizes convolution to halve the depth dimension. We repeat the decoder blocks three times to upsample the feature dimensions from $\frac{H}{32} \times \frac{W}{32} \times 8C$ to $\frac{H}{16} \times \frac{W}{16} \times 4C$, $\frac{H}{8} \times \frac{W}{8} \times 2C$ and $\frac{H}{4} \times \frac{W}{4} \times C$ successively. The last patch-expanding layer is incorporated to perform 4× up-sampling to restore the resolution of the feature maps to the resolution $H \times W \times 4C$. We also concatenation operation through the skip-connection features from SFEA and each decoder block's outputs. We explain our proposed Spatial feature aggregation and Expansion block in Subsection 2.3. At last we apply a 2D convolution to get the output feature dimension $H \times W \times 1$ for binary mass segmentation. Here, $H = 256$, $W = 256$ and $C = 128$.

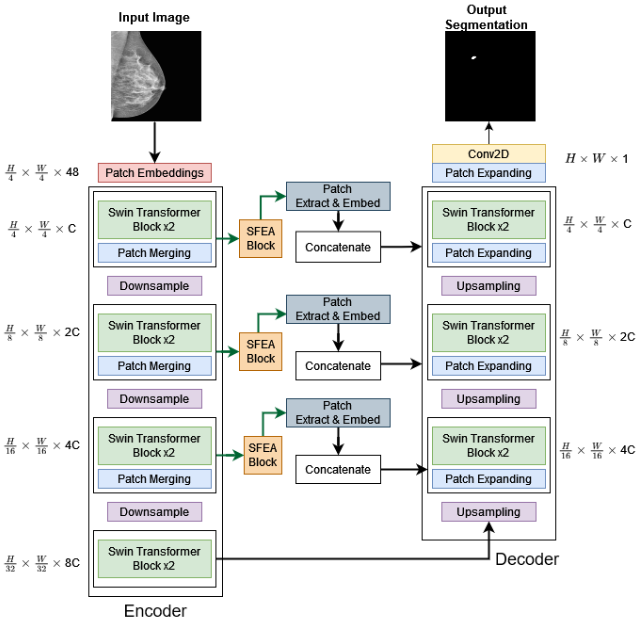
Input Image
Output Segmentation
$\frac{H}{4} \times \frac{W}{4} \times 48$
Patch Embeddings
Conv2D
Patch Expanding
$H \times W \times 1$
$\frac{H}{4} \times \frac{W}{4} \times C$
Swin Transformer Block x2
Patch Merging
SFEA Block
Patch Extract & Embed
Concatenate
Swin Transformer Block x2
Patch Expanding
$\frac{H}{4} \times \frac{W}{4} \times C$
Downsample
Upsampling
$\frac{H}{8} \times \frac{W}{8} \times 2C$
Swin Transformer Block x2
Patch Merging
SFEA Block
Patch Extract & Embed
Concatenate
Swin Transformer Block x2
Patch Expanding
$\frac{H}{8} \times \frac{W}{8} \times 2C$
Downsample
Upsampling
$\frac{H}{16} \times \frac{W}{16} \times 4C$
Swin Transformer Block x2
Patch Merging
SFEA Block
Patch Extract & Embed
Concatenate
Swin Transformer Block x2
Patch Expanding
$\frac{H}{16} \times \frac{W}{16} \times 4C$
Downsample
Upsampling
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Swin Transformer Block x2
Decoder
Encoder

Fig. 1. An overview of the proposed Swin-SFTNet consisting of Swin-transformer, Patch Merging, Patch Expanding, Patch Embedding and Spatial Feature Expansion & Aggregation Blocks.

2.2. Swin-Transformer Block

Traditional window-based multi-head self-attention (W-MSA) proposed in Vision Transformer (ViT) [6] utilizes a single low-resolution window for building feature-map and has quadratic computation complexity. In contrast, the Swin Transformer block incorporates shifted windows multi-head self-attention (SW-MSA), which builds hierarchical local feature maps and has linear computation time. Swin transformer block can be described in the following Eq. 1 and Eq. 2.

$$\begin{aligned} x^l &= W\text{-}MSA(\phi(x^{l-1})) \\ x^l &= \delta(\phi(x^l)) + x^l) \end{aligned} \tag{1}$$

$$\begin{aligned} x^{l+1} &= SW\text{-}MSA(\phi(x^l)) \\ x^{l+1} &= \delta(\phi(x^{l+1})) + x^{l+1}) \end{aligned} \tag{2}$$

In Eq. 1, we illustrate the first sub-block of swin transformer consisting of LayerNorm (ϕ) layer, multi-head self attention module (W-MSA), residual connection (+) and 2-layer MLP with GELU non-linearity (δ). In similar way Eq. 2 illustrates the second sub-block of swin transformer consisting of LayerNorm (ϕ) layer, shifted window multi-head self attention module (SW-MSA), residual skip-connection (+) and MLP with GELU activation (δ). Additionally, l notifies layer number and x is the feature-map.

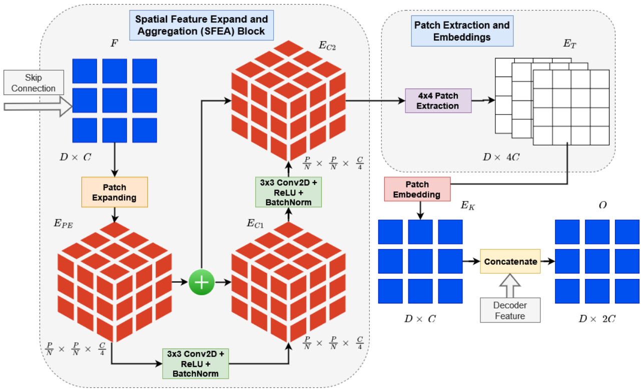
Spatial Feature Expand and Aggregation (SFEA) Block
F
E_{C2}
Skip Connection
$D \times C$
Patch Expanding
E_{PE}
E_{C1}
$\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}$
3x3 Conv2D + ReLU + BatchNorm
3x3 Conv2D + ReLU + BatchNorm
$\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}$
$\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}$
Patch Extraction and Embeddings
E_T
4x4 Patch Extraction
$D \times 4C$
Patch Embedding
E_K
O
Concatenate
Decoder Feature
$D \times C$
$D \times 2C$

Fig. 2. Visualization of the proposed SFEA Block

2.3. Spatial Feature Expansion and Aggregation

Although the multi-head self-attention module can capture local contextual information to understand inherent feature representations, consecutive patch merging and expanding layers can degrade the overall global context of the task. The existing skip connection concatenation cannot solve this problem as they apply dense layers on sequential patches to create linear projections. To create spatial projections of learnable features, we propose the Spatial Feature Expansion and Aggregation block illustrated in Fig. 2. We start with the top-most skip connection that comes out of the first encoder layer and has a feature output of $F \in \mathbb{R}^{D \times C}$, where $D = 4096$ and $C = 128$. We apply a patch expanding layer with patch-size 4×4 which gives us the feature output $E_{PE} \in \mathbb{R}^{\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}}$. Here, spatial dimension $P = 256$ and $N = 1$, so the resultant spatial dimension becomes $256 \times 256 \times 32$. In a similar manner, from the 2nd and 3rd skip connections with 1024×256 and 256×512 dimensional feature input we can get $128 \times 128 \times 64$ and $64 \times 64 \times 128$ feature outputs with $[N_2, N_3] = [2, 4]$. Next, we apply a 3×3 2D Convolution, ReLU activation, and Batch-Normalization operation followed by element-wise addition of features from E_{PE} to get output feature $E_{C1} \in \mathbb{R}^{\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}}$. In a similar manner, we apply another same 2D Convolution Block on E_{C1} to get feature output and add element-wise features from E_{PE} to get final output $E_{C2} \in \mathbb{R}^{\frac{P}{N} \times \frac{P}{N} \times \frac{C}{4}}$. These two convolution operation helps with extracting global spatial context information that we further combine with our decoder's local patch-level information. Following this operation, we utilize the 4×4 Patch-extraction operation to convert it the feature into 2D sequence feature output, $E_T \in \mathbb{R}^{D \times 4C}$. After that, we use patch-embedding layer to make the feature dimension same as the decoder's paired output, so the output feature map becomes $E_K \in \mathbb{R}^{D \times C}$. Next, we concatenate the feature from the decoder's patch expanding layer with the E_K. We do this for all of our skip connections, so the output feature map becomes $O \in \mathbb{R}^{D \times 2C}$. Here, we use three different values for $C = [128, 256, 512]$, for the three skip connections.

2.4. Objective Function and Embedding Similarity Loss

For binary output of background and masses we use binary cross-entropy loss given in Eq. 3. We also use Dice-coefficient loss given in Eq. 4 for better segmentation output. For dice-coefficient we use $\varepsilon = 1.0$ in numerator and denominator for addressing the division by zero. Here, $\mathbb{E}$ symbolizes expected values given, p (prediction) and y (ground-truth).

$$\mathcal{L}_{bce} = \mathbb{E}_{p,y}\big[\frac{1}{N}\sum_{i=1}^{N} -(y_i * log(p_i) + (1-y_i) * log(1-p_i))\big] \tag{3}$$

$$\mathcal{L}_{dsc} = \mathbb{E}_{p,y}\big[1 - \frac{2\sum_{i=1}^{N} p_i y_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \varepsilon}\big] \tag{4}$$

Finally, the embedding feature loss is calculated by obtaining positional and patch features from the transformer encoder layers E and decoder layers D by inserting the image, as shown in Eq. 5. Here, Q stand for the number of features extracted from the embedding layers of the transformer-encoder.

$$\mathcal{L}_{emb} = \mathbb{E}_{x,y}\sum_{i=1}^{k}\frac{1}{Q} \parallel E_{em}^{i}(x) - D_{em}^{i}(x) \parallel \tag{5}$$

We combine Eq. 3, 4, and 5 to configure our ultimate cost function as provided in Eq. 6. Here, λ is the weight for each loss.

$$\mathcal{L} = \lambda_{dsc} * \mathcal{L}_{dsc} + \lambda_{bce} * \mathcal{L}_{bce} + \lambda_{emb}\mathcal{L}_{emb} \tag{6}$$

3. EXPERIMENTS

3.1. Dataset

We evaluated our model with three publicly available datasets. We used CBIS-DDSM [10] and InBreast [11], two whole mammography segmentation datasets. All images are resized to 256×256 dimension using bilinear interpolation, and the masks are resized to the same size using the nearest-neighbor technique. Both dataset contains craniocaudal (CC) and mediolateral oblique (MLO) views of breasts. From the CBIS-DDSM dataset, we separate 849 training and 69 test images based on the subtlety of 4 and 5. The masses on the test images are less than 200 pixels in size, which is 0.3% of the whole image. The subtlety defines the visual challenge to annotate the masses for the clinician, with 1-5 grading where 1= ungradable and 5=most gradable. We use OpenCV's contour-based technique to remove artifacts, and for enhancement, we use CLAHE. The Inbreast dataset contains 107 images, which we split into 90 training and 17 test images. The test images are separated based on any mass being less than 100 pixels or smaller.

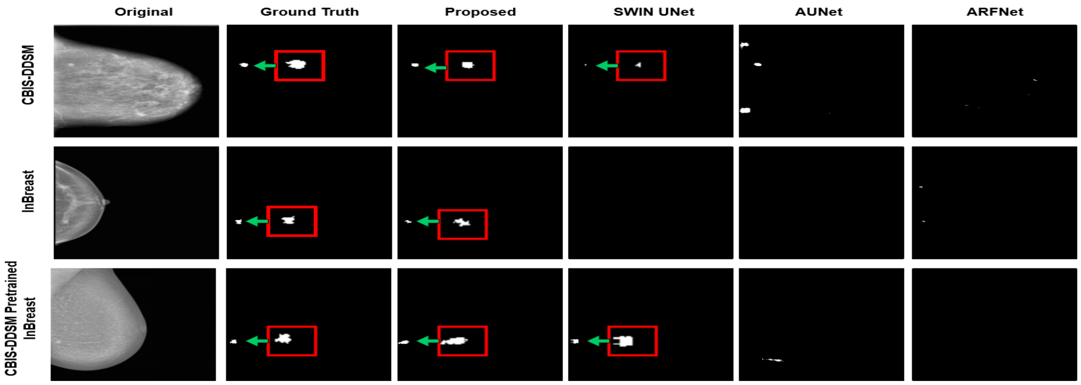

Fig. 3. Qualitative performance of Swin-SFTNet vs. other architectures on three datasets. (Red box) are zoomed-in images.

3.2. Hyper-parameter

We chose $\lambda_{bce} = 0.4$, $\lambda_{dice} = 0.6$ and $\lambda_{emb} = 0.01$ (Eq. 6). For optimizer, we used Adam with a learning rate of $\alpha = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We used Tensorflow 2.8 to train the model in mini-batches with the batch size, $b = 8$ in 100 epochs which took around 1 hour to train on NVIDIA A30 GPU. The inference time is 41 millisecond per image.

3.3. Quantitative Evaluation

For micro-mass segmentation tasks, we compare our model with three state-of-the-art architectures, AUNet [4], ARF-Net [5], and SWIN-UNet [7] for CBIS-DDSM and InBreast, as given in Table. 1. AUNet utilizes attention-guided dense up-sampling to retain important spatial features lost due to bilinear up-sampling. In contrast, ARF-Net uses a Selective Receptive Filed Module (SRFM) module to fuse multi-scale and multi-receptive field information and the current state-of-the-art for both breast segmentation datasets. Finally, Swin-UNet combines swin transformer blocks presented in [12] with a UNet-like structure to reach high precision in multiple organ segmentation. ARFNet and AUNet show high-performance gains against previous approaches. However, the prediction is skewed because the test set contains images with more large masses and few micro-masses. We designed the experiment to emphasize micro-mass segmentation, so we sorted the images based on the tumor size. For InBreast, we chose the last portion for testing (small than 100 px), which has the smallest breast masses. And for CBIS-DDSM, we discarded the leftover large mass test images (larger than 200 px), as we had separate training images.

For metrics, we use Dice score = $\frac{2\times TP}{2\times TP+FP+FN}$, Mean IOU (mIOU) = $\frac{TP}{TP+FP+FN}$, Sensitivity (SEN) = $\frac{TP}{TP+FN}$, and Specificity (SPE) = $\frac{TN}{TN+FP}$. We can see from Table. 1, our model achieves the best score compared to others for all the metrics. We reach higher segmentation dice score over the state-of-the-art by 3.10% on CBIS-DDSM, 3.81% on InBreast, and 3.13% on CBIS pre-trained model tested on InBreast (Given in Red). Moreover, in qualitative comparison in Fig. 3, our model can segment harder and smaller masses than other architectures.

Table 1. Comparison for **CBIS-DDSM**, **InBreast** and **CBIS-DDSM Pretrained InBreast**

Dataset	Model	Dice(%)	mIoU(%)	SEN(%)	SPE(%)
CBIS-DDSM	AUNet	14.20	9.14	29.81	99.40
	ARFNet	2.54	1.35	1.44	99.98
	SWIN UNet	21.03	15.53	31.36	99.70
	Proposed	**24.13**	**17.44**	**33.31**	**99.72**
InBreast	AUNet	12.11	9.35	15.28	99.94
	ARFNet	13.95	8.69	21.21	99.68
	SWIN UNet	14.12	9.46	28.17	99.55
	Proposed	**17.93**	**13.10**	**20.56**	**99.81**
CBIS-DDSM Pretrained InBreast	AUNet	17.24	12.41	23.49	99.74
	ARFNet	2.84	1.53	65.04	86.30
	SWIN UNet	20.25	14.27	29.46	99.58
	Proposed	**23.38**	**17.40**	**34.54**	**99.87**

We also did ablation study for the embedding loss for two datasets, which are provided in Table. 2. With the novel loss function we have 3.14%, 6.33%, and 0.86% gain for CBIS-DDSM, InBreast, and CBIS pre-trained model consecutively.

Table 2. Dice score for with or w/o Feature Matching Loss

Feature Matching Loss	CBIS-DDSM	InBreast	CBIS-DDSM Pretrained InBreast
With	**24.13**	**17.93**	**23.38**
Without	20.72	11.60	22.52

4. CONCLUSION

In this paper, we proposed Swin-SFTNet, with a novel Spatial Feature Expansion and Aggregation Block (SFEA) block, which captures the global context of the images and fuses it with the local patch-wise features. Moreover, we also integrate a novel embedding loss that computes the similarities between the encoder and decoder block's patch-level features. Our model outperforms other architectures in micro-mass segmentation tasks in two popular datasets.

5. ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation under Grant No. OIA- 2148788 and OAC- 2201599.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [10, 11]. Ethical approval was not required, as confirmed by the license attached with the open-access data.

7. REFERENCES

[1] Vivek Kumar Singh, Hatem A Rashwan, Santiago Romani, Farhan Akram, Nidhi Pandey, Md Mostafa Kamal Sarker, Adel Saleh, Meritxell Arenas, Miguel Arquez, Domenec Puig, et al., "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network," *Expert Systems with Applications*, vol. 139, pp. 112855, 2020.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] Hui Sun, Cheng Li, Boqiang Liu, Zaiyi Liu, Meiyun Wang, Hairong Zheng, David Dagan Feng, and Shanshan Wang, "Aunet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms," *Physics in Medicine & Biology*, vol. 65, no. 5, pp. 055005, 2020.

[5] Chunbo Xu, Yunliang Qi, Yiming Wang, Meng Lou, Jiande Pi, and Yide Ma, "Arf-net: An adaptive receptive field network for breast mass segmentation in whole mammograms and ultrasound images," *Biomedical Signal Processing and Control*, vol. 71, pp. 103178, 2022.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[8] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, and Salah A Baker, "Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3235–3245.

[9] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[10] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[11] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.