IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 February 2024, accepted 18 February 2024, date of publication 21 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368170

RESEARCH ARTICLE

BGRD-TransUNet: A Novel TransUNet-Based Model for Ultrasound Breast Lesion Segmentation

ZHANLIN JI[1,2], (Member, IEEE), HAORAN SUN[1], NA YUAN[3], HAIYANG ZHANG[4], JIAXI SHENG[1], XUEJI ZHANG[5], AND IVAN GANCHEV[2,6,7], (Senior Member, IEEE)

[1]Hebei Key Laboratory of Industrial Intelligent Perception, North China University of Science and Technology, Tangshan 063210, China
[2]Telecommunications Research Centre (TRC), University of Limerick, Limerick, V94 T9PX Ireland
[3]Intelligence and Information Engineering College, Tangshan University, Tangshan 063000, China
[4]Department of Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China
[5]School of Biomedical Engineering, Shenzhen University Health Science Center, Shenzhen, Guangdong 518060, China
[6]Department of Computer Systems, Plovdiv University "Paisii Hilendarski," 4000 Plovdiv, Bulgaria
[7]Institute of Mathematics and Informatics—Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

Corresponding authors: Ivan Ganchev (Ivan.Ganchev@ul.ie) and Xueji Zhang (zhangxueji@szu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0135700; in part by the Bulgarian National Science Fund (BNSF) under Grant КП-06-ИП-КИТАЙ/1 (KP-06-IP-CHINA/1); and in part by the Telecommunications Research Centre (TRC) of the University of Limerick, Ireland.

ABSTRACT Breast UltraSound (BUS) imaging is a commonly used diagnostic tool in the field of counter fighting breast diseases, especially for early detection and diagnosis of breast cancer. Due to the inherent characteristics of ultrasound images such as blurry boundaries and diverse tumor morphologies, it is challenging for doctors to manually segment breast tumors. In recent years, the Convolutional Neural Network (CNN) technology has been widely applied to automatically segment BUS images. However, due to the inherent limitations of CNNs in capturing global contextual information, it is difficult to capture the full context. To address this issue, the paper proposes a novel BGRD-TransUNet model for breast lesion segmentation, based on TransUNet. The proposed model, first, replaces the original ResNet50 backbone network of TransUNet with DenseNet121 for initial feature extraction. Next, newly designed Residual Multi-Scale Feature Modules (RMSFMs) are employed to extract features from various layers of DenseNet121, thus capturing richer features within specific layers. Thirdly, a Boundary Guidance (BG) network is added to enhance the contour information of BUS images. Additionally, newly designed Boundary Attentional Feature Fusion Modules (BAFFMs) are used to integrate edge information and features extracted through RMSFMs. Finally, newly designed Parallel Channel and Spatial Attention Modules (PCSAMs) are used to refine feature extraction using channel and spatial attention. An extensive experimental testing performed on two public datasets demonstrates that the proposed BGRD-TransUNet model outperforms all state-of-the-art medical image segmentation models, participating in the experiments, according to all evaluation metrics used (except for few separate cases), including the two most important and widely used metrics in the field of medical image segmentation, namely the Intersection over Union (IoU) and Dice Similarity Coefficient (DSC). More specifically, on the BUSI dataset and dataset B, BGRD-TransUNet achieves IoU values of 76.77% and 86.61%, and DSC values of 85.08% and 92.47%, respectively, which are higher by 7.27 and 3.64, and 5.81 and 2.54 percentage points, than the corresponding values achieved by the baseline (TransUNet).

INDEX TERMS Breast disease, breast ultrasound (BUS), tumor segmentation, medical image segmentation, TransUNet.

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

I. INTRODUCTION

According to the latest global cancer data report [1], from 2015 to 2019, the number of women breast cancer

incidences has continued to increase, resulting in a quite high mortality rate reaching second place after the lung cancer. It has become the deadliest cancer among women, posing a serious threat to their health. Early diagnosis and treatment of breast cancer are crucial for improving the cure rate and survival rate. Breast X-ray examinations and magnetic resonance imaging (MRI) are methods used in breast screening. However, due to radiation concerns, both doctors and patients often approach these types of examinations with caution. Breast UltraSound (BUS) imaging technology, on the other hand, is a radiation-free, real-time, and cost-effective means for detection, [2]. Therefore, it has been rapidly accepted by both medical professionals and patients and is increasingly being used in breast tumor detection. Segmenting tumors from BUS images is a crucial step in assisting doctors in accurately locating and describing tumor regions. However, BUS tumor images inherently present uncontrollable issues such as blurry lesion boundaries and diverse tumor morphologies. These challenges make it difficult for doctors to accurately segment breast tumors manually. To enhance the objectivity and accuracy of diagnosis, creating computer-aided diagnostic systems to assist doctors in segmenting BUS images is an urgent need in current research.

To date, research in image segmentation primarily has followed two directions utilizing traditional methods and deep learning methods, respectively. Relying on manually designed features and rules, traditional methods, such as threshold segmentation, region growing, edge detection, etc., can generally only handle images where the target and background have significant differences [2]. Consequently, traditional methods have limitations such as high demand for image quality and sensitivity to parameters. With the advancement of machine learning to deep learning, more and more researchers are turning to using deep learning methods to improve segmentation accuracy. Deep learning convolutional neural networks (CNNs) have gained widespread application in medical image segmentation as they offer significant advantages over traditional segmentation methods [3]. Boosted by the recent developments, CNNs have been repeatedly applied to BUS image segmentation [4]. Among these, the most representative example is U-Net [5]. Characterized by its symmetric encoder-decoder U-shape, U-Net is a prevalent model in medical segmentation. Multiple variations of U-Net, such as U-Net++ [6], Attention UNet [7], Res-UNet [8], and others, have been developed and used to date.

However, due to the inherent limitations of CNN operations, U-Net often struggles to capture global contextual information, [9]. This limitation arises because the size of the convolutional kernel in CNN determines its local perception ability, restricting a single convolutional kernel to focus on a limited local region, [10]. While CNN models can gradually expand the receptive field of convolutional kernels by stacking numerous convolutional layers (which allows higher-level convolutional kernels to focus on global features), lower and intermediate-level convolutional kernels remain constrained and can only focus on local regions. In contrast to the CNN architecture, the Transformer architecture [11] excels at capturing global contextual information. Consequently, some researchers have combined Transformer with U-Net, giving rise to the representative work of TransUNet [9]. TransUNet draws inspiration from the Vision Transformer (ViT) model [12] and incorporates Transformer into the encoder part of U-Net. Although TransUNet addresses the issue of CNN's difficulty in capturing global information, it presents some new challenges. Compared to other image segmentation datasets, medical image datasets (especially BUS datasets) are often extremely limited in size, typically comprising only a few hundred or thousand data samples. Training TransUNet directly on these small-size datasets can easily lead to model overfitting. Another issue is that due to the semantic differences between the shallow encoder and decoder in TransUNet, shallow-level features with less semantic information may impair the final model performance through simple skip connections. Furthermore, the problem of fuzzy lesion boundaries in BUS images remains unaddressed.

To cope with the aforementioned issues, we propose an improved model based on TransUNet, named BGRD-TransUNet, with the following main contributions:

1) In the encoder, the original ResNet50 backbone network [13] of TransUNet is replaced with DenseNet121 [14]. This leverages DenseNet's features, such as parameter sharing and feature reuse, to address the model overfitting problem caused by small-size BUS image datasets;
2) During the skip connection process, newly designed Residual Multi-Scale Feature Modules (RMSFMs) are employed to expand the receptive field, allowing to capture richer features at specific layers;
3) A newly designed Deformable Atrous Spatial Pyramid Pooling Module (DASPPM) is introduced between the encoder and decoder to enhance the extraction of complex shapes;
4) The original skip connection mechanism of TransUNet is enhanced by introducing a newly designed Boundary Guidance (BG) network that separately trains boundary features to address the issue of fuzzy lesion boundaries in BUS images;
5) Newly designed Boundary Attentional Feature Fusion Modules (BAFFMs), incorporating a multi-scale channel attention mechanism, are used by the proposed model to obtain attention-based fusion features. This mitigates the semantic differences between the shallow encoder and decoder. Additionally, within each BAFFM, boundary features, extracted by BG, are integrated;
6) In the decoder, the proposed model uses newly designed Parallel Channel and Spatial Attention Modules (PCSAMs) to better capture essential features in

the images by simultaneously considering both channel information and spatial information.

The remaining structure of this paper is the following. Section II describes related work done in this area. Section III describes the proposed BGRD-TransUNet model in detail. Section IV presents the conducted experiments and analyses the obtained results, followed by conclusions presented in Section V.

II. REALATED WORK

A. APPLICATION OF TRANSFORMERS IN MEDICAL IMAGE SEGMENTATION

Influenced by the significant success of Transformer models in the field of Natural Language Processing (NLP), Transformers have been further extended into the realm of computer vision. Dosovitskiy et al. [12] introduced an improved model based on Transformers, known as ViT, which was designed to address the shortcomings of CNNs in capturing distant semantic dependencies [15]. This innovation garnered widespread attention in the field of computer vision. In the domain of medical image segmentation, the application of ViT has also demonstrated immense potential. One of the most noteworthy developments was the integration of ViT with the U-Net architecture, which can be categorized into two main approaches: (i) supplementing a U-shaped CNN network structure with Transformers; and (ii) constructing an independent U-shaped Transformer architecture.

Among the models that supplement the U-shaped CNN network structure with Transformers, one of the most representative examples is TransUNet [9]. Within the TransUNet encoder, there are 12 stacked Transformer layers, which employ tokenized pathways to extract abstract features from the original input, thereby capturing global context information. In the decoder pathway, encoded features are combined with high-resolution CNN feature maps and used for upsampling. This process enables precise localization [15].

Subsequently, some researchers proposed an alternative approach that does not rely on traditional CNNs but rather constructs segmentation networks solely using Transformers. A notable example in this group is Swin-UNet, where Swin Transformer blocks [16] serve as a primary structure of a U-shaped network [17]. Swin Transformer plays a crucial role in visual tasks, operating as an efficient linear transformer with the ability to support hierarchical architectures. A key design feature of Swin Transformer is its shift window scheme, which enables the Transformer to calculate relationships among various patches within the same window. Subsequently, the window shifts across patches and calculates attention within it. This continuous shifting operation, along with the capture of local contextual information among patches within the window, can be stacked multiple times. Finally, a patch merging layer is used to construct hierarchical feature maps by merging deep-level patches. This concept shares similarities with the U-Net structure and has been successfully applied in the field of medical image segmentation, resulting in the Swin-UNet architecture.

In recent years, Transformer-based models have continued attracting much attention in the field of medical image segmentation and achieved remarkable results in solving problems caused by blurred lesion boundaries. A series of models, represented by TransDeepLab [18] and HiFormer [19], make full use of the Transformer's self-attention mechanism and cross-context attention mechanism, as well as utilize the Swin-Transformer module to achieve multi-scale feature fusion. These models effectively improve segmentation performance and efficiency by capturing long-distance dependencies and spatial correlations. By cleverly combining the advantages of CNNs and Transformers, they achieve detailed fusion of global features and local features, thus further improving the processing of boundary areas. Different from these models that rely on Transformers, DBGA-Net [20] proposes a novel idea, using a dual-branch global-local attention network. This model focuses on the global features of the entire image by designing a global channel attention module (GCAM). Although Transformer is not used directly, through GCAM's global average pooling and weight vector mapping, effective extraction of global information is successfully achieved, bringing new ideas for use in the image segmentation task. Overall, these two new groups of models introduce global information in different ways and fuse local information, effectively improving the their performance and robustness in medical image segmentation tasks. These studies provide useful inspiration for further developments in the field.

In this paper, we have made improvements to the TransUNet architecture by replacing the network model in the CNN part. This change assists the model in retaining more features, reducing the risk of overfitting, and improving the segmentation performance.

B. MULTI-SCALE FEATURE EXTRACTION

In natural scenes, the presence of multi-scale visual patterns is a common occurrence. Firstly, objects within a single image may vary significantly in size, such as the notable size difference between a table and a cup. Secondly, crucial contextual information about objects can extend far beyond the boundaries of the objects themselves. For example, to accurately determine whether a small black spot on a large table is a cup or a pencil holder, we rely on the context provided by the large table. Thirdly, gathering information from different scales is incredibly important for understanding object parts and wholes, especially in tasks such as fine-grained classification and semantic segmentation. Therefore, designing an effective multi-scale feature extraction method is essential for visual tasks.

In previous research, many state-of-the-art multi-scale feature extraction networks were proposed. For example, AlexNet [21] introduced a stacked convolution operation approach, enabling data-driven multi-scale feature learning.

Subsequently, the efficiency of multi-scale capabilities was further improved by using convolutional layers with different kernel sizes, as seen in the Inception architecture. Additionally, Gao et al. [22] introduced a simple yet effective multi-scale processing method called Res2Net. It divides the input feature maps into several groups. Initially, a group of filters extracts features from one set of input feature maps. Then, the output features from the previous group are combined with another set of input feature maps and sent to the next group of filters. This process repeats several times until all input feature maps are processed. Finally, features from all groups are concatenated and sent to another set of 1×1 filters for information fusion.

In the case of BUS images, cancer can vary in terms of size and position. Therefore, building upon Res2Net, we have designed a novel Residual Multi-Scale Feature Module (RMSFM) to extract more detailed features. This module is described in detail in Subsection III-B.

C. BOUNDARY GUIDANCE (BG)

For BUS images, accurate lesion segmentation remains challenging due to poor image quality, such as low contrast and high speckle noise, as well as edge blurring [23]. In previous research, several methods were proposed to address this issue. MLMSNet [24] used supervised foreground boundary detection and edge detection. AFNet [25] developed a meticulous feedback module to better explore target structures. CPDNet [26] introduced a partial decoder to refine high-level features for generating precise saliency maps. CF2-Net [27] employed two consecutive convolutional layers with kernel sizes of 1×1 and 3×3, respectively, to extract edge information from fused features generated from skip connections. Subsequently, a 1×1 convolutional layer further extracted high-level edge semantic features, which were then connected with the fused features as input for a mini U-Net. In this context, we further propose a boundary prediction approach to enhance lesion segmentation accuracy. This is explained in detail in Subsection III-D.

D. DenseNet121

DenseNet121 is a deep neural network, part of the Dense Convolutional Network series (DenseNet) [14], Figure 1. The "121" in the name represents the number of layers in the network, including 121 layers of convolutional and fully connected layers.

Unlike traditional CNNs, DenseNet121 adopts a dense connectivity design, where each layer is connected to all preceding layers. This enables thorough information propagation and sharing within the network, effectively alleviating the issue of gradient vanishing. It also makes the network more compact and efficient. The basic building blocks of DenseNet121 consist of multiple dense blocks, each containing several convolutional layers, batch normalization layers, and dense connections. Additionally, to further reduce the dimensionality of feature maps, DenseNet121 introduces

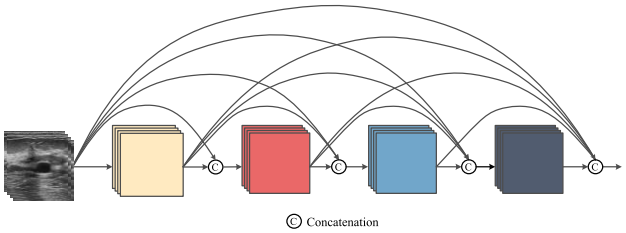

FIGURE 1. The DenseNet structure.

transition layers. These layers employ 1×1 convolutions and average pooling to compress the size of feature maps, thereby reducing computational complexity. DenseNet121 achieved outstanding results in the ImageNet image classification competition and has been widely applied in various computer vision tasks, including object detection and image segmentation. Its design philosophy of dense connectivity provides crucial insights and references for constructing efficient and accurate deep neural networks.

III. PROPOSED MODEL: BGRD-TransUNet

In this section, the overall structure of the proposed BGRD-TransUNet model is first introduced, followed by a detailed description of its main modules.

A. OVERALL STRUCTURE

The BGRD-TransUNet model is an improvement upon the TransUNet model. Currently, some models based on TransUNet, such as DA-TransUNet [28] and IB-TransUNet [10], introduce bottleneck modules between the CNN and Transformer to optimize the feature maps fed into the Transformer. However, the CNN components of these models still utilize the original ResNet50 [13] structure of TransUNet without modification. Due to the specific nature of our research and the limited size of the utilized datasets, using ResNet50 as the CNN backbone is insufficient for our needs. To mitigate the risk of model overfitting due to the scarcity of data, we propose, for the first time, replacing the ResNet50 in the CNN backbone of TransUNet with DenseNet121 [14].

The proposed BGRD-TransUNet model consists of an encoder, skip connections, a decoder, and a boundary guidance (BG) network, as shown in Figure 2.

In the encoder, the original ResNet50 backbone network of TransUNet is replaced with DenseNet121, where each layer has access to information from all previous layers in the network, which helps alleviate the gradient vanishing problem. Additionally, due to using dense connections, DenseNet121 achieves more efficient parameter sharing. This means that compared to ResNet50 with the same depth, DenseNet121 typically requires fewer parameters, reducing the risk of model overfitting and improving parameter efficiency. As previously mentioned, the BUS dataset is small in size, but Transformer models have a large number of parameters, which can pose a risk of model overfitting in

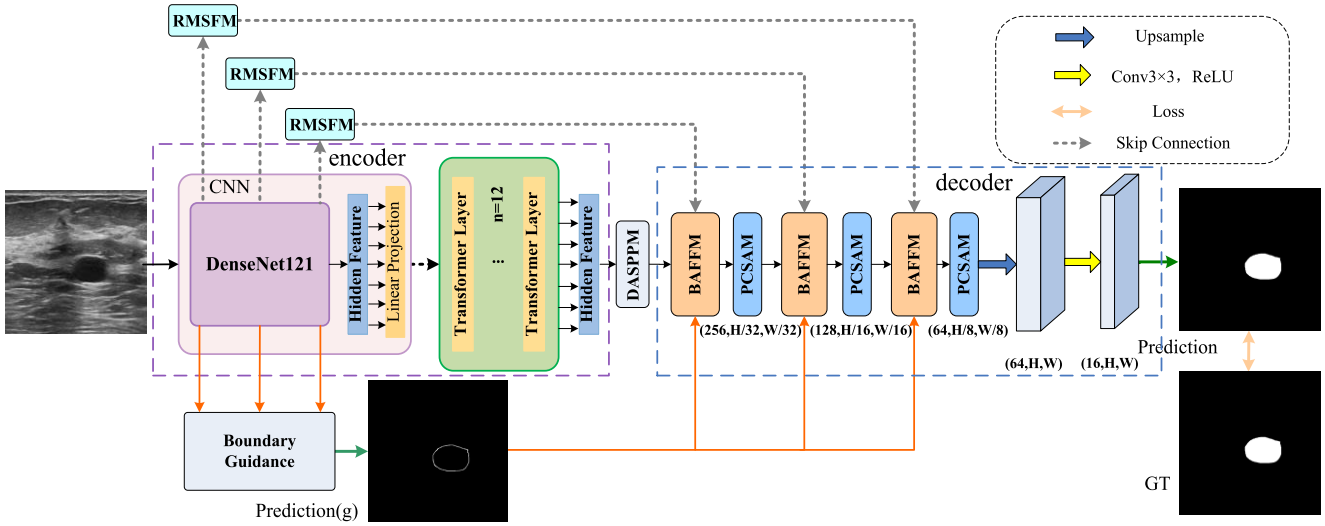

FIGURE 2. The proposed BGRD-TransUNet model.

segmentation tasks. Therefore, by replacing ResNet50 with DenseNet121, we leverage the inherent characteristics of the latter to mitigate model overfitting issues. Additionally, a newly designed module, DASPPM, is introduced after ViT to expand the receptive field.

In skip connections, newly designed modules, RMSFMs, are introduced to help alleviate semantic differences between the shallow encoder and decoder, and extract richer feature information. Different from the commonly used multi-resolution fusion, e.g., in IB-TransUNet, or adding attention mechanism, e.g., in DA-TransUNet, to modify the skip link, we propose to use the residual multi-scale feature extraction method to strengthen the skip connections. Convolutions with different convolution kernel sizes are used to extract features of each layer in parallel to achieve multi-scale feature extraction in the encoding stage. Compared with previous works, this operation not only reduces the semantic difference between the encoding stage and decoding stage, but also more fully extracts the information of lesions of different sizes in the images, making the model more suitable for segmenting images in datasets with different target sizes. In addition, unlike the residual multi-scale feature fusion method mentioned by Qin et al. [29], we did not choose to share feature information in the residual connection of different scale feature information, but adopted a top-down cumulative transfer method, as shown in Figure 3, to prevent the generation of redundant features and reduce the impact of redundant feature information on the model performance.

A separate boundary guidance (BG) network was designed by us and used to enhance boundary information. In the decoder, newly designed modules, BAFFMs, are inserted to better fuse skip connection information and BG information. Specifically, we designed a simple boundary detection network that connects three layers of features from the DenseNet121 backbone and uses convolution to obtain a boundary map supervised by the GT boundary map. Finally, the predicted boundary map is fused with the trunk features as boundary features to enhance the sensitivity of the model to boundaries. Unlike previous works of boundary guidance, we do not fuse the boundary map with the backbone features directly by channel splicing or pixel-by-pixel point summation. In order to achieve fuller use of the boundary information, we adopt the affine operation to fuse the boundary features with the backbone features, which is performed by the BF module utilized by BAFFMs, described in Subsection III-E. More specifically, the affine operation is conducted with two parameters – a scaling parameter and a translation parameter, which learn to adjust the scaling and translation of the input data during the training process, so as to improve the expressive ability and adaptability of the model. The purpose of fusing boundary features with backbone features using the affine operation is to adaptively adjust the contribution of boundary information through learning parameters, strengthen the boundary features, make the network more focused on the task requirements, and make full use of important information about the shape and location of the lesion during the fusion process.

Additionally, after each BAFFM, another newly designed module, PCSAM, is added to further enhance the model's segmentation performance. Finally, through a 3×3 convolutions and upsampling, the original image size is restored.

B. RMSFM

Due to the limitations of traditional sequential convolution operations, it is not possible to extract rich contextual information. Inspired by the Inception module [30] and Res2Net [22], a newly designed module, called RMSFM (Figure 3), is proposed here to enhance the model segmentation performance by extracting richer feature

information and alleviating the semantic gap between the shallow encoder and decoder.

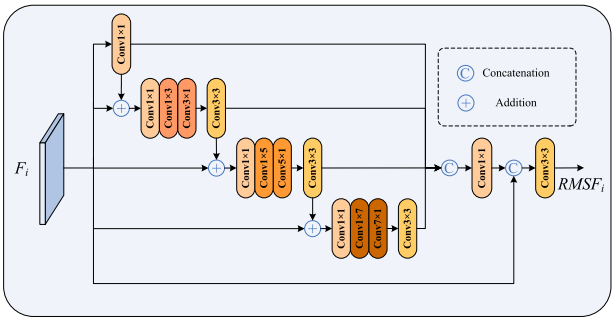

FIGURE 3. The designed residual multi-scale feature module (RMSFM).

RMSFM employs parallel convolution operations, progressively enlarging the receptive field using residual blocks. RMSFM utilizes five branches to capture different scales of information. Four of these branches start with a 1×1 convolution layer. The subsequent two layers consist of $1\times(2k\text{-}1)$ and $(2k\text{-}1)\times 1$ convolutional layers, employing asymmetric convolutions to reduce computational complexity. The final layer is a 3×3 convolutional layer with a dilation rate of $(2k\text{-}1)$ when $k > 2$. The general operation of these four branches is performed according to the following formula:

$$Branch_k^i = \begin{cases} Conv_{1\times1}(F_i); & k = 1 \\ Conv_{1\times1,1\times3,3\times1,3\times3}(F_i \oplus Branch_1^i); & k = 2 \\ Conv_{1\times1,1\times5,5\times1,3\times3}(F_i \oplus Branch_2^i); & k = 3 \\ Conv_{1\times1,1\times7,7\times1,3\times3}(F_i \oplus Branch_3^i); & k = 4 \end{cases} \quad (1)$$

where F_i denotes the i-th input feature map generated by DenseNet121, k denotes the output of the k-th branch in $Branch_k^i$, $\oplus$ denotes element-wise addition, and $Conv()$ refers to the stacked convolutional layers mentioned above. The outputs of these branches are then concatenated, and 1×1 convolutions are applied to adjust their channel dimensions to match the input channel size. The fifth branch, which does not undergo any operations, is combined with the other four branches. Then, a 3×3 convolution operation is performed, followed by a ReLU activation, to obtain the ultimate feature. Finally, the output feature $RMSF_i$ is obtained as follows:

$$RMSF_i = Conv_{3\times3}(Cat(F_i, Conv_{1\times1}(Cat_k^4(Branch_k^i)))) \quad (2)$$

where Cat_k^4 denotes the concatenation of all four branches and Cat denotes the concatenation of the input feature F_i and fused features from the four branches.

C. DASPPM

Traditional convolution operations divide the feature map into sections of the same size as the convolutional kernel and perform convolution. Each section's position on the feature map is fixed. This approach may work well for objects with regular shapes, but it becomes less effective for complex objects like tumors, which have varying shapes, sizes, and positions. Therefore, a newly designed module, called DASPPM (Figure 4), is introduced between the encoder and decoder of the proposed model to enhance the extraction of complex shapes.

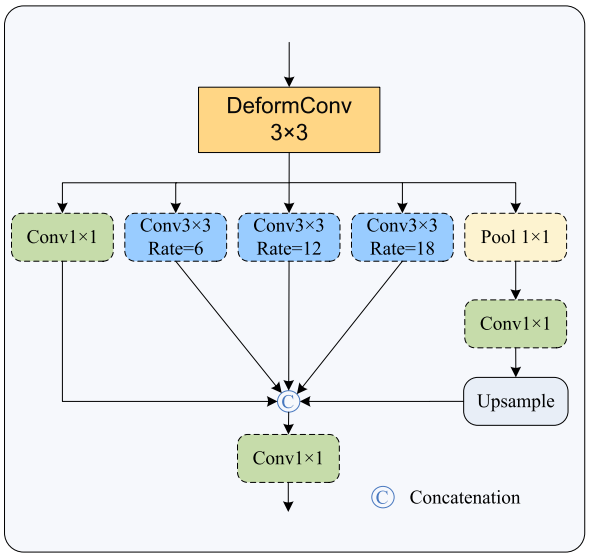

FIGURE 4. The designed deformable atrous spatial pyramid pooling module (DASPPM).

DASPPM consists of a 3×3 deformable convolution followed by a serial connection with Atrous Spatial Pyramid Pooling (ASPP) [31]. The deformable convolution uses additional offsets to increase spatial sampling positions within the module. These offsets are learned from the target task, allowing the entire module to better consider changes in the target's shape. ASPP, on the other hand, employs multiple parallel void convolutional layers with different sampling rates. It processes the features extracted for each sampling rate in separate branches to capture multi-scale object information and then combines them to generate the final result. In DASPPM, the deformable convolution and ASPP are serially connected, leveraging the dual advantages of deformable convolution and ASPP to address the challenges posed by the complex and diverse features of targets.

D. BG NETWORK

For the BUS image segmentation task, it is known that pixels near the object boundaries are complex, [32]. Since the encoder uses multiple convolutional layers and Transformers to extract features, a significant amount of upsampling is needed to restore the resolution, which results in some degree of spatial information loss. This issue is particularly pronounced in BUS data segmentation because the boundaries of objects in ultrasound images are already unclear, and further information loss makes the boundaries even more ambiguous. Therefore, we attempted to extract boundary information and integrate it into the feature space of the decoder to enhance

the model's sensitivity to boundaries. With prior knowledge of object boundaries, one can easily identify objects with patterns similar to the background. Thus, boundary information plays a crucial role in the image segmentation task. However, during the model training phase, we cannot directly use the gradient map of the ground truth as guidance. Instead, we need to train a separate network to obtain the boundary map.

We designed a simple Boundary Guidance (BG) network (Figure 5) that takes the three feature maps from the backbone and processes them through a 3×3 convolution operation. Subsequently, the images from the first two layers are upsampled with scale_factors of 4 and 2, respectively, to ensure that all three feature maps have the same size. These three feature maps are then fused and (through a linear layer and a bilinear interpolation layer) optimized during the training phase with the corresponding loss to obtain the final boundary feature map.

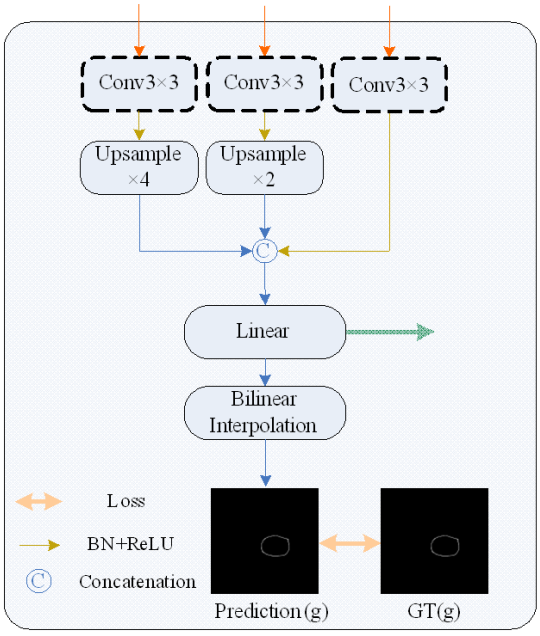

FIGURE 5. The designed boundary guidance (BG) network.

The ultimate boundary prediction is calculated as follows:

$$E = BI \{f[Cat(up(Conv(e_1), 4), up(Conv(e_2), 2), Conv(e_3))]\}, \quad (3)$$

where $e_i (i = 1, 2, 3)$ denote the three feature maps generated by the backbone network, $Conv()$ denotes a 3×3 convolution, $up()$ denotes the upsampling layer with its second parameter being the sampling rate, Cat denotes a concatenation operation, f denotes the linear layer operation, and BI denotes a bilinear interpolation.

E. BAFFM

To better fuse the skip-connection information with edge feature information and alleviate further the semantic gap between the shallow encoder and decoder, a newly designed module, called BAFFM, is proposed here, as shown in Figure 6a.

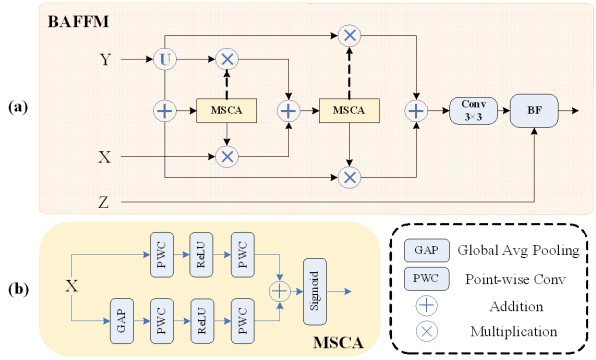

FIGURE 6. (a) The designed boundary attentional feature fusion module (BAFFM); (b) The MSCA block [33], utilized by BAFFM.

This module effectively fuses multi-level features using AFF that introduces Multi-Scale Channel Attention (MSCA) [33], shown in Figure 6b, which possesses strong adaptability to targets of different scales. MSCA consists of two branches – one branch employs global average pooling to capture global contextual information, while the other branch retains the original feature size to obtain local contextual information. Both branches utilize pointwise convolution operations to compress and restore features in the channel dimension, facilitating the fusion of multi-scale channel contextual information.

The AFF computation is performed as follows:

$$A = M(X \oplus up(Y)) \otimes X \oplus (1 - M(X \oplus up(Y))) \otimes up(Y), \quad (4)$$

$$AFF = Conv(M(A) \otimes X \oplus (1 - M(A)) \otimes up(Y)), \quad (5)$$

where M denotes MSCA, X denotes the feature maps generated by the main network, Y denotes the feature maps generated by skip connections, $up()$ denotes upsampling, $\oplus$ denotes the initial fusion of X and Y (Y is upsampled and pixel-wise added to X), $\otimes$ denotes element-wise multiplication, and $(1-M(X \oplus up(Y)))$ corresponds to the dashed line in Figure 6a. Finally, the features are passed through a 3×3 convolutional layer, followed by batch normalization and ReLU activation.

After AFF, the main features are already fused with the skip connection features. Then, the fused features are passed to a Boundary Fusion (BF) module, shown in Figure 7. The BF module has two inputs – one accepts the previously fused main features and the other accepts the edge features that are separately trained as described in Subsection IV-D. The edge prediction is used as a condition. Through this module,

spatial information is integrated into the feature maps, aiding in capturing boundary features more effectively.

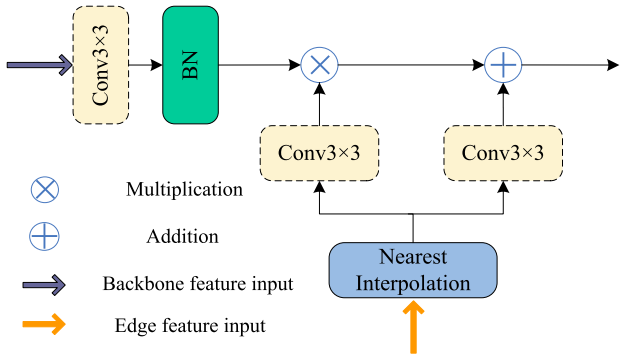

FIGURE 7. The BF module, utilized by BAFFM.

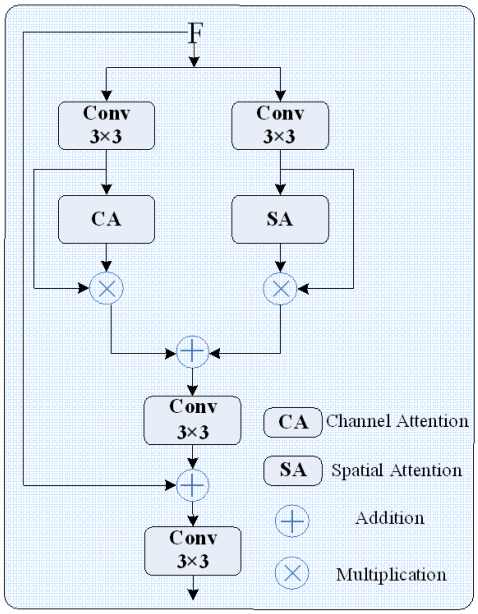

FIGURE 8. The designed parallel channel and spatial attention module (PCSAM).

The BF operation is defined as follows:

$$BF = CB\,(AFF) \otimes NC\,(E) \oplus NC\,(E)\,, \tag{6}$$

where CB denotes a 3×3 convolution followed by batch normalization (BN), AFF denotes the main features generated, NC denotes nearest-neighbor interpolation and the 3×3 convolutional layer, used for encoding the edge map information.

F. PCSAM

Inspired by the Convolutional Block Attention Module (CBAM) idea [34], in order to enhance the model's ability to analyze complex scene information, a newly designed module, called PCSAM, is proposed here, as shown in Figure 8.

In this module, first, the BAFFM output is passed in parallel through two 3×3 convolutions, feeding two branches. One branch employs channel attention to generate attention feature maps, which are then element-wise multiplied with the original input feature maps for adaptive feature refinement, producing the final feature maps. The other branch processes the feature information generated by spatial attention in parallel to the other branch, following the same steps. Subsequently, the feature maps produced by these two branches are pixel-wise added and then passed through a 3×3 convolution operation. These are further fused with the BAFFM output and, finally, another 3×3 convolution operation is applied to obtain the ultimate feature representation.

G. LOSS FUNCTIONS

The Binary Cross-Entropy (BCE) loss function [35] is widely used in many tasks, including image segmentation. BCE is well-suited for pixel-level binary classification and can directly measure the difference between each pixel's prediction and the true label, thereby encouraging the model to make accurate classifications. This helps achieve precise segmentation boundaries, enhancing the quality of segmentation results. Additionally, the BCE loss function is compatible with deep learning models and can be used in conjunction with CNNs, enabling a model to learn rich image features. As a result, it is widely popular in practical applications and provides an effective solution for segmentation problems. The BCE loss is calculated as follows:

$$L_{\text{BCE}} = -\sum_{i=1}^{N} [G_i \ln(P_i) + (1 - G_i) \ln(1 - P_i)], \tag{7}$$

where Gi denotes the value of pixel i in the ground-truth labels and Pi denotes the value of pixel i in the segmentation prediction results.

However, BCE has a notable drawback – when the number of target pixels is significantly lower than background pixels, the model tends to heavily bias towards the background, resulting in poor segmentation performance. To achieve better results, we combined the BCE loss function with the Dice loss function as proposed in [36].

The Dice loss function plays a crucial role in image segmentation. It is used to measure the similarity between the segmentation results and the ground-truth masks, encouraging the model to produce more precise segmentation boundaries.

The Dice loss function is robust to small objects and class imbalance issues, which helps improve segmentation performance. Additionally, it tends to lead to more stable model training and faster convergence, making it a preferred loss function for many segmentation tasks. Moreover, using the Dice loss function can reduce model overfitting. The Dice loss is calculated as follows:

$$L_{Dice} = 1 - 2\frac{\sum_{i=1}^{N} G_i P_i}{\sum_{i=}^{N} G_i^2 + \sum_{i=1}^{N} P_i^2} \tag{8}$$

The combined use of the BCE and Dice loss functions allows to leverage their respective advantages effectively.

This approach encourages accurate pixel classification, enhances segmentation boundary precision, manages class imbalance issues, and ensures stable convergence, thereby improving the quality of image segmentation. The combined BCE+Dice loss function is calculated as follows:

$$L_{BCE+Dice} = \frac{1}{2}L_{BCE} + L_{Dice} \tag{9}$$

Due to the fact that the proposed model has two supervised outputs, namely the segmentation map and the boundary map, the final overall loss function used is shown below:

$$L = \alpha L_t(P, G) + \beta(EP, EG), \tag{10}$$

where P denotes the segmentation prediction result, G denotes the ground-truth label image for segmentation, EP denotes the boundary prediction result, EG denotes the ground-truth boundary image, $\alpha = 0.8$, and $\beta = 0.2$.

IV. EXPERIMENTS AND RESULTS

A. DATASETS AND IMAGE PRE-PROCESSING

Two widely used public datasets were employed to evaluate the segmentation performance of the proposed BGRD-TransUNet model. The first dataset, referred to as Breast UltraSound Image (BUSI) dataset, was established by Al-Dhabyani et al. [37]. It consists of 780 BUS images with an average size of 500 × 500 pixels, accompanied by corresponding segmentation masks from 600 female patients, with 210 malignant cases, 437 benign cases, and 133 normal cases. These images were acquired at Bahaya Hospital using two different ultrasound devices – LOGIQ E9 and LOGIQ E9-Agile. The second dataset used, referred to as dataset B, was curated by Yap et al. [38]. It comprises 163 images (53 images with cancerous masses and 110 images with benign lesions) with an average size of 760 × 570 pixels, captured using the Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHz).

In the experiments, the BUSI dataset was used in two different ways: (i) including normal cases; and (ii) excluding normal cases. In both of these, we randomly split the BUSI dataset into training, validation, and test sets in an 8:1:1 ratio, as shown in Table 1. Due to the limited number of images contained in the dataset B, we used an 8:2 random split into training and validation sets, with no separate test set (Table 1). To address the data scarcity in dataset B, we applied data augmentation techniques, including random flips, random rotations, and random cropping, to the training set. The validation set remained unchanged during this process.

TABLE 1. Splitting the datasets in the experiments.

Dataset	Training set (images)	Validation set (images)	Test set (images)	Total images
BUSI	517	64	66	647
BUSI (incl. normal cases)	623	77	80	780
B	130	33	-	163

B. EVALUATION METRICS

In the experiments, five metrics were used to evaluate the model segmentation performance, including Intersection over Union (IoU), Dice Similarity Coefficient (DSC), recall, precision, and accuracy.

One of the most common metrics in semantic segmentation is IoU, a.k.a. Jaccard index, used to measure the degree of overlap between two regions, as follows:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{11}$$

where TP, FP, and FN respectively denote the accurate segmentation of breast lesions, incorrect segmentation of background regions as breast lesions, and incorrect segmentation of breast lesions as background regions.

DSC is the other widely used metric in the field of medical image segmentation. It measures the overlap between a model's segmentation results and the ground-truth labels, as follows:

$$DSC = \frac{TP}{2TP + FP + FN}, \tag{12}$$

Recall (Rec) is used to assess the proportion of true positives correctly identified by the model within the actual positive instances, which represents the model's coverage of the target region, as follows:

$$Rec = \frac{TP}{TP + FN}, \tag{13}$$

Precision (Pre) examines how many of the predictions made by a model are correct when it predicts positive instances, as follows:

$$Pre = \frac{TP}{TP + FP}. \tag{14}$$

Accuracy (Acc) measures the overall pixel-level segmentation performance, as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{15}$$

where TN denotes correct segmentation of background regions.

Together these five metrics form the confusion matrix, providing a comprehensive evaluation of the model segmentation performance.

C. EXPERIMENTAL ENVIRONMENT

Using PyTorch version 2.0.1 [40] and Python version 3.9.16, the experiments were conducted on a Windows 10 operating system. The hardware setup included a computer with a 13th Gen Intel® Core™ i5-13600KF CPU, 32GB of RAM, and an NVIDIA GeForce RTX 3060 GPU equipped with 12GB of RAM. Model training extended over 100 epochs, employing the Adam optimizer [41] with an initial learning rate of 1e-4. We adopted the CosineAnnealingLR scheduler [42], a

TABLE 2. Ablation study: compositions of components and corresponding versions of the developed model, along with their segmentation performance results.

Model Versions	Parameter Count (*million*)	Components							Evaluation Metrics				
		TransUNet	DensNet121 (used as a replacement of ResNet50 in TransUNet)	DASPPM	RMSFM	BD	BAFFM	PCSAM	IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)
v0 (baseline)	184.02	√							69.50	79.27	81.54	81.37	96.32
v1	**57.83**	√	√						72.21	81.90	85.99	81.94	96.56
v2	63.77	√	√	√					73.19	82.03	86.44	82.70	96.50
v3	86.17	√	√	√	√				74.82	82.91	87.34	83.60	96.70
v4	86.79	√	√	√	√	√			74.60	83.28	85.65	85.02	96.98
v5	94.11	√	√	√	√	√	√		74.90	83.73	**88.63**	82.64	96.81
v6 (*proposed*)	109.65	√	√	√	√	√	√	√	**76.77**	**85.08**	87.62	**85.89**	**97.14**

TABLE 3. Segmentation performance comparison of the proposed model with state-of-the-art models, based on experiments conducted on BUSI dataset and dataset B.

Models	Parameter Count (*million*)	BUSI dataset					Dataset B				
		IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)	IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)
U-Net [5]	7.85	62.21	72.93	69.41	83.53	95.53	72.24	81.23	79.35	85.74	97.79
FPN [39]	46.14	73.56	82.93	87.73	81.14	96.52	85.70	91.78	90.25	**94.47**	99.01
SegNet	29.44	65.33	75.78	75.87	82.78	95.79	76.99	86.08	84.31	89.60	98.32
DeepLabV3+	58.74	73.10	82.61	85.05	83.98	96.86	84.31	90.49	90.00	93.15	98.95
DoubleU-Net	29.29	72.44	82.02	**87.89**	80.76	96.42	81.92	89.17	88.64	91.34	98.64
UNet++ [6]	9.16	62.58	72.89	71.10	83.49	95.40	72.95	82.25	86.08	81.60	97.81
BiSeNetV2	5.10	67.00	77.33	82.16	77.94	95.86	76.76	85.48	87.84	86.29	98.45
DCSAU-Net	**2.59**	74.03	83.57	83.86	85.52	96.87	81.92	89.17	88.64	91.34	98.64
Attention-Unet [7]	57.16	61.98	72.62	70.00	83.15	95.51	70.96	81.23	80.16	85.97	97.82
TransUNet [9]	184.02	69.50	79.27	81.54	81.37	96.32	82.97	89.93	89.34	92.23	98.83
BGRD-TransUNet (*proposed*)	109.65	**76.77**	**85.08**	87.62	**85.89**	**97.14**	**86.61**	**92.47**	**92.78**	93.01	**99.15**

momentum of 0.9, and a batch size of 4. In terms of input image dimensions, all experiments utilized a consistent image size of 512 × 512 pixels.

D. ABLATION STUDY EXPERIMENTS

To assess the performance of different components of the developed model, we conducted ablation study experiments on the BUSI dataset, utilizing TransUNet as a baseline.

Table 2 provides a detailed breakdown of the various compositions of components resulting in different model versions, along with their segmentation performance results (the best results are shown in **bold**). The presented results indicate that the newly designed modules, presented in the previous section, all contributed positively to enhancing the segmentation performance of the developed model. As shown in Table 2, the addition of these modules (one after the other) to the baseline resulted in a gradual increase in all metrics, except for few cases of IoU (v4 vs. v3), recall (v4 vs. v1-3 & v6 vs. v5), precision (v5 vs. v2-4), and accuracy (v2 vs. v1 & v5 vs. v4). In particular, v1 (with DenseNet121 as a backbone) outperforms v0 (with ResNet50 as a backbone) according to all evaluation metrics, demonstrating that image segmentation with DensNet121 as a backbone performs better than when using ResNet50 as a backbone, which proves the feasibility of replacing the backbone network. After integrating all designed modules, the sixth version (v6) of the developed model, i.e., the proposed BGRD-TransUNet model, demonstrated an increase of 7.27 percentage points

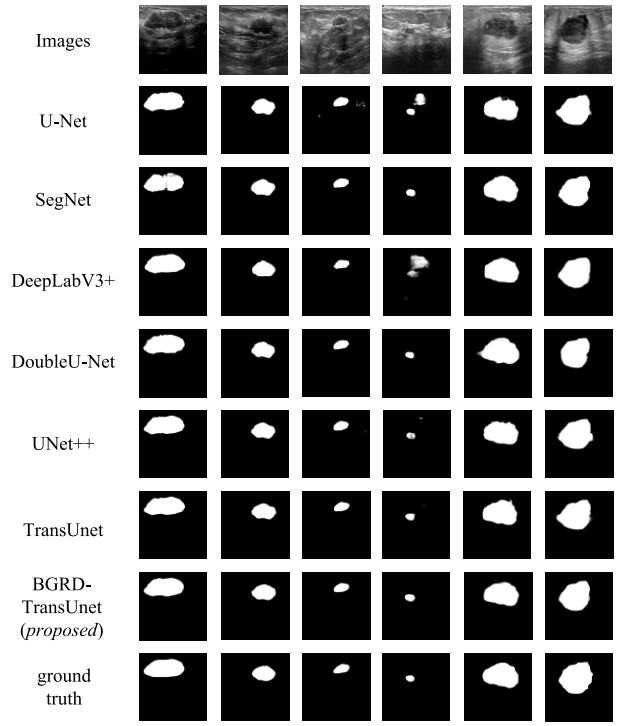

FIGURE 9. Illustration of segmentation performance comparison of the proposed model with state-of-the-art models, based on experiments conducted on BUSI dataset.

in IoU, 5.81 percentage points in DSC, 6.08 percentage points in recall, 4.52 percentage points in precision, and 0.82 percentage points in accuracy, compared to the baseline (TransUNet). This suggests that combining all these components together aids the model in learning more robust feature representations from BUS images.

Table 2 also shows the number of parameters of each version of the developed model as an indicator of its computational complexity. As can be seen, the parameter count gradually increases with each additional type of module added to the model. In the final version (v6) of the model, named BGRD-TransUNet, the number of parameters rose to 109.65 million. However, this number is still much smaller than the parameter count of the baseline (TransUnet), equal to 184.02 million.

E. SEGMENTATION PERFORMANCE COMPARISON EXPERIMENTS

Next, we compared the image segmentation performance of the proposed BGRD-TransUNet model with that of state-of-the-art models on both datasets, whereby BUSI was used for training, validation, and testing, while dataset B was used for training and validation only. The results obtained on the BUSI dataset are summarized in the left-hand part of Table 3 (the best results are shown in **bold**). As can be seen, BGRD-TransUNet outperforms all other models according to all metrics (including IoU and DSC which are the two most important and widely used metrics in the field of medical image segmentation), except for recall where it takes third place by closely following the leader (DoubleU-Net [43]) and first runner-up (FPN [39]). More specifically for all other four metrics, BGRD-TransUNet achieves values that are respectively higher by 2.74 percentage points for IoU, 1.51 percentage points for DSC, 0.37 percentage points for precision, and 0.27 percentage points for accuracy, compared to the second-best performing model (DCSAU-Net [44]). Visual comparisons of breast lesion segmentation results of the compared models are presented in Figure 9.

TABLE 4. Segmentation performance comparison of the proposed model with state-of-the-art models, based on BUSI dataset results reported in literature (years 2022-2023).

Models	IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)
AAU-Net [46]	-	77.51	81.10	79.61	-
AMS-PAN [2]	68.53	80.71	79.30	83.50	97.13
ATFE-Net [47]	69.73	82.46	82.78	-	96.32
CSwin-PNet [48]	75.11	83.68	85.87	85.71	-
DPCTN [49]	57.97	70.22	**87.97**	61.53	-
HCTNet [50]	-	82.00	82.14	83.24	96.94
NU-net [51]	-	78.62	82.46	79.56	-
U-Netmer [52]	-	74.82	76.47	-	95.96
BGRD-TransUNet (*proposed*)	**76.77**	**85.08**	87.62	**85.89**	**97.14**

The right-hand part of Table 3 contains image segmentation performance results of the proposed BGRD-TransUNet model and state-of-the-art models, obtained on dataset B (the best results are shown in **bold**). As can be seen, the proposed model outperforms all other models according to four (out of five) metrics, including IoU and DSC which are the two most important and widely used metrics in the field of medical image segmentation. The only exception is precision, where BGRD-TransUNet takes third place by closely following the leader (FPN [39]) and the first runner-up (DeepLabV3+ [45]). For the other four metrics, BGRD-TransUNet achieves values that are respectively higher by 0.91 percentage points for IoU, 0.69 percentage points for DSC, 2.53 percentage points for recall, and 0.14 percentage points for accuracy, compared to the second-best performing model (FPN [39]).

Table 3 also shows the parameter count of each model as an indicator of the model computational complexity. As can be observed, the proposed model has more parameters compared to other models, which is due primarily to addressing the demands of complex tasks, including the incorporation of the Transformer architecture and larger embedding dimensions. However, compared to the baseline (TransUNet), the proposed BGRD-TransUNet model has a much smaller parameter count.

Next, the proposed model was compared to other most recently proposed models according to their results achieved on the BUSI dataset, as reported in the corresponding literature sources, as shown in Table 4 (the best results are

TABLE 5. External validation results of segmentation performance of the proposed model and state-of-the-art models, trained on BUSI dataset and tested on datasets B and BUSI (including normal cases).

Models	Dataset B					BUSI dataset (including normal cases)				
	IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)	IoU (%)	DSC (%)	Rec (%)	Pre (%)	Acc (%)
U-Net [5]	55.09	65.85	62.34	**85.53**	97.54	51.41	61.23	60.13	70.09	96.07
SegNet [53]	60.14	71.16	74.51	77.29	**97.79**	53.19	61.87	63.05	66.33	95.90
DeepLabV3+ [45]	61.33	70.49	79.45	72.18	97.66	54.53	62.95	63.71	69.15	96.03
DoubleU-Net [43]	51.81	63.78	84.99	57.38	96.34	56.00	64.77	67.53	67.44	95.80
UNet++ [6]	53.98	64.11	62.33	85.37	97.47	50.83	60.22	59.46	67.63	95.77
BiSeNetV2 [54]	45.87	57.23	78.02	52.85	95.14	55.83	64.52	68.66	65.90	96.18
DCSAU-Net [44]	61.21	72.31	87.98	67.01	96.68	59.21	67.67	70.55	69.97	96.32
Attention-UNet [7]	54.13	63.82	61.16	78.45	97.54	51.46	60.35	58.80	69.60	96.02
TransUNet [9]	60.89	67.82	77.81	71.90	97.47	57.91	65.20	65.47	71.90	96.92
BGRD-TransUNet (*proposed*)	**64.92**	**75.27**	**91.41**	69.15	97.38	**69.76**	**71.54**	**74.27**	**72.96**	**97.52**

shown in **bold**). Again, the proposed model outperforms all other models according to four of the metrics (including IoU and DSC which are the two most important and widely used metrics in the field of medical image segmentation), except for recall, where BGRD-TransUNet takes second place by closely following the leader (DPCTN [49]). More specifically, for the other four metrics, compared to the second-best performing model, BGRD-TransUNet achieves values that are respectively higher by 1.66 percentage points for IoU (compared to CSwin-PNet [48]), 1.40 percentage points for DSC (compared to CSwin-PNet [48]), 0.18 percentage points for precision (compared to CSwin-PNet [48]), and 0.01 percentage points for accuracy (compared to AMS-PAN [2]).

F. ROBUSTNESS ANALYSIS EXPERIMENTS

Finally, we conducted robustness analysis experiments consisting of two parts: (i) external validation experiments; and (ii) model segmentation performance comparison experiments on the BUSI (normal) dataset.

Due to the significant differences in data collected from various sites, substantial variability among the acquired data may be present, [23]. This variability may lead to models performing well on the training set but experiencing a drop in performance when dealing with external data, [46]. So, in the external validation experiments, we used dataset B (without data augmentation) as an external data source to test the segmentation performance of models trained on the BUSI dataset. The obtained results are summarized in the left-hand part of Table 5 (the best results are shown in **bold**). As can be seen, the proposed model still demonstrates the best performance according to three (out of five) evaluation metrics (including the two most important in the field of medical image segmentation, i.e., IoU and DSC). At the same time, one can also notice that precision, achieved by BGRD-TransUNet in this experiment, is significantly lower. This could be attributed to the fact that the annotation information in the BUSI dataset and dataset B was provided by different sets of doctors. Subjective personal factors and different annotation focuses may have resulted in errors in the annotation of the edges of lesions in dataset B. In addition, the proposed BGRD-TransUNet model adds steps of boundary information extraction and fusion, which makes it focus more on the boundary labeling habits of the training dataset, making the problem of boundary labeling differences more prominent when validating the model on another dataset. Finally, the value of precision depends on TP and FP values, where TP represents the number of correct pixel predictions, and FP represents the number of incorrect pixel predictions. As can be deducted from (14), when the number of pixel prediction errors increases, this leads to low precision. The reason for the significant increase of FP could be due to errors in boundary prediction, which is also the fundamental reason for low precision. Other metrics, such as DSC and IoU, may be relatively tolerant to the accuracy of the predicted boundaries because these metrics consider the overlap between the predicted value and ground-truth value. Judging from these two metrics, the generalization performance of the proposed model is still excellent. However, precision is sensitive to boundaries. If a model shows errors near the predicted boundaries, precision will be low naturally.

Next, we conducted an assessment of the impact of ultrasound images, containing normal cases, on the model segmentation performance. The right-hand part of Table 5 contains the results obtained on the BUSI dataset (including normal cases). In comparison to the previous findings (c.f., Table 3), one can note a significant influence of introducing ultrasound images, containing normal cases, on the model segmentation performance. Similarly, Xue et al. [55] pointed out that including normal-case ultrasound images in the experiments is unfavorable for breast lesion segmentation. Nonetheless, the comparison between Table 3 and Table 5 indicates that regardless of the presence of normal-case ultrasound images, the proposed BGRD-TransUNet model demonstrates the best segmentation performance among the compared models in both cases – with and without inclusion of BUSI normal-case ultrasound images in the experiments. This suggests that the proposed model is capable, to a certain extent, of mitigating the interference caused, by surrounding tissues with similar intensity distributions.

V. CONCLUSION

This paper has introduced a novel TransUNet-based model, named BGRD-TransUNet, to address the challenges of breast lesion segmentation. BGRD-TransUNet starts by replacing the ResNet50 backbone network of TransUNet with DenseNet121. This substitution is aimed at mitigating the problem of model overfitting caused by small-size Breast UltraSound (BUS) image datasets. Secondly, to reduce the semantic gap between the shallow encoder and decoder, the proposed model employes newly designed Residual Multi-Scale Feature Modules (RMSFMs) in skip connections to extract features from various layers of DenseNet121, thus capturing richer features within specific layers. Thirdly, a newly designed Deformable Atrous Spatial Pyramid Pooling Module (DASPPM) is introduced between the encoder and decoder to enhance the extraction of complex shapes. Additionally, given that BUS images often have low resolutions, which can result in unclear lesion boundaries and affect the segmentation performance, a Boundary Guidance (BG) network was designed and used to enhance the contour information of BUS images for improved segmentation. In the decoder, the proposed model utilizes two newly designed types of modules. The first one is the Boundary Attentional Feature Fusion Module (BAFFM), which leverages Multi-Scale Channel Attention (MSCA) to fuse features from the main network, skip connections, and the BG network, thus further reducing the semantic gap between the shallow encoder and decoder. In each fusion process, a second type of modules, Parallel Channel and Spatial Attention Modules (PCSAM), combining both channel and spatial attention, is added to enhance the model's ability to analyze complex scene information.

Regarding the loss function, the Binary Cross-Entropy (BCE) and Dice loss functions were jointly used to address the issue of highly imbalanced positive and negative samples. Furthermore, multiple sets of experiments were conducted using two publicly available datasets. These experiments included an ablation study, image segmentation performance comparisons with state-of-the-art models, and robustness analysis. In the ablation study, the presented experimental results have demonstrated that the segmentation performance reached its peak when all newly designed components were integrated into the model. Additionally, the performance of the model improved with the addition of each individual component. In the performance comparisons with state-of-the-art models, it became evident that the proposed BGRD-TransUNet model outperformed all other models according to the majority of the evaluation metrics (including IoU and DSC which are the two most important and widely used metrics in the field of medical image segmentation), regardless of the dataset used. For the final robustness analysis, external validation experiments were conducted, along with additional model segmentation performance comparison experiments on the BUSI dataset (including normal cases), providing strong evidence that BGRD-TransUNet exhibits excellent generalization capabilities.

Even though the proposed model demonstrated superior results in breast lesion segmentation, there are still some aspects for improvement, as indicated by Tables 3–5, namely addressing the following limitations: (i) as the BG network is not yet perfect, the accurate obtaining of target contours still remains a challenging task; (ii) the quantity of available datasets is still limited, and they may not cover all types of breast lesions. To address these limitations, we plan to further enhance the BG network and incorporate foreground-background information to improve contour recognition. Additionally, we aim to explore more comprehensive datasets to enhance feature diversity and boost the model's generalization capabilities.

In conclusion, the proposed BGRD-TransUNet model is both feasible and effective, and holds the potential to serve as a reference point for further integration of artificial intelligence into early-stage breast cancer clinical diagnostics. Moving forward, our future research will focus on exploring the applicability of this model in other medical domains as well.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: Cancer J. Clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[2] Y. Lyu, Y. Xu, X. Jiang, J. Liu, X. Zhao, and X. Zhu, "AMS-PAN: Breast ultrasound image segmentation model combining attention mechanism and multi-scale features," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104425.

[3] J. Civit-Masot, F. Luna-Perejón, J. M. R. Corral, M. Domínguez-Morales, A. Morgado-Estévez, and A. Civit, "A study on the use of edge TPUs for eye fundus image segmentation," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104384.

[4] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a deep-learning-based method for breast ultrasound image segmentation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1103–1108.

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

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[7] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[8] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 327–331.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[10] G. Li, D. Jin, Q. Yu, and M. Qi, "IB-TransUNet: Combining information bottleneck and transformer for medical image segmentation," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 3, pp. 249–258, 2023.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–21.

[13] 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.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[15] R. Azad, E. Khodapanah Aghdam, A. Rauland, Y. Jia, A. Haddadi Avval, A. Bozorgpour, S. Karimijafarbigloo, J. Paul Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," 2022, *arXiv:2211.14830*.

[16] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3202–3211.

[17] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv-Yafo, Israel: Springer, 2022, pp. 205–218.

[18] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based deeplab v3+ for medical image segmentation," in *Proc. Int. Workshop Predictive Intell. Med.* Singapore: Springer, 2022, pp. 91–102.

[19] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2023, pp. 6202–6212.

[20] J. Xia, Y. Zhou, and L. Tan, "DBGA-Net: Dual-branch global–local attention network for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 7502305.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[22] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[23] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "SMU-Net: Saliency-guided morphology-aware U-Net for breast lesion segmentation in ultrasound image," *IEEE Trans. Med. Imag.*, vol. 41, no. 2, pp. 476–490, Feb. 2022.

[24] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8150–8159.

[25] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[26] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3907–3916.

[27] Z. Ning, K. Wang, S. Zhong, Q. Feng, and Y. Zhang, "CF2-Net: Coarse-to-fine fusion convolutional network for breast ultrasound image segmentation," 2020, *arXiv:2003.10144*.

[28] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "DA-TransUNet: Integrating spatial and channel dual attention with transformer U-Net for medical image segmentation," 2023, *arXiv:2310.12570*.

[29] J. Qin, Y. Huang, and W. Wen, "Multi-scale feature fusion residual network for single image super-resolution," *Neurocomputing*, vol. 379, pp. 334–342, Feb. 2020.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[31] M. Weber, H. Wang, S. Qiao, J. Xie, M. D. Collins, Y. Zhu, L. Yuan, D. Kim, Q. Yu, D. Cremers, L. Leal-Taixe, A. L. Yuille, F. Schroff, H. Adam, and L.-C. Chen, "DeepLab2: A TensorFlow library for deep labeling," 2021, *arXiv:2106.09748*.

[32] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, "I can find you! Boundary-guided separated attention network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 3608–3616.

[33] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 3560–3569.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[35] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, pp. 19–67, Feb. 2005.

[36] M. Montazerolghaem, Y. Sun, G. Sasso, and A. Haworth, "U-Net architecture for prostate segmentation: The impact of loss function on system performance," *Bioengineering*, vol. 10, no. 4, p. 412, Mar. 2023.

[37] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[38] 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 J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[42] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[43] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[44] Q. Xu, Z. Ma, H. E. Na, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106626.

[45] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[46] G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap, "AAU-Net: An adaptive attention U-Net for breast lesions segmentation in ultrasound images," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1289–1300, May 2023.

[47] Z. Ma, Y. Qi, C. Xu, W. Zhao, M. Lou, Y. Wang, and Y. Ma, "ATFE-Net: Axial Transformer and Feature Enhancement-based CNN for ultrasound breast mass segmentation," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106533.

[48] H. Yang and D. Yang, "CSwin-PNet: A CNN-Swin transformer combined pyramid network for breast lesion segmentation in ultrasound images," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119024.

[49] P. Song, Z. Yang, J. Li, and H. Fan, "DPCTN: Dual path context-aware transformer network for medical image segmentation," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106634.

[50] Q. He, Q. Yang, and M. Xie, "HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106629.

[51] S. Jin, S. Yu, J. Peng, H. Wang, and Y. Zhao, "A novel medical image segmentation approach by using multi-branch segmentation network based on local and global information synchronous learning," *Sci. Rep.*, vol. 13, no. 1, p. 6762, 2023.

[52] S. He, R. Bao, P. Ellen Grant, and Y. Ou, "U-netmer: U-Net meets transformer for medical image segmentation," 2023, *arXiv:2304.01401*.

[53] 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.

[54] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, Sep. 2021.
[55] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, and P.-A. Heng, "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101989.

ZHANLIN JI (Member, IEEE) received the M.Eng. degree from Dublin City University, in 2006, and the Ph.D. degree from the University of Limerick, Ireland, in 2010. He is currently a Professor with the North China University of Science and Technology, China, and an Associate Researcher with the Telecommunications Research Centre (TRC), University of Limerick. He has authored/coauthored more than 100 research papers in refereed journals and conferences. His research interests include ubiquitous consumer wireless world (UCWW), the Internet of Things (IoT), cloud computing, big data management, and data mining.

HAORAN SUN was born in 1999. He received the B.S. degree from the North China University of Science and Technology, in 2022, where he is currently pursuing the master's degree. His research interests include machine vision and graphic image processing.

NA YUAN received the bachelor's degree from Heilongjiang University, in July 2011, and the master's degree from the Hebei University of Technology, in January 2014. She is currently a Lecturer with Tangshan University. Her research interests include intelligent control, machine vision, and graphic image processing.

HAIYANG ZHANG received the B.S. degree from the School of Software Engineering, Jilin University, China, in 2013, and the Ph.D. degree from the Department of Electronic and Computer Engineering, University of Limerick, Ireland, in 2018.

She is currently a Lecturer with the Xi'an Jiaotong-Liverpool University, Suzhou, China. Her current research interests include recommender systems, data mining, collaborative filtering, and natural language processing.

JIAXI SHENG born in Tangshan, Hebei, in 1982. She is currently pursuing the Master of Medicine degree. She attending a Physician with the Endocrinology Department of the Affiliated Hospital of the North China University of Technology. She is good at the diagnosis and treatment of diabetes, acute and chronic complications of diabetes, thyroid diseases, adrenal diseases, pituitary, and other common diseases in the Endocrine Department. Her main research interests include endocrine and metabolic diseases.

XUEJI ZHANG is currently the Vice President of Shenzhen University and a Professor with the School of Biomedical Engineering, China. His research interests include the disciplines of chemistry, biology, materials, and medicine, with an emphasis on studies of biosensing, biomedicine, and biomaterials. He has been an editorial member of 24 international journals. He has received numerous national and international awards and honors, including a member of the Russian Academy of Engineering, a fellow of the American Institute for Medical and Bioengineering and the Royal Chemical Society, the National Innovation Award in China, the Scientist of the Year in China, and the Simon Fellow of ICSC-World Laboratory. He serves as the Co-Editor-in-Chief for *Sensors & Diagnostics*.

IVAN GANCHEV (Senior Member, IEEE) received the Engineering and Ph.D. degrees (summa cum laude) from the Saint Petersburg State University of Telecommunications, in 1989 and 1995, respectively. He is an International Telecommunications Union (ITU-T) Invited Expert and an Institution of Engineering and Technology (IET) Invited Lecturer, currently affiliated with the University of Limerick, Ireland, Plovdiv University "Paisii Hilendarski," Bulgaria, and the Institute of Mathematics and Informatics—Bulgarian Academy of Sciences, Bulgaria. He participated in more than 40 international and national research projects. He has served on the TPC of more than 400 prestigious international conferences/symposia/workshops and has authored/coauthored one monographic book, three textbooks, four edited books, and more than 300 research papers in refereed international journals, books, and conference proceedings. He serves on the editorial board and as a guest editor for multiple reputable international journals.