Journal of Artificial Intelligence for Medical Sciences
Vol 5 (1–2), June 2024, pp. 01–08
Journal home: http://2022.oapublishing-jaims.com/JAIMS
DOI: https://doi.org/10.55578/joaims.240226.001; eISSN: 2666-1470

RESEARCH ARTICLE

D-TransUNet: A Breast Tumor Ultrasound Image Segmentation Model Based on Deep Feature Fusion

Yiying Wan[1,†], Yaru Yang[1,†], Hongjiang Guo[1], Yangtian Yan[1], Tongtong Liu[1], Wenpei Liu[1], Yiru Wang[1], Wenhang Wang[1], Hao Dang[1,2,*]

[1]*School of Information Technology, Henan University of Chinese Medicine, Zhengzhou 450046, P.R. China*
[2]*Zhengzhou Key Laboratory of Intelligent Analysis and Utilization of Traditional Chinese Medicine Information, Zhengzhou 450046, P.R. China*

ARTICLE DATA

Article History

Received 28 November 2023
Revised 19 January 2024
Accepted 22 February 2024

Keywords
Breast tumor ultrasound image segmentation
Central dense connection
TransUNet
Deep learning

ABSTRACT

Breast ultrasound is a widely utilized modality for breast cancer screening since its noninvasive, radiation-free, low-cost, and easy-to-operate characteristics. The segmentation of breast tumor ultrasound images aims to accurately delineate the lesion area, thereby enhancing the usability and reliability of auxiliary diagnosis. In the realm of deep learning, U-Net and its variants based on fully convolutional networks have demonstrated outstanding performance in various medical image segmentation tasks. TransUNet also has achieved significant breakthroughs in medical image segmentation by introducing a global self-attention mechanism to overcome the limitations of traditional U-Net in handling long-range dependencies. In this paper, we propose a D-TransUNet to implement breast tumor ultrasound image segmentation. This model explores the introduction of a central dense connection module to more effectively fuse multi-scale features between the encoder and the upsampling. Finally, we conducted a series of comprehensive experiments on the BUSI dataset. The results demonstrate that D-TransUNet achieves an accuracy (Acc) of 0.9621, precision (Pre) of 0.9062, Recall of 0.9073, F1-score of 0.9033, mIoU of 0.8403, Dice of 0.8934, and 95% HD of 23.3299. These results show that the proposed method exhibits excellent accuracy in challenging scenarios, including complex shapes and blurred boundaries in breast tumor ultrasound image segmentation tasks, and providing a robust and reliable support for auxiliary breast cancer diagnosis.

1. INTRODUCTION

Breast cancer is one of the most prevalent malignant tumors in women, with its incidence ranking first among female malignancies [1]. Early diagnosis and precise segmentation of breast cancer are of paramount significance for the treatment and prognosis. Image segmentation technology plays a vital role in the diagnosis and treatment of breast cancer, aiding doctors in accurately locating and classifying tumor areas. However, the segmentation of breast cancer images encounters several challenges. Due to the complex structure and rich texture features of breast tissue, traditional image segmentation methods often struggle to accurately delineate tumor regions.

In recent years, the rapid advancement of deep learning has significantly improved the accuracy of semantic segmentation. Convolutional neural networks (CNN) have consistently held a dominant position in image analysis [2]. By progressively extracting semantic features from images, CNNs can comprehensively capture local spatiotemporal features, leading to high-precision detection and recognition. In 2015, Ronneberger et al. introduced a U-Net network based on fully convolutional neural networks (FCN) [3], which has profoundly influenced the development of the medical imaging field through its classic encoder-decoder structure and skip connections. Furthermore, it delivers precise and rapid segmentation results, making it well-suited for scenarios with a limited number of images. Almajalid et al. employed U-Net networks for breast cancer ultrasound image segmentation, resulting in segmentation images that were more accurate than previous methods [4]. Yap et al. presented an end-to-end solution, the fully convolutional network (FCN-AlexNet) [5], for identifying breast ultrasound lesions using deep learning techniques. Gour et al. developed a deep residual neural network model (DeepRNNetSeg) for automatic nucleus segmentation of histopathological breast cancer images [6]. Guan et al. were the first to proposed integrating the concept of dense connections into U-Net, creating a fully dense connected network FD-Unet [7]. In the retinal blood vessel segmentation task, Zhang et al. introduced the DenseInception U-Net, which integrates the residual idea, the Inception module, and dense connections [8].

In various medical image segmentation tasks, numerous improved versions of U-net networks have demonstrated strong

*Corresponding author. Email: danglee@hactcm.edu.cn
†These authors contributed equally to this work and should be considered co-first authors.

performance, including Unet++ [9], Unet3+ [10], Attention U-Net [11], and nnU-Net [12]. However, owing to the inductive bias, weight sharing, and local perception characteristics of convolution, CNN-based segmentation models lack the modeling ability for long-range dependency problems and exhibit limited spatial perception ability, hindering the further advancement of segmentation networks.

To tackle the aforementioned challenges, researchers have proposed Transformer models [13]. The Transformer effectively compensates for the limitations of convolution. Currently, some researchers have adapted the Transformer structure for the computer vision domain. Wang et al. introduced TransBTS in the brain tumor segmentation task, utilizing multi-scale information and the attention mechanism to improve the accuracy of breast tumor segmentation [14]. Building on the SwinTransformer concept [15], Cao et al. divided the network into multiple stages to generate features of different scales, proposing a pure Transformer U-shaped network Swin-Unet [16]. In specific experiments, this network outperforms traditional Transformer and convolution-combined networks. Chen et al. [17] proposed TransUNet, effectively improving the accurate segmentation of lesions in medical images by combining the global modeling capability of Transformer with the local feature extraction advantage of U-Net. Zhang et al. [18] proposed a parallel branch TransFuse network, incorporating parallel branches of Transformer and CNN architectures to simultaneously capture global dependencies and low-level spatial details. Wang et al. [19] presented UCTransNet based on the U-Net. By combining cross-channel transformer (CCT) and channel-wise cross attention (CCA) for feature fusion in the decoder, they achieved state-of-the-art results on multiple benchmark datasets for medical image segmentation in an end-to-end manner. Liu et al. [20] proposed the TransUNet+, drawing inspiration form UCTransNet and TransUNet. They designed a feature enhancement module to improve the features of the skip connection. To fuse multi-level feature information from the encoder and simultaneously address long-distance dependencies to bridge the semantic gap between the encoder and decoder, Li et al. [21] proposed the UCFilTransNet network. They designed the Cross-FilterTrans block in the skip connection to effectively alleviate issues related to semantic information loss and information asymmetry caused by continuous down-sampling. However, these transformer architectures also suffer from some limitations. First, compared to CNN, it is difficult to capture local context information. Therefore, we can reasonably assume that combining CNN to the benchmark transformer can enhance the extraction ability of feature. Moreover, the existing transformer model is actually composed of residual and normalization modules. There is no direct connection path between the final output layer and the previous transformer layers, and the gradient disappearance is prone to occur. Therefore, we consider that introducing additional dense connections to relieve this problem.

To address the aforementioned limitations, we proposed a new segmentation framework named D-TransUNet (introducing a Central Dense Block in TransUNet). Compared with the TransUNet, D-TransUNet mainly has the following advantages:

(1) Combining CNN and Transformer: D-TransUNet establishes an end-to-end network framework by integrating a Transformer encoder and a U-Net configuration. Within this framework, the CNN is tasked with extracting low-level features, while the Transformer encoder processes global context information. This combined approach fully leverages the advantages of the Transformer model to enhance the performance of medical image segmentation.

(2) Self-attention mechanism: The Transformer encoder in the D-TransUNet network incorporates self-attention mechanism, allowing it to adaptively learn relationships between different positions in the input image. This mechanism is effective in capturing long-distance dependencies within the image, thereby contributing to the improved accuracy of segmentation results.

(3) Skip Connections and Central Dense Block: In order to strengthen the information transmission and feature fusion between different layers, D-TransUNet combines skip connections with the proposed Central Dense Block. This integration effectively combines multi-scale features from the encoder with up-sampled features, leading to improved segmentation precision, particularly in terms of details and edges.

(4) Moreover, extensive experimental results on public breast ultrasound datasets show that D-TransUNet has better robustness in breast tumors segmentation.

2. RELATED WORK

2.1. The Combination of CNN and Transformer

In the field of computer vision, attention mechanisms are widely employed to capture crucial information in images. The traditional convolution method achieves feature extraction through the weight aggregation function on the local receptive field, and then distributing it across the entire feature map. In recent years, researchers have explored integrating attention mechanisms into CNNs to enhance feature representation. Among them, Hu et al. [22] proposed SE Net, applying the attention mechanism to the channel dimension attribute of the image and improving the model performance. Additionally, Woo et al. [23] introduced the spatial attention module SAM based on SE Net, incorporating spatial dimension modeling on top of channel attention. These studies show that the introducing attention mechanism can enhance the performance of convolution module. However, excessive reliance on attention mechanisms causes the model excessive sensitive and compromise its ability to perceive global feature. Therefore, it is necessary to explore how to reasonably use and integrate attention mechanisms to trade off the expression of local and global features for improved performance.

Transformer, initially proposed by Vaswani et al. [13] and applied to natural language processing, comprises stacked encoders and decoders with a multi-head self-attention mechanism and residual structure. To adapt Transformer for computer vision, Vision Transformer applies the self-attention mechanism to global images for image classification tasks [24]. Through the weighted average operation based on the input feature context, this method has achieved performance equivalent to or even superior to CNNs in numerous visual tasks. And then Chen et al. proposed a new network, TransUNet, establishing a self-attention mechanism from the perspective of sequence-to-sequence prediction. However, the inductive bias ability of Self-Attention is weaker than that of CNN and requires a substantial amount of data. Therefore, TransUNet integrates the CNN module. First, the image undergoes convolution to capture detailed high-resolution spatial information, and then the tokenized image block in the feature map is encoded as

an input sequence to extract the global context [5]. In medical image segmentation tasks, this approach shows superior performance, surpassing previous models.

2.2. Breast Cancer Ultrasound Image Segmentation

Breast cancer have surpassed lung cancer to become the cancer type with the highest global incidence [25]. Ultrasonography is a routine examination for breast diseases. However, ultrasound images of breast tumors face several challenges, including severe speckle noise, artifacts, low image resolution and contrast, and the intricate shapes of tumors.

Traditional automatic segmentation methods, including threshold segmentation, edge detection, and Markov random field, generally have average performance when dealing with ultrasound images characterized by low contrast and strong noise. In the pursuit of enhancing breast cancer ultrasound image segmentation, these methods based on traditional convolution networks has been widely employed. For instance, Cho et al. [26] proposed the Breast Tumor Integrated Classification Network (BTEC Net) for the classification of breast tumors in ultrasound images. It utilized the Residual Feature Selection UNet (RFS UNet) for exclusive segmentation of images with abnormal BTEC Net classification. Wu et al. [27] proposed a context level set method for breast tumor segmentation. Wang et al. [28] proposed the PDPNet, a progressive dual priori network specifically designed for segmenting breast tumors from dynamically enhanced images. Li et al. [21] designed the MultiIB-Transformer structure within the MultiIB-TransUNet. This structure is composed of a single Transformer layer and multiple Information Bottleneck (IB) blocks. It serves to reduce the number of model parameters and has demonstrated good performance on breast cancer datasets.

In the realm of breast cancer ultrasound image segmentation research, the high-efficiency extraction of edge features is the key to improving model performance. The fusion of multi-scale local context and global context features, theoretically leads to better segmentation accuracy. Therefore, in the exploration of breast cancer ultrasound image segmentation methods, the improvement of edge features remains a promising direction to improve segmentation accuracy.

3. METHODS

3.1. D-TransUNet Model Architecture

The proposed D-TransUNet network is depicted in Figure 1. This network comprises three main modules: the CNN-Transformer hybrid encoding module, the central dense connection module, and the decoder module. For an input image, the feature extraction begins by feeding images into the hybrid encoder module. Subsequently, the Transformer module encodes the feature maps into an input sequence for extracting global context information. Then, the central dense connection module facilitates multi-scale features fusion depending on dense connections. Finally, the upsample module is employed to upsample the encoded features and combine them with high-resolution features from the encoder. This results in a lightweight end-to-end U-shaped network structure that effectively leverages local context information from the feature extraction stage, enhances the receptive field in the upsampling stage, and improves detail preservation. By combining these three modules and examining the segmentation results, it's evident that D-TransUNet achieves superior edge prediction and outperforms other networks in terms of segmentation performance.

3.2. Hybrid Encoder

The encoder section initially with three layers of convolutional downsampling on the input image. It embeds image patches extracted from the CNN features and adds positional encoding. And then the embedded patches, as one-dimensional vectors, are input into a 12-layer Transformer structure. Concretely, an improved ResNet-50 and Vision Transformer (ViT) serve as the backbones for CNN and Transformer, respectively. The convolutional layers utilize StdConv2d, replacing the original BatchNorm with GroupNorm. The original ResNet-50 structure contains stage1 and stage2, while stage3 and stage4 are merged into a new stage3 in the enhanced ResNet-50, resulting in three stage1, four stage2, and nine stage3 components.

As illustrated in Figure 2, the internal processing of the hybrid encoder involves dimension changes. After the Stem operation, the resolution becomes 1/4 of the original image, transforming from [512, 512, 3] to [108, 108, 64]. The Stage1 operation maintains the resolution, changing from [108, 108, 64]

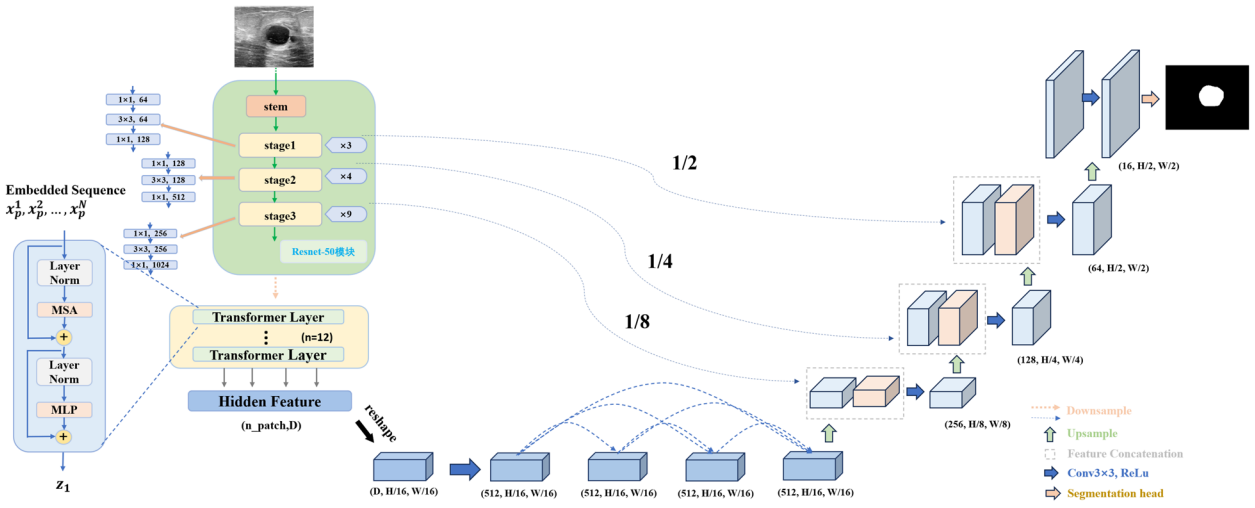

Figure 1 | D-TransUNet network model.

to [108, 108, 256]. Stage2 reduces the resolution to 1/8, altering from [108, 108, 256] to [56, 56, 512]. Stage3 further reduces the resolution to 1/16, converting from [56, 56, 512] to [28, 28, 1024]. Subsequently, dimension reduction is performed by a 1×1 convolution, resulting in a serialized input to the Transformer: [28, 28, 1024] → [28, 28, 768] → [784, 768]. Here, the dimensions correspond to the sequence required by the Transformer, transforming from [*H*, *W*, 3] to $\left[\frac{H\times W}{P^2}, P^2\times C\right]$, where *X* represents the sequence length. Finally, after Patch Embedding and positional encoding, the input is processed through Patch Embedding, yielding [784, 768] × [768, 768] → [784, 768], or $\left[\frac{H\times W}{P^2}, P^2\times C\right]\times[P^2\times C, \mathrm{D}]\rightarrow\left[\frac{H\times W}{P^2}, D\right]$, which is then fed into the Transformer module for global feature extraction. The entire process is summarized in Table 1.

3.3. Central Dense Connection

Similar to TransUNet, skip connections are employed to fuse multi-scale features from the encoder with upsampled features. Shallow and deep features are concatenated to alleviate spatial information loss caused by downsampling. However, downsampling in the encoding part only captures shallow features of the image. Directly fusing these features with the final layer of upsampled features can result in a loss of crucial information. Therefore, this paper introduces the central dense connection module to fill the core of the TransUNet network architecture, ensuring connections between each scale in the horizontal direction.

As shown in Figure 3, in this module, the input of each convolution layer encompasses the outputs of all previous convolution layers, which achieve comprehensive feature reuse. The fusion of high and low-level features enhances the network resistance to overfitting. After concatenating two feature maps, a 1×1 convolution is used to halve the number of channels while retaining the same number of channels as in the pre-concatenation feature maps. The paper introduces four central dense connections.

In deep learning networks, as the network depth increases, the problem of gradient disappearance becomes more pronounced, and high-level features obtained in the deep layers may suffer from insufficient precision. Therefore, introducing central dense modules at the lowest layer of the TransUNet network is crucial to improving the convolutional layer structure of conventional U-shaped network. Dense connections establish multiple connections that span distant front and back layers, combining both long and short connection strategies. In recent years, many researchers have introduced dense connection mechanisms in U-Net. For instance, Li et al. combined U-Net and dense skip connections in the nested segmentation network (attention-based Nested U-Net, ANU-Net) to obtain full-resolution feature maps at different semantic levels. Numerous experiments indicate that adding dense connection modules can enhance feature fusion.

3.4. Decoder

The decoding section introduces the Cascade Upsampler (CUP), which consists of multiple upsampling blocks and multi-level skip connections for decoding hidden features and obtaining the final segmentation mask. Upon reconstructing the hidden feature sequence into $\frac{H}{P}\times\frac{W}{P}\times D$, the complete resolution restoration of $\frac{H}{P}\times\frac{W}{P}\times D$ to $H\times W$ is accomplished by cascading multiple upsampling blocks and incorporating skip connections.

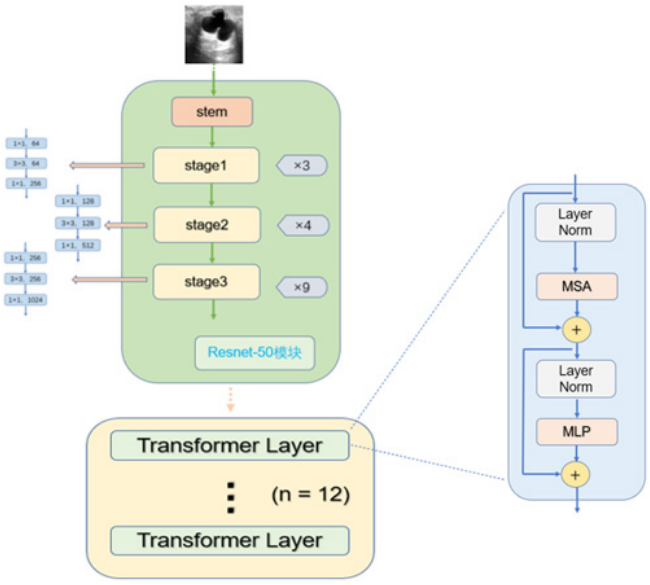

Figure 2 | Internal structure of the hybrid encoder.

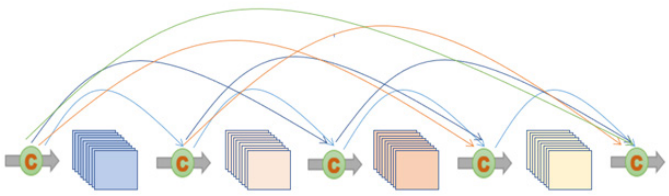

Figure 3 | Central dense connection module.

Table 1 | **Hybrid encoder processing flow**

Step	Processing	Input	Output
stem	3-layer convolutional downsampling	[512, 512, 3]	[108, 108, 64]
stage1	ResNet-50 stage1	[108, 108, 64]	[108, 108, 256]
stage2	ResNet-50 stage2	[108, 108, 256]	[56, 56, 512]
stage3	ResNet-50 stage3 and stage4	[56, 56, 512]	[28, 28, 1024]
Convolutional downsampling	conv 1×1	[28, 28, 1024]	[28, 28, 768]
Serialization input	Serialize for transformer input	[28, 28, 768]	[784, 768]
Patch embedding	Embed patches	[784, 768] × [768, 768]	[784, 768]

Each upsampling block is sequentially composed of feature map concatenation, convolution functions, and ReLU activation functions, with the goal of restoring the full resolution of $H \times W$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The experiment is conducted on a server environment with an NVIDIA RTX 3090 GPU (24GB), Python 3.8, PyTorch 1.11.0, and CUDA 11.3. To expedite the training process, adaptive learning rate and optimal gradient descent techniques are employed to ensure the fast and smooth decrease in loss. The input image size is configured as 512×512, with a patch size set to 16. For achieving full resolution, the decoder part utilizes four convolution upsampling blocks in TransUnet. The framework adopts D-TransUnet and incorporated pre-trained weights from ViT trained on the ImageNet dataset. Regarding hyperparameter settings, the training spans 50 epochs, the initial learning rate is set at 1E-4, and the batch size is 8.

Breast cancer tumor lesions are categorized as either benign or malignant. Benign tumor ultrasound images typically feature relatively smooth lesion areas with a relatively regular distribution of semantic features. In contrast, malignant tumor ultrasound images often exhibit irregular lesion distributions characterized by fuzzy boundaries and uneven brightness. For achieving accurate segmentation between benign and malignant tumors, it is vital to consider the different boundaries feature. Hence, this study employs the Dice loss function to refine the segmentation results for lesions with fuzzy boundaries, with the goal of improving accuracy and predictive performance. The specific form of the Dice loss function is as follows:

$$L_{\text{Dice}} = 1 - \frac{2\sum y_{\text{pred}} \times y_{\text{true}}}{\sum_{\text{pix}} y_{\text{pred}}^2 + \sum_{\text{pix}} y_{\text{true}}^2} \tag{1}$$

where y_{pred} represents the predicted pixel probability values, and y_{true} is the actual label values. The Dice coefficient ranges from 0 to 1, with higher values indicating more accurate predictions.

4.2. Dataset

The paper leverages the BUSI dataset (Breast Ultrasound Images), which was compiled by Al-Dhabyani W in 2020 [29]. This publicly available dataset consists of 780 breast ultrasound images from females aged 25 to 75 years, with an average image size of 500×500 pixels. Among these images, there are 133 normal images without lesions, 437 images with benign lesions, and 210 images with malignant lesions. It's worth noting that some benign and malignant samples may contain two or more lesions.

To evaluate the network's performance, this experiment exclusively utilizes samples with benign and malignant data. The dataset is partitioned into training and testing sets in a 17:3 ratio. The training set, which undergoes data augmentation techniques such as rotation, mirroring, and brightness enhancement, is further split into a training subset and a validation subset in a 5:1 ratio. The final dataset comprises 4199 images in the training set, 734 images in the validation set, and 98 images in the test set.

4.3. Evaluation Metrics

To evaluate the segmentation performance of the model, we compare the performance of segmentation in breast ultrasound research by using Accuracy (Acc), Precision (Pre), Recall, balanced F-score, Dice Similarity Coefficient (Dice), Mean Intersection over Union (mIoU), and Hausdorff distance (HD95). They are defined in Eqs. (2)–(8).

$$\text{Acc} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}} \tag{5}$$

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{7}$$

$$\text{HD}(A, B) = \max \max_{a \in A} \min_{b \in B} d(a, b), \max_{b \in B} \min_{a \in A} d(a, b) \tag{8}$$

The definition of correctly segmented lesion areas is termed True Positive (TP), correctly segmented normal tissue regions are termed True Negative (TN), normal tissue regions segmented as lesion areas are defined as False Positive (FP), and lesion areas segmented as normal regions are defined as False Negative (FN). Here, N represents the total number of classes, and d (a, b) represents the Euclidean distance between points a and b.

4.4. Results

D-TransUNet is a novel model based on TransUNet that introduces dense connections to the bottom layer of the original network. This effectively increases feature reuse, leading to improve the segmentation accuracy for ultrasound images. To demonstrate the reliability of the proposed model, six groups of experiments are compared with those of U-Net, U-Net++, Attention U-Net, and Swin-UNet. As depicted in Figure 4, the segmentation results of D-TransUNet are significantly more accurate than those of the other models. Particularly, D-TransUNet show superiority in capturing boundary information of the lesion area, closely aligning with the ground truth. This observation suggests that the proposed D-TransUNet enables extract efficient features from multiple scales and levels, enhancing the ability of boundary perception, and capturing more details information of the lesion area.

The quantitative results of the model are presented in Table 2. As observed from the table, the Pre, F1-score, mIoU, and Dice coefficients of D-TransUNet excess those of other methods (including U-Net, U-Net++, Attention U-Net, Swin-UNet, and TransUNet). Hausdorff Distance (HD95) measures the maximum distance between the boundary generated by segmentation and the true subset of the actual tumor area. The evaluation index of proposed method reaches 23.3299 lower than other results. This indicates that D-TransUNet is closer to

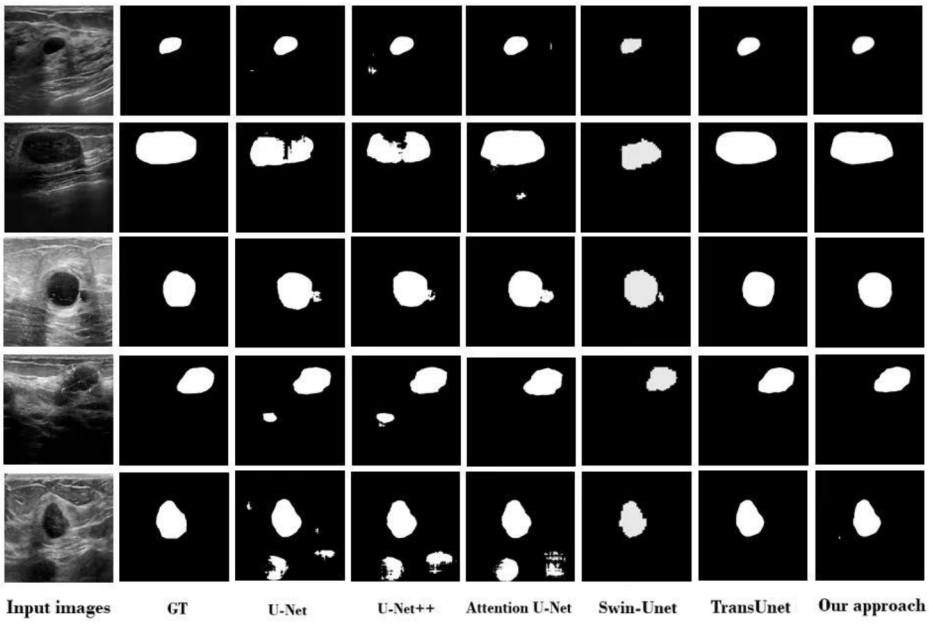

Figure 4 | The segmentation comparison results of SOTA models.

Table 2 | The segmentation results of different models in dataset BUSI

Model	Acc	Pre	Recall	F1-score	mIoU	Dice	HD95
U-Net [8]	0.9503	0.8730	0.8438	0.8516	0.7686	0.8353	85.8279
U-Net++ [9]	0.9507	0.8542	0.8688	0.8544	0.7754	0.8346	58.5642
Attention U-Net [11]	0.9430	0.8763	0.8196	0.8383	0.7409	0.8129	97.3960
Swin-UNet	**0.9656**	0.8857	**0.9093**	0.8946	0.8283	0.8858	26.2631
TransUNet	0.9501	0.8820	0.8515	0.8570	0.7630	0.8322	38.6687
D-TransUNet	0.9621	**0.9062**	0.9073	**0.9033**	**0.8403**	**0.8934**	**23.3299**

the ground truth in overall segmentation performance than other models and can more accurately match the overlap between the tumor area and the ground truth. Additionally, the performances of D-TransUNet are superior on the details and boundaries of the tumor area to other approaches. Concretely, the evaluation indicators of Acc, Pre, Recall, F1-score, mIOU, and Dice respectively are 0.9621, 0.9062, 0.9073, 0.9033, 0.8403, and 0.8934, which is 1.2%, 2.42%, 5.58%, 4.63%, 7.73%, and 6.12% higher than TransUNet. The demonstrates that the dense connections can effectively improve the performance of breast tumor segmentation of TransUNet. In brief, relying on the advantage of capturing the shape and boundary of the tumor, D-TransUNet can provide more accurate segmentation results, holding potential clinical value for tasks such as medical image analysis and pathological diagnosis.

According to the aforementioned analysis, the segmentation performance of D-TransUNet is notably closer to the ground truth. It is evident that D-TransUNet is proven to be feasible, allowing it more effectively focus on the tumor area through integrating the main features from different lays.

5. CONCLUSION

This paper investigated the limitations of transformer method and a comprehensive evaluation of the most representative tumor segmentation approaches for breast ultrasound. And then a breast tumor ultrasound image segmentation method based on TransUNet is proposed. The model called D-TransUNet is improved by introducing dense connections to avoid the disappearance of gradient caused by deep network architecture. The model achieves deep feature fusion by inputting feature maps into dense blocks. The experimental results on the BUSI dataset also showcase the proposed method can realize the analysis of actual complex breast tumor ultrasound image and lay the foundation for extending its application to practical scenarios.

DECLARATIONS

Competing Interests

The authors declare that they have no competing interests.

Author Hao Dang is a member of the Editorial Board of Journal of Artificial Intelligence for Medical Sciences. The paper was handled by another Editor and has undergone a rigorous peer review process. Author Hao Dang was not involved in the journal's peer review of, or decisions related to, this manuscript.

Authors' Contribution

HD designed the work; WYY achieved the method; YYR and GHJ accomplish the experiments; YYT, LTT, and LWP accomplish the writing assignment; WYR drafted the figures and tables about this paper; WWH substantively revised it.

Acknowledgments

All authors are acknowledged for their valuable contributions to this research. On behalf of the entire authorship, our sincere gratitude is extended to National Cancer Institute in Cairo University, Egypt for granting permission to utilize their data in this study.

Availability of Data and Materials

The datasets analysed during the current study are available in the SNLI repository, https://scholar.cu.edu.eg/?q=afahmy/pages/dataset.

Funding

This work is partially supported by the Key Research and Development Project (Science and Technology Development) of Henan Province under Grant 222102210028.

REFERENCES

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, CA Cancer J. Clin. 73 (2023), 17–48.

[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS, NY, United States, 2012, pp. 1097–1105.

[3] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, In: N. Navab, J. Hornegger, W. Wells, A. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Lecture Notes in Computer Science, Vol. 9351, Springer, Cham, 2015, pp. 234–241.

[4] R. Almajalid, J. Shan, Y. Du, M. Zhang, Development of a deep-learning-based method for breast ultrasound image segmentation, 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Orlando, FL, USA, 2018, pp. 1103–1108.

[5] M.H. Yap, M. Goyal, F.M. Osman, R. Martí, E. Denton, A. Juette, et al., Breast ultrasound lesions recognition: end-to-end deep learning approaches, J. Med. Imaging. 6 (2019), 011007.

[6] M. Gour, S. Jain, R. Agrawal, DeepRNNetSeg: deep residual neural network for nuclei segmentation on breast cancer histopathological images, In: N. Nain, S. Vipparthi, B. Raman (Eds.), Computer Vision and Image Processing, CVIP 2019, Communications in Computer and Information Science, Vol. 1148, Springer, Singapore, 2019, pp. 243–253.

[7] S. Guan, A.A. Khan, S. Sikdar, P.V. Chitnis, Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal, IEEE J. Biomed. Health Inform. 24 (2020), 568–576.

[8] Z. Zhang, C. Wu, S. Coleman, D. Kerr, DENSE-INception U-net for medical image segmentation, Comput. Methods Programs Biomed. 192 (2020), 105395.

[9] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, In: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, et al., (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, DLMIA ML-CDS 2018, Lecture Notes in Computer Science, Vol. 11045, Springer, Cham, 2018, pp. 3–11.

[10] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., UNet 3+: A full-scale connected UNet for medical image segmentation, ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 2020, pp. 1055–1059.

[11] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention U-Net: learning where to look for the pancreas, arXiv preprint, arXiv:180403999, 2018.

[12] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P.F. Jaeger, S. Kohl, et al., nnU-Net: self-adapting framework for U-Net-based medical image segmentation, In: H. Handels, T. Deserno, A. Maier, K. Maier-Hein, C. Palm, T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2019, Informatik aktuell, Springer Vieweg, Wiesbaden, 2019, p. 22.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, NY, United States, 2017, pp. 6000–6010.

[14] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, TransBTS: multimodal brain tumor segmentation using transformer, In: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, et al., (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Lecture Notes in Computer Science, Vol. 12901, Springer, Cham, 2021, pp. 109–119.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 10012–10022.

[16] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-Unet: Unet-like pure transformer for medical image segmentation, In: L. Karlinsky, T. Michaeli, K. Nishino (Eds.), Computer Vision–ECCV 2022 Workshops, ECCV 2022, Lecture Notes in Computer Science, Vol. 13803, Springer, Cham, 2023, pp. 205–218.

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., TransUNet: transformers make strong encoders for medical image segmentation, arXiv preprint, arXiv:2102.04306, 2021.

[18] Y. Zhang, H. Liu, Q. Hu, TransFuse: fusing transformers and CNNs for medical image segmentation, In: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, et al., (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Lecture Notes in Computer Science, Vol. 12901, Springer, Cham, 2021, pp. 14–24.

[19] H. Wang, P. Cao, J. Wang, O.R. Zaiane, UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer, Proceedings of the AAAI Conference on Artificial Intelligence, 36 (2022), 2441–2449.

[20] Y. Liu, H. Wang, Z. Chen, K. Huangliang, H. Zhang, TransUNet+: redesigning the skip connection to enhance features in medical image segmentation, Knowl. Based Syst. 256 (2022),109859.

[21] L. Li, Q. Liu, X. Shi, Y. Wei, H. Li, H. Xiao, UCFilTransNet: cross-filtering transformer-based network for CT image segmentation, Expert Syst. Appl. 238 (2024), 121717.

[22] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[23] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision–ECCV 2018, Lecture Notes in Computer Science, Vol. 11211, Springer, Cham, 2018, pp. 3–19.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929, 2020.

[25] W. Cao, H.D. Chen, Y.W. Yu, N. Li, W.Q. Chen, Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020, Chin. Med. J. 134 (2021), 783–791.

[26] S.W. Cho, N.R. Baek, K.R. Park, Deep learning-based multistage segmentation method using ultrasound images for breast cancer diagnosis, J. King Saud Univ. Comput. Inf. Sci. 34 (2022), 10273–10292.

[27] S. Hussain, X. Xi, I. Ullah, Y. Wu, C. Ren, Z. Lianzheng, et al., Contextual level-set method for breast tumor segmentation, IEEE Access. 8 (2020), 189343–189353.

[28] L. Wang, L. Wang, Z. Kuai, L. Tang, Y. Ou, C. Ye, et al., Progressive dual priori network for generalized breast tumor segmentation, arXiv preprint, arXiv:2310.13574, 2023.

[29] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data in Brief. 28 (2020), 104863.