Mathematical Biosciences and Engineering

MBE, 20(10): 18173–18190.
DOI: 10.3934/mbe.2023807
Received: 15 May 2023
Revised: 29 May 2023
Accepted: 30 May 2023
Published: 21 September 2023

http://www.aimspress.com/journal/MBE

Research article

TS-GCN: A novel tumor segmentation method integrating transformer and GCN

Haiyan Song[1,*], Cuihong Liu[2,3,*], Shengnan Li[1,*] and Peixiao Zhang[1]

[1] The Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China
[2] Affiliated Eye Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China
[3] School of Nursing, Shandong University of Traditional Chinese Medicine, Jinan, China.

* **Correspondence:** Email: shy5646@163.com, semlch@163.com, snqsh@163.com; Tel: 13954195646.

Abstract: As one of the critical branches of medical image processing, the task of segmentation of breast cancer tumors is of great importance for planning surgical interventions, radiotherapy and chemotherapy. Breast cancer tumor segmentation faces several challenges, including the inherent complexity and heterogeneity of breast tissue, the presence of various imaging artifacts and noise in medical images, low contrast between the tumor region and healthy tissue, and inconsistent size of the tumor region. Furthermore, the existing segmentation methods may not fully capture the rich spatial and contextual information in small-sized regions in breast images, leading to suboptimal performance. In this paper, we propose a novel breast tumor segmentation method, called the transformer and graph convolutional neural (TS-GCN) network, for medical imaging analysis. Specifically, we designed a feature aggregation network to fuse the features extracted from the transformer, GCN and convolutional neural network (CNN) networks. The CNN extract network is designed for the image's local deep feature, and the transformer and GCN networks can better capture the spatial and context dependencies among pixels in images. By leveraging the strengths of three feature extraction networks, our method achieved superior segmentation performance on the BUSI dataset and dataset B. The TS-GCN showed the best performance on several indexes, with Acc of 0.9373, Dice of 0.9058, IoU of 0.7634, F1 score of 0.9338, and AUC of 0.9692, which outperforms other state-of-the-art methods. The research of this segmentation method provides a promising future for medical image analysis and diagnosis of other diseases.

Keywords: medical image segmentation; medical image processing; transformer; graph convolutional

networks; breast tumor diagnosis

1. Introduction

Breast cancer is a common and serious health concern for women worldwide, resulting in increased mortality rates. Medical imaging, such as ultrasound, mammography and MRI, is a widely used tool for breast cancer detection and diagnosis [1,2]. Various features such as texture and smoothness captured through ultrasound scans can help to identify abnormalities in breast tumors. Manually analyzing ultrasound scans and distinguishing abnormal from normal breast tissue can be challenging and time-consuming, leading to delays in the diagnosis process [3]. However, segmenting small tumors in ultrasound images is challenging due to low-resolution scans, varying tumor shapes and sizes, and the presence of noise in the images. Therefore, automatic segmentation of tumor regions using computer-aided diagnosis systems is crucial for the early detection of cancer and for reducing mortality rates [4–6]. Traditional image segmentation methods have played a significant role in the field of medical imaging. These methods primarily rely on handcrafted features and heuristics to delineate the boundaries of structures or regions of interest in an image. Several well-established techniques fall under this category, including thresholding, edge-based methods, region growing, and watershed segmentation. They may struggle with handling irregular shapes, leading to inaccuracies segmentation results. However, with the advancements in deep learning, learning-based methods have gained prominence due to their ability to automatically learn complex features and adapt to diverse imaging conditions. Huang [7] proposed a fuzzy FC network to perform ultrasound image segmentation. Lei et al. proposed a boundary-regularized deep convolutional encoder-decoder network to alleviate the challenge of segmenting whole breast ultrasound images [8]. Therefore, it is essential to develop targeted tumor segmentation schemes that take into account the unique characteristics of the images.

Deep learning-based computer-aided diagnosis systems are developed for the early detection of breast tumors for faster diagnosis and treatment, especially for the detection of ultrasound images. U-Net is a popular CNN-based segmentation framework that has shown impressive performance in this regard, with several studies comparing its efficacy to other methods [9,10]. Wang et al. used deep supervision strategy constraints on the feature maps captured at each stage of U-net to segment breast lesions [11]. However, the existing mainstream methods cannot effectively extract the features of the small lesion. When segmenting breast tumors, CNN-based methods do not need precise image feature definitions, in contrast to conventional feature-oriented methods. Cheng et al. [18] presented the deepest semantically guided multi-scale feature fusion network (DSGMFFN), the SC-attention module is meant to incorporate both rich semantic information and finer-grained spatial information to reduce performance deterioration brought on by ambiguous boundaries and different tumor sizes. However, many existing segmentation methods rely on generic frameworks that may not effectively extract lesion information from ultrasound images or discriminate between relevant and irrelevant features.

Inspired by human visual attention, many attention algorithms have been developed to strengthen the representation ability of CNNs [12–16]. The Transformer is a type of neural network architecture that was originally introduced for natural language processing tasks [17]. However, its self-attention mechanism has since been successfully applied to many other domains, including image processing. The self-attention mechanism allows the network to selectively focus on different parts of the input,

which is particularly useful for tasks such as object detection or segmentation [18]. This can result in feature redundancy and a loss of discriminant features, hindering the accuracy of the tumor segmentation process. But they still face some challenges, such as the imbalance of data distribution and the inability to capture the spatial relations among pixels. Among these methods, the transformer and graph convolutional neural network (GCN) have attracted considerable attention due to their ability to capture global and local features of images, respectively [19–21]. GCN is a powerful model for semantic image representation, which can encode structural information by featuring a pre-defined adjacency matrix [22–24]. Huang et al. proposed a boundary-rendering network for breast lesions segmentation by a differentiable boundary selection module and a GCN-based boundary rendering module [25]. As GCN models are based on fixed-size graph structures, they may fail to effectively capture the features of tumor regions at different scales. To visualize our overall research idea conveniently. The block diagram of the overall research route is shown in Figure 1, which contains in order the dataset, data pre-processing, automatic segmentation method design, the segmentation module optimization and end-to-end training, and the final output segmentation results.

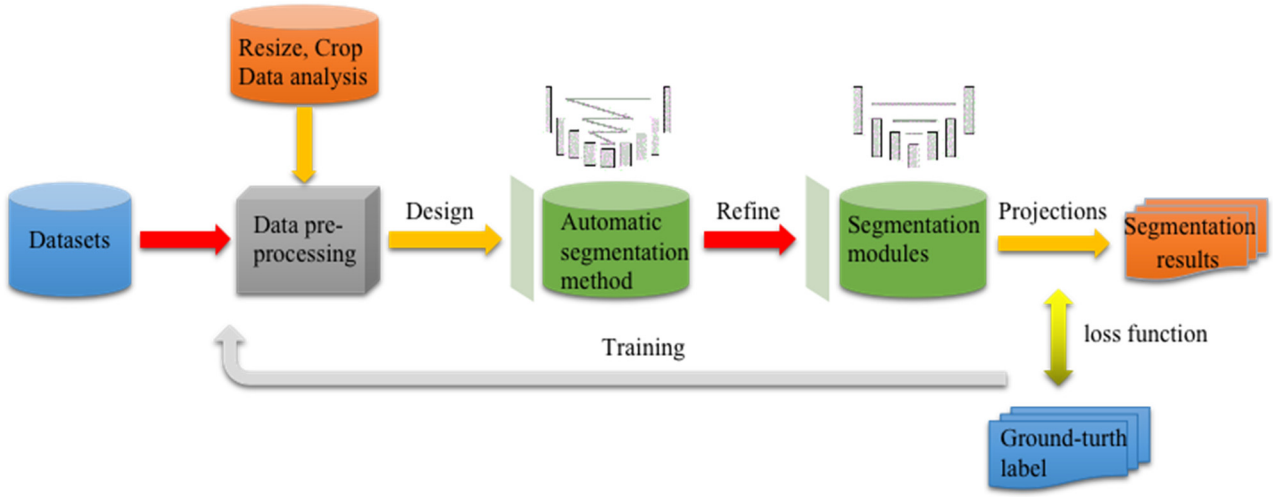

Figure 1. A block diagram of the overall study route.

To overcome these challenges, we propose a novel approach that combines two powerful deep learning architectures, transformer, and graph convolutional networks, named TS-GCN. The transformer model has been proven effective in natural language processing and image recognition tasks, while GCN has demonstrated its ability to capture the relationships between nodes in graphs. By combining the strengths of these two models, our proposed approach can better capture the spatial dependencies among pixels in mammograms and improve the accuracy of breast cancer segmentation. Our work is to segment the tumor area, and the chief purpose of tumor segmentation is tumor assessment, change tracking, and distribution identification in clinical applications. In addition to its application in breast imaging, our proposed segmentation method has the potential for extension to other important imaging modalities, such as optical coherence tomography (OCT) [36]. OCT imaging procession provides valuable information for retinal diseases, cardiovascular conditions, and dermatological disorders. Our proposed method can be adapted and applied to this image analysis. The ability to capture spatial dependencies and accurately segment target structures, such as blood vessels or pathological features, can greatly assist in OCT-based [37,38] disease diagnosis and treatment

planning. This broader application scope highlights the versatility and potential impact of our proposed method beyond breast imaging.

The contribution of this study is threefold.

- The proposed TS-GCN method is the first to design improved tumor segmentation by transformer learning of medical image blocks, and this representation can be presented at GCN for lesion region information enhancement.
- Segmentation branching is used as a novel architecture that combines transformer and GCN modules to learn more information and discriminate representations.
- The segmentation performance of the proposed model is validated on two modalities of breast imaging using standard segmentation evaluation metrics, where it outperformed the other state-of-the-art segmentation models.

2. Materials and methods

2.1. Overview

In this section, we propose a tumor segmentation method, called TS-GCN, which integrates the strengths of both Transformer and GCN for accurate and efficient segmentation of breast cancer images. The overall architecture of TS-GCN is illustrated in Figure 2. Our proposed method consists of five steps: image deep feature extraction, image blocks feature, graph representation learning, feature fusion, and segmentation. Specifically, we first employ a pre-trained Transformer model to extract features from the input medical images. It is capable of capturing both global and local features by combining the transformer and local self-attention mechanisms. Secondly, the GCN is designed to further capture the local structure of the medical images, we construct a graph using the extracted image block features as nodes. The GCN model is capable of capturing the local features of the medical images and their relationships. Then, after obtaining the graph representation of the images, we aggregate the features of all the nodes to generate a global feature. We use the inverse degree of each node as the weight for feature aggregation. Finally, we use the global feature vector as the input to predict breast cancer diagnosis. It is mainly included four parts, Transformer, GCN, CNN and the combined method.

2.2. Transformer feature encoder

In our method, we use the Transformer as an encoder to extract features from the breast cancer images. The encoder is composed of multiple layers, each consisting of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The self-attention mechanism allows the network to attend to different parts of the input image, while the feed-forward network provides a non-linear mapping to a higher-dimensional space. The output of each layer is then passed through a layer normalization and residual connection before being fed to the next layer. The final output of the encoder is a set of feature maps that capture the high-level semantic information of the input image.

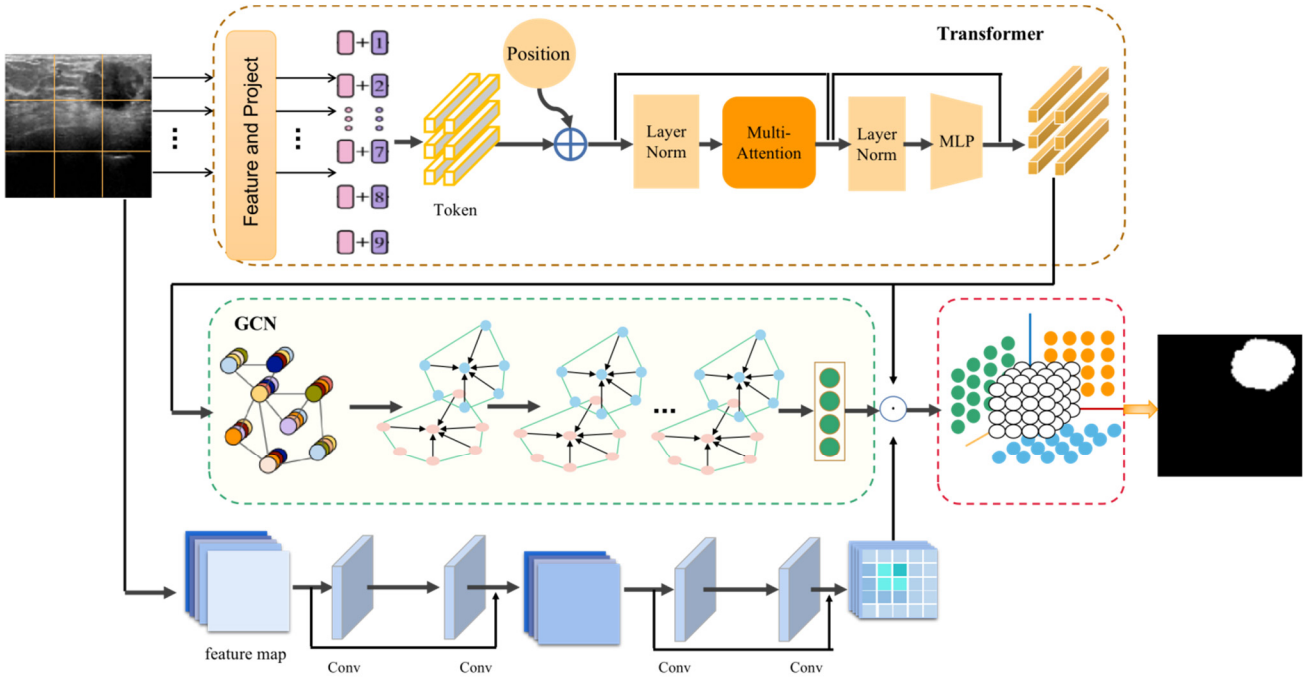

Figure 2. An overview of the TS-GCN method. It consists of five steps: image deep feature extraction, image blocks feature, graph representation learning, feature fusion, and segmentation.

The Transformer encoder [5] first extracts a set of feature maps from the input image using a series of convolutional layers. These feature maps are then transformed into a set of key-value (K, V) pairs, and the query (Q) vectors are generated by applying another convolutional layer on the input feature maps. The Transformer encoder then computes the self-attention scores between the query and key vectors, and applies a Softmax function to obtain the attention weights:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where T denotes the transpose operation. The attention weights are used to compute a weighted sum of the value vectors, which is then fed into a feed-forward network to obtain the final output of the transformer encoder:

$$FFN(Attention(Q, K, V)) = ReLU(W_2 ReLU(W_1 Attention(Q, K, V) + b_1) + b_2) \tag{2}$$

where W_1, W_2, b_1 and b_2 denote the parameters of the feed-forward network.

The transformer module consists of multiple self-attention layers, which allow the model to attend to different parts of the image when making predictions. Each self-attention layer takes as input a feature map F and produces a new feature map F' as follows:

$$F' = LayerNorm\big(F + MultiHead(F)\big) \tag{3}$$

where $LayerNorm$ is a layer normalization function, and $MultiHead(F)$ is a multi-head self-attention function defined as:

$$\mathrm{MultiHead(F)} = \mathrm{Concat(h_1, \ldots, h_i)w^0} \tag{4}$$

where $h_i = Attention(FW_i^Q, FW_i^K, FW_i^v)$ is the result of the $\mathrm{i_{th}}$ attention head, and W_i^Q, W_i^K, W_i^v and w^0 are learnable weight matrices. The output feature maps of the Transformer encoder are then fed into the GCN module for further processing.

2.3. GCN module

The GCN module [22] is designed to capture the contextual relationships between the different pixels in the image. This is achieved by representing the feature of image blocks as a graph, the GCN module then performs graph convolution operations to extract features from the graph. In our method, we use the output feature maps of the Transformer encoder as the input to the GCN module. The feature maps are first transformed into a graph representation, where each pixel is a node in the graph and the edges represent the spatial relationships between the nodes. The graph convolution operation is then performed to capture the contextual relationships between the different pixels in the image. The output of the GCN module is a set of refined feature maps that encode both the high-level semantic information from the Transformer encoder and the contextual relationships between the pixels.

First, the features of each image are represented as a matrix $X_i \in R^{N_i \times D}$, where N_i is the number of pixels in the $\mathrm{i_{th}}$ image and D is the dimensionality of the feature vector. Then, to fuse the features of multiple images, we define a graph structure $G = (V, E)$, where $V = v_1, v_2, ..., v_K$ denotes K images and E denotes the relationship between images blocks, which can be any metric based on image similarity or distance.

Next, for each image i, we construct the adjacency matrix $A_i \in R^{N_i \times N_i}$ based on the nodes around each node, where $A_{ij} = 1$ means that node v_j is a neighbor of node v_i, otherwise, it is 0.

Then, we use the GCN module for feature extraction. For each image i, we calculate the new features for each node using the following formula:

$$\mathrm{H_i^{l+1}} = \sigma(\mathrm{D_i^{-\frac{1}{2}}\widehat{A}_l\widehat{D}_l^{-\frac{1}{2}}H_i^lW_l}) \tag{5}$$

where $H_i^l \in R^{N_i \times F_i}$ is the node identity matrix of the I_{th} layer of the GCN, $\widehat{A_l} = A_i + IN_i$ is the adjacency matrix plus the self-loop, $\widehat{D_l}$ is the degree matrix $\widehat{D_l} = diga(\widehat{A_l}I)$, W_l is the weight matrix of the I_{th} layer of the GCN, and $\sigma(.)$ is the activation function.

Next, the features of the nodes can be extracted by passing information between the GCN layers. Specifically, assuming that the feature of node i is represented as $h_v^{(l)}$ in the l_{th} GCN layer, the feature $h_u^{(l-1)}$ of node u at the $l + 1_{th}$ layer can be calculated by the following equation:

$$\mathrm{h_v^{(l)}} = \sigma(\textstyle\sum_{\mathrm{u \in N_i}} \frac{1}{\mathrm{c_{u,v}}} \mathrm{W^{(l)} h_u^{(l-1)}}) \tag{6}$$

where N_i denotes the set of neighboring nodes of node i, and $W^{(l)}$ denotes the parameter matrix used in the GCN layer at layer l.

For feature extraction of multiple images, the feature representation of each image can be considered as a node, and the similarity between neighboring nodes can be calculated using the following formula:

$$\mathrm{c_{u,v}} = \exp(-\frac{\|\mathrm{x_u - x_v}\|^2}{\sigma^2}) \tag{7}$$

where x_v and x_u denote the feature representations of node u and node v, and σ is a

hyperparameter to control the weight of similarity.

Through the iteration of multiple layers of GCN, we can obtain the final hidden representation vector $h_v^{(l)}$ of each node, where l represents the number of layers of the GCN model. These hidden representation vectors can be used as a new representation of image features for subsequent tasks, such as classification or segmentation.

2.4. Combining transformer and GCN

The Transformer and GCN modules are combined in a two-stage training process. In the first stage, the Transformer encoder is trained to extract high-level features from the input image. In the second stage, the GCN module is trained to capture the contextual relationships between the pixels in the image, using the output feature maps of the Transformer encoder as input. To combine the Transformer and GCN, we first pass the image patches through the Transformer to obtain the feature vectors. The adjacency matrix for the graph is then constructed based on the spatial relationships between the patches. The feature vectors are then passed through the GCN to refine their representation based on the graph structure. The TS-GCN can be represented by the following equations:

$$\mathrm{F} = \mathrm{Transformer}(\mathrm{X}) \tag{8}$$

$$\mathrm{A_{i,j}} = \exp\left(-\frac{\|\mathrm{p_i}-\mathrm{p_j}\|^2}{\sigma^2}\right) \tag{9}$$

$$\mathrm{H} = \mathrm{GCN}(\mathrm{F}, \mathrm{A}) \tag{10}$$

where X is the input image, F is the feature matrix obtained from the transformer, A is the adjacency matrix, and H is the refined feature matrix.

The loss function used during training is Dice loss, which is a commonly used loss function for image segmentation tasks. The Dice loss measures the overlap between the predicted segmentation mask and the ground truth mask, and is given by:

After obtaining the final embedding from the transformer and GCN layers, we concatenate them and feed them through a fully connected layer to obtain the final segmentation map. The final loss function is a combination of the Dice loss and binary cross-entropy loss, given by:

$$\mathrm{L} = \mathrm{L_{Dice}} + \alpha \mathrm{L_{BCE}} \tag{11}$$

where L_{Dice} is the Dice loss, L_{BCE} is the binary cross-entropy loss, and α is a hyperparameter to balance the two losses.

L_{BCE} here is based on sigmoid to do binary classification, where N is the number of samples, as this loss is equal to the average of the categorical cross-entropy loss on the two-category task.

$$\mathrm{L_{BCE}} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i}\ln(\overline{\mathrm{y_i}}) + (1-\mathrm{y_i})\ln(1-\mathrm{y_i}) \tag{12}$$

L_{Dice} appeared frequently with outstanding performance in tumor segmentation networks. For tumor segmentation of medical images, some small-size tumors occupy only a small area of the scanned image, often resulting in loss or partial loss of detection of the foreground, while network prediction is heavily biased toward the background. To solve the problem, the loss function based on

the dice coefficient is used to reweight the sample and enhance the importance of the foreground area, making it higher than the background area. The dice D is written as follows:

$$\mathrm{D} = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{13}$$

where N is the total number of pixels, p_i is a single component of the predicted binary segmentation area P, and g_i is that of the ground truth binary area G.

2.5. *Decoder network*

The decoder network takes the refined feature maps from the GCN module as input and generates a pixel-wise segmentation mask for the input image. The decoder is composed of multiple layers, each consisting of two sub-layers: a 2D transposed convolution layer and a position-wise fully connected feed-forward network. The transposed convolution layer is used to up-sample the feature maps, while the feed-forward network provides a non-linear mapping to a higher-dimensional space. The output of each layer is then passed through a layer normalization and residual connection before being fed to the next layer. The final output of the decoder is a pixel-wise segmentation mask that indicates the probability of each pixel belonging to the breast cancer region. In summary, the proposed method combines the GCN and Transformer models to classify breast cancer images. The GCN model captures the spatial relationships between pixels in the image, the prediction facilitates clinical interpretation.

3. Results

3.1. Datasets

To evaluate the effectiveness of our proposed method, we conducted experiments on two publicly available datasets of breast images, including the public breast ultrasound Dataset BUSI [26] and dataset B [27]. Each image was equipped with the ground truth mask of the lesion to automatically interpret and analyze the breast ultrasound images. (1) BUSI, as a publicly available breast ultrasound dataset, contains both markers and annotation data. For a fair comparison with other methods, no additional processing is performed on this dataset, which is consistent with previous works. BUSI has 780 breast ultrasound images, including 547 tumor images. (2) dataset B, consists of 163 images collected by Siemens ACUSON Sequoia C512 system, with 110 benign and 53 malignant tumor images. Specifically, we selected 109 images as the train set, 13 images as the validation set, and 41 images as the test set in dataset B. The images were preprocessed by resizing them to a fixed size of 224 × 224 pixels and normalizing them to have a mean of 0.5 and a standard deviation of 0.5. The links to the BUSI and dataset B datasets of our paper are https://scholar.cu.edu.eg/?q=afahmy/pages/dataset and https://ieeexplore.ieee.org/abstract/ document/goo.gl/SJmoti, respectively. Additionally, we randomly resize, flip and rotate the training images for datasets augmentation.

3.2. Evaluation metrics

To evaluate the performance of our proposed method, we employ several widely-used evaluation metrics in the field of image segmentation. In the breast cancer segmentation model, the classifier

converts the logarithmic values into a probability distribution and uses the one with the highest probability value as the model prediction class. Specifically, we use the following metrics: the Dice similarity coefficient (Dice) measure the overlap between the predicted segmentation mask and the ground truth mask, while the measures the ratio of the intersection to the union (IoU) of the predicted and ground truth masks. All methods were trained using the same training, validation, and test sets, and were evaluated using those metrics. In addition, we also use two standard measures IoU and F1-score to verify the effectiveness of the network design. F1-score reflects the comprehensive performance of Precision and Recall. The corresponding equations are as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{16}$$

$$\text{F1} - \text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{17}$$

where *FN* represents the total number of false negatives, *TP* represents the total number of true positives, *FP* represents the total number of occurrences of false positive samples, *TN* represents the total number of true negative samples, and *N* represents the total number of samples.

Then, the model is further evaluated using the Dice, which take values in the range of [0, 1] and is usually used to measure the similarity of the prediction mask to the true value.

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{18}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{19}$$

The predicted mask represents the area of the tumor detected by an algorithm and the ground truth mask represents the actual area of the tumor as labeled by a medical expert. The IoU value ranges from 0 to 1, with 1 indicating perfect overlap between the two sets and 0 indicating no overlap. In addition, the AUC metric is also used to evaluate the performance of the model, which takes values in the range [0.5, 1.0]. the closer the AUC is to 1.0, the better the performance of the model is and the more correctly it can distinguish between positive and negative samples.

3.3. Experimental setups

The entire model is trained end-to-end using binary cross-entropy loss between the predicted mask and the ground truth mask. The model is used PyTorch and is trained on a single NVIDIA Tesla V100 GPU. We used the Adam optimizer with a learning rate of 0.001 and trained the model for 200 epochs. For 1 to 100 epochs, we set the initial learning rate as 0.0001 which is attenuated by multiplying 0.88 after every epoch. We trained the model for 100 epochs and selected the model with the best validation performance for testing. The proposed network has 55 M trainable parameters. In the testing stage, the inference time was 0.039 s per image.

We compared our proposed method with several state-of-the-art methods, including mask RCNN [28], DeepLab-v3+ [29], GCN-based [30], U-Net [31], FCN [32], inception-UNet [33] and attention-UNet [34,35] methods. We used the same experimental setup for all the methods and

evaluated the models on the same test set. Mask RCNN is a deep-learning model for object detection and instance segmentation. U-Net is a popular deep-learning architecture for medical image segmentation. FCN is a fully convolutional neural network designed for semantic segmentation. Attention U-Net and inception U-Net is a variant of U-Net that uses an attention mechanism to improve segmentation performance. DeepLab-v3+ combines the benefits of atrous convolution and the spatial pyramid module to achieve high accuracy and efficiency. GCN-UNet is a variant of U-Net that uses GCN to model spatial dependencies.

3.4. Result analysis

3.4.1. Results of BUSI datasets

The experiment aimed to evaluate the performance of different segmentation models on different datasets. The results of the experiment are summarized in the table provided. Table 1 shows the evaluation results of different methods on the BUSI breast dataset. The methods include mask RCNN, Deeplab-v3, GCN-based, U-Net, FCN, inception-UNet, attention-UNet and the proposed method TS-GCN. From Table 1, we can see that the proposed method TS-GCN achieves the highest performance in terms of IoU, F1 and AUC. It achieves an accuracy of 0.9373, a Dice of 0.9058, IoU of 0.7634, F1 score of 0.9338 and an AUC of 0.9692. The results can accurately capture the target object and achieve good segmentation results. Meanwhile, attention-UNet and Deeplab-v3 also have relatively high performance, indicating that their attention mechanism and deep learning architecture can also contribute to the segmentation accuracy. The second-best performing attention-UNet with an accuracy of 0.8883, Dice of 0.9027, IoU of 0.7338, F1 score of 0.9242 and AUC of 0.9605. Our proposed method achieved state-of-the-art performance on the BUSI dataset.

Table 1. Shows the evaluation results of different methods on the BUSI breast dataset.

Method	Acc	Dice	IoU	F1	AUC
Mask RCNN	0.8761	0.8055	0.7049	0.8760	0.9010
Deeplab-v3	0.8717	0.8790	0.7509	0.8602	0.9187
GCN-based	0.8422	0.8520	0.7300	0.8499	0.9229
U-Net	0.8641	0.8924	0.7523	0.9001	0.9208
FCN	0.8708	0.8827	0.7409	0.9110	0.9162
Inception-UNet	0.8728	0.8956	0.7480	0.9184	0.9370
Attention-UNet	0.8883	0.9027	0.7338	0.9242	0.9605
Ours	0.9373	0.9058	0.7634	0.9338	0.9692

The Deeplab-v3 model achieved an accuracy of 0.8717, a Dice of 0.8790, an IoU of 0.7509, an F1 score of 0.8602 and an AUC of 0.9187. These results suggest that the Deeplab-v3 model performed better than mask RCNN in terms of Dice and IoU, but not in terms of accuracy, F1 score and AUC. However, some methods such as GCN-based have lower performance, which suggests that their architectures may not be suitable for the specific task or need further improvement.

Other methods performed moderately, such as U-Net with Acc of 0.8641, having high Dice and

IoU scores. FCN had a relatively high F1 score of 0.9110, while inception-UNet had a high F1 score of 0.9184. Mask RCNN had the highest accuracy among all methods with a value of 0.8761. The inception-UNet model achieved an accuracy of 0.8728, a Dice of 0.8956, an IoU of 0.7480, an F1 score of 0.9184 and an AUC of 0.9370. Overall, the proposed method TS-GCN outperforms other methods in terms of most evaluation metrics. The proposed method TS-GCN leverages both transformer and GCN in a novel way, allowing it to achieve superior performance compared to other methods on most evaluation metrics. This suggests that the combination of these two powerful architectures has a synergistic effect, enabling more accurate and efficient segmentation results.

3.4.2. Results of dataset B

Table 2 presents the results of various methods evaluated on dataset B for breast cancer tumor segmentation. It can be concluded that the proposed method (referred to as "Ours" in the table) outperformed other methods in most evaluation metrics, including accuracy, Dice, IoU, F1 score and AUC. Specifically, our method achieved the highest accuracy of 0.9501, and the second-highest Dice of 0.9139. Compared to other state-of-the-art methods, TS-GCN, which combines the Transformer and GCN techniques, achieved superior segmentation results.

Table 2. Presents the results of various methods evaluated on dataset B.

Method	Acc	Dice	IoU	F1	AUC
Mask RCNN	0.8882	0.8135	0.7158	0.9015	0.9152
Deeplab-v3	0.8920	0.8890	0.7309	0.9259	0.9376
GCN-based	0.9090	0.8780	0.7656	0.9299	0.9504
U-Net	0.9371	0.9885	0.7632	0.9358	0.9315
FCN	0.9127	0.9115	0.7521	0.9321	0.9435
Inception-UNet	0.9358	0.9051	0.7422	0.9340	0.9519
Attention-UNet	0.9254	0.9110	0.7388	0.9297	0.9556
Ours	0.9501	0.9139	0.7821	0.9479	0.9739

It is worth noting that some of the other methods, such as U-Net and attention-UNet. The U-Net and attention-UNet methods achieved Dice scores of 0.9885 and 0.9110, respectively. U-Net performed the best out of all the methods in terms of the Dice score. Attention-UNet, on the other hand, achieved the highest score, demonstrating its effectiveness in segmenting breast cancer tumors. However, U-Net outperformed attention-UNet in other metrics such as accuracy, F1 score and AUC. Therefore, both U-Net and attention-UNet are effective methods for breast cancer tumor segmentation, with U-Net performing better overall. Overall, these results provide important insights into the effectiveness of different deep-learning methods for breast tumor segmentation and demonstrate the potential of the proposed method in this task. This indicates that the combination of GCN and transformer can effectively model spatial and channel dependencies for breast ultrasound image segmentation. These results demonstrate that our proposed method is effective for the segmentation task of breast cancer tumors, and it can potentially be used as a useful tool for the diagnosis and treatment of breast cancer in clinical practice.

3.4.3. Ablation studies

Additionally, we conduct ablation studies to investigate the effectiveness of several different modules for medical image segmentation, including CNNs, GCNs and transformers. As shown in Table 3, we compare the performance of the TS-GCN model with three variants: (1) the baseline model using only CNN without transformer and GCN(U-Net-only); (2) the baseline model using only Transformer without GCN and CNN (Transformer-only); (3) the baseline model using only GCN without Transformer and CNN (GCN-only); (4) the model using both CNN and GCN and Transformer(CNN-GCN); 5) the model using both CNN and Transformer (CNN-TS); (6) the model using CNN, GCN and Transformer (TS-GCN).

Table 3. Conducts an ablation study to investigate the effectiveness of different modules.

Method	Acc	Dice	IoU	F1	AUC
CNN-only	0.8641	0.8924	0.7523	0.9001	0.9208
GCN-only	0.8422	0.8520	0.7300	0.8499	0.9229
Transformer-only	0.8701	0.8808	0.7690	0.9030	0.9276
CNN + Transformer	0.9045	0.9205	0.8009	0.9240	0.9487
CNN + GCN	0.8956	0.9132	0.7876	0.9155	0.9357
TS-GCN (our method)	0.9373	0.9058	0.7634	0.9338	0.9692

Firstly, we evaluated the model with only CNN layers, which achieved an accuracy of 0.8641, Dice of 0.8924, IoU of 0.7523, F1 of 0.9001 and AUC of 0.9208. Then, we tested the GCN-only model, which resulted in lower performance than the CNN-only model, with an Acc of 0.8422. Next, we evaluated the Transformer-only model, which outperformed the GCN-only model but was inferior to the CNN-only model in terms of segmentation performance, achieving an accuracy of 0.8701. Subsequently, we tested the CNN + Transformer and CNN + GCN models, which showed significant improvements over the single-component models. The CNN + Transformer model achieved an accuracy of 0.9045, while the CNN+GCN model achieved an accuracy of 0.8956. When CNNs are combined with either GCNs or Transformers, there is a further performance improvement, indicating that the combination of these different types can effectively capture both local and global features. Finally, we compared our proposed TS-GCN method with the other models, which achieved the best segmentation performance with an accuracy of 0.9373 and AUC of 0.9692. This indicates that combining CNNs, GCNs, and Transformers in a single model can capture local and global features effectively and improve segmentation performance. By incorporating these strengths, TS-GCN can achieve state-of-the-art results in the segmentation task.

3.4.4. Quality analysis

We conducted a qualitative analysis to evaluate the quality of the segmentation masks generated by our proposed method. As shown in Figure 3, we randomly selected 5 images from the BUSI and dataset B and visually inspected the segmentation masks generated by our method and the ground truth masks provided by the radiologists. The ground truth mask is presented in the last column, while the remaining columns represent the segmentation outputs of different methods. As can be observed, our

proposed TS-GCN method achieves superior segmentation accuracy compared to the other methods, as evidenced by the high degree of overlap between predicted masks and ground truth masks.

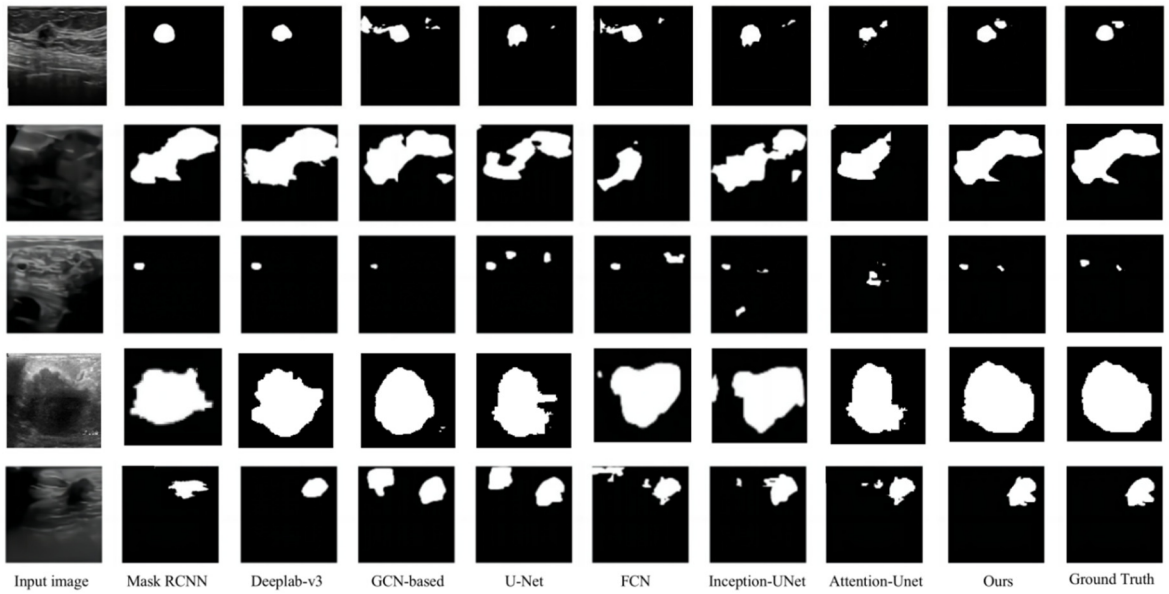

Figure 3. The quality of the segmentation masks generated by our proposed method.

Additionally, the series of U-Nets and GCN-based methods tend to produce blurry and fragmented segmentation results, particularly around the edges of the tumor. The TS-GCN method effectively captures the spatial dependencies between the pixels in the image and generates more precise segmentation results. It is visible from the results that compared to existing techniques, TS-GCN techniques have demonstrated significant dominance in quality and accuracy regarding the segmentation of complex regions. Our approach will handle the glitches noticeably present in other approaches' output. Our method smoothes the edge of tumor segmentation very well and achieves accurate segmentation of small tumor areas without leakage. The TS-GCN method demonstrates strong performance in accurately segmenting breast tumors, making it a promising tool for clinical applications in breast cancer diagnosis and treatment planning.

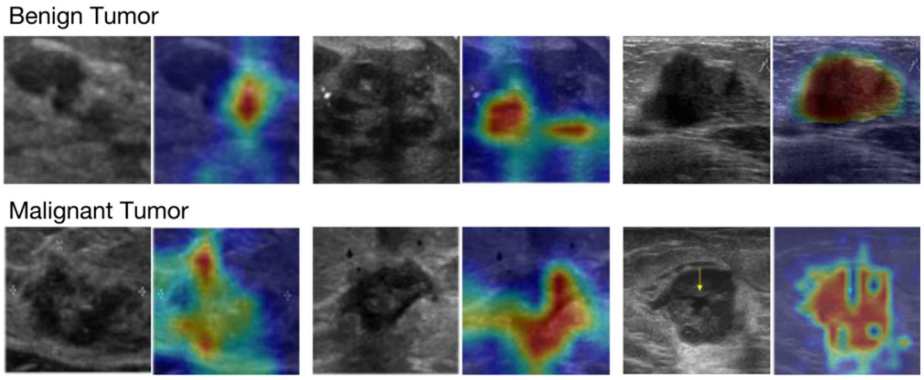

Figure 4. Visual comparison of the proposed method.

In addition, to show the effectiveness of the feature aggregation network based on the TS-GCN more intuitively, we analyzed the attention maps of breast images. To visualize class-specific attention, we applied class activation mapping to generate attention maps of breast images, and the results were shown in Figure 4. We presented grad-cam maps of benign and malignant tumors respectively. It can be seen that TS-GCN can notice the key areas of tumor classification, including boundary, calcification, and so on. The attention map shows that TS-GCN pays more attention to edge-related areas. Although the interpretability of deep learning is still a difficult research field, from the attention maps, we can see that the proposed feature aggregation can effectively integrate the regions of interest of the Transformer and GCN networks. It can effectively fuse the features extracted from the two networks for breast tumor classification and segmentation.

4. Discussion

The TS-GCN method proposed in our work offers several advantages and also has some limitations. The method leverages the strengths of both the Transformer and GCN models, enabling it to capture spatial dependencies among pixels and improve the accuracy of tumor segmentation. The fusion of these models allows for more comprehensive feature integration and better representation of tumor regions. Also, based on the diffusion model theory, the TS-GCN considers the multimodality of labels, text, or images in the hidden space. This innovative approach enhances the synthesis of logo images by considering multiple modalities, leading to more diverse and high-quality samples. While our focus is on tumor segmentation in breast images, the TS-GCN method has the potential to be applied to other medical imaging tasks, as well as image super-resolution, deblurring and text-to-image translation. There are some Limitations of the method. The performance of TS-GCN heavily relies on the availability and quality of training data. Like many deep learning models, TS-GCN may lack interpretability in terms of understanding the specific decision-making process. While attention maps and visualizations can provide some insights, the exact reasoning behind the model's segmentation decisions may not be readily explainable. It is important to note that these advantages and limitations are specific to the TS-GCN method proposed in our work and may vary in different contexts and applications [39–41].

Society will benefit from this segmentation method in several ways. Firstly, the improved accuracy of tumor segmentation enables more precise and reliable disease diagnosis. This can lead to early detection of cancers and other diseases, allowing for timely intervention and treatment, ultimately saving lives and improving patient outcomes. Secondly, the computer-aided diagnostic system based on this segmentation method can enhance the efficiency of medical professionals. By automating the segmentation process, medical practitioners can save valuable time and resources, enabling them to focus on other critical tasks, such as treatment planning and patient care. Furthermore, the application of this segmentation method can contribute to advancing medical research and knowledge. This knowledge can drive further advancements in cancer research, personalized medicine, and the development of novel therapeutic approaches. Overall, the societal impact of this segmentation method lies in its potential to improve healthcare outcomes, enhance medical professionals' efficiency, and advance medical research, ultimately benefiting individuals, healthcare systems, and society as a whole.

5. Conclusions

In this paper, we proposed a novel breast tumor segmentation method, TS-GCN, based on the fusion of Transformer and GCN networks. Specifically, we designed a feature aggregation network to integrate the complementary features extracted from the Transformer, GCN and CNN networks, which improves the segmentation performance. The CNN feature extractor uses the common ResNet-50 backbone network. The Transformer component enables the model to capture long-range dependencies and contextual information, while the GCN component allows for effective information propagation and aggregation across the graph structure. Experimental results on two publicly available datasets demonstrate that our proposed method achieves state-of-the-art performance in terms of various evaluation metrics. The attention map analysis also shows that the proposed method can effectively fuse the features of the two networks and highlight the important regions of the breast tumor. In addition, the ablation studies further verify the effectiveness of each module of our method. The application of this segmentation method can contribute to advancing medical research and knowledge. Accurate tumor segmentation allows for more accurate analysis and quantification of tumor characteristics, leading to a better understanding of disease progression, response to treatment, and potential biomarkers. In summary, the proposed TS-GCN method achieves superior performance in breast tumor segmentation and provides a promising direction and basic reference for future research in medical image analysis.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, et al., Automated breast ultrasound lesions detection using convolutional neural networks, *IEEE J. Biomed. Health Inf.*, **22** (2018), 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873
2. J. Gao, Q. Jiang, B. Zhou, D. Chen, Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview, *Math. Biosci. Eng.*, **16** (2019), 6536–6561. https://doi.org/10.3934/mbe.2019326
3. C. Xu, Y. Qi, Y. Wang, M. Lou, J. Pi, Y. Ma, ARF-Net: An adaptive receptive giel network for breast mass segmentation in whole mammograms and ultrasound images, *Biomed. Signal Process. Control*, **71** (2022), 103178. https://doi.org/10.1016/j.bspc.2021.103178
4. Y. Wang, N. Wang, M. Xu, J. Yu, C. Qin, X. Luo, et al., Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound, *IEEE Trans. Med. Imaging*, **39** (2019), 866–876. https://doi.org/10.1109/TMI.2019.2936500

5. S. Jiang, J. Li, Z. Hua, Transformer with progressive sampling for medical cellular image segmentation, *Math. Biosci. Eng.*, **19** (2022), 12104–12126. https://doi.org/10.3934/mbe.2022563
6. A. Iqbal, M. Sharif, MDA-Net: Multiscale dual attention-based network for breast lesion segmentation using ultrasound images, *J. King Saud Univ. Comput. Inf. Sci.*, **34** (2022), 7283–7299. http://dx.doi.org/10.1016/j.jksuci. 2021.10.002
7. R. Bi, C. Ji, Z. Yang, M. Qiao, P. Lv, H. Wang, Residual-based attention-Unet combing DAC and RMP modules for automatic liver tumor segmentation in CT, *Math. Biosci. Eng.*, **19** (2022), 4703–4718. https://doi.org/10.3934/mbe.2022219
8. B. Lei, S. Huang, R. Li, C. Bian, H. Li, Y. H. Chou, et al., Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder-decoder network, *Neurocomputing*, **321** (2018), 178–186. https://doi.org/10.1016/j.neucom.2018.09.043
9. Y. Ouyang, Z. Zhou, W. Wu, J. Tian, F. Xu, S. Wu, et al., A review of ultrasound detection methods for breast microcalcification, *Math. Biosci. Eng.*, **16** (2019), 1761–1785. https://doi.org/10.3934/mbe.2019085
10. M. N. S. K. B. Soulami, N. Kaabouch, A. Tamtaoui, Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using U-net model based semantic segmentation, *Biomed. Signal Process. Control*, **2021** (2021), 102481. https://doi.org/10.1016/j.bspc.2021.102481
11. Y. Wang, N. Wang, M. Xu, J. Yu, C. Qin, X. Luo, et al., Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound, *IEEE Trans. Med. Imaging*, **39** (2019), 866–876. https://doi.org/10.1109/TMI.2019.2936500
12. E. H. Houssein, M. M. Emam, A. A. Ali, P. N. Suganthan, Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review, *Exp. Syst. Appl.*, **167** (2021), 114161. https://doi.org/10.1016/j.eswa.2020.114161
13. M. Xian, Y. Zhang, H. D. Cheng, F. Xu, B. Zhang, J. Ding, Automatic breast ultrasound image segmentation: A survey, *Pattern Recognit.*, **79** (2018), 340–355. https://doi.org/10.1016/j.patcog.2018.02.012
14. Y. Tong, Y. Liu, M. Zhao, L. Meng, J. Zhang, Improved U-net MALF model for lesion segmentation in breast ultrasound images, *Biomed. Signal Process. Control*, **68** (2021), 102721. https://doi.org/10.1016/j.bspc.2021.102721
15. D. Mishra, S. Chaudhury, M. Sarkar, A. S. Soin, Ultrasound image segmentation: A deeply supervised network with attention to boundaries, *IEEE Trans. Biomed. Eng.*, **66** (2018), 1637–1648. https://doi.org/10.1109/TBME.2018.2877577
16. G. Chen, Y. Dai, J. Zhang, C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation, *Comput. Methods Programs Biomed.*, **2022** (2022), 107086. https://doi.org/10.1016/j.cmpb.2022.107086
17. N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, et al., Fanet: A feedback attention network for improved biomedical image segmentation, *Trans. Neural Networks Learn. Syst.*, **2022** (2022). https://doi.org/10.1109/TNNLS.2022.3159394
18. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *Trans. Pattern Anal. Mach. Intell.*, **2018** (2018), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

19. Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging CNN and Transformer for 3d medical image segmentation, in *Medical Image Computing and Computer Assisted Intervention MICCAI*, (2021), 171–180. https://doi.org/10.1007/978-3-030-87199-4_16
20. N. S. Punn, S. Agarwal, RCA-IUnet: A residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging, *Mach. Vision Appl.*, **33** (2022), 1–10. ttps://doi.org/10.1007/s00138-022-01280-3
21. N. Abraham, N. M. B. T. Khan, A novel focal tversky loss function with improved attention U-Net for lesion segmentation, *Int. Symp. Biomed. Imaging*, **2019** (2019), 683–687. https://doi.org/10.1109/ISBI.2019.8759329
22. W. Jin, T. Derr, Y. Wang, Y. Ma, Z. Liu, J. Tang, Node similarity preserving graph convolutional networks, in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, (2021), 148–156. https://doi.org/10.1145/3437963.3441735
23. B. Wu, X. Liang, X. Zheng, Y. Guo, H. Tang, Improving dynamic graph convolutional network with fine-grained attention mechanism, in *ICASSP International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2022), 3938–3942. https://doi.org/10.1109/ICASSP43922.2022.9746009
24. Y. Lu, Y. Chen, D. Zhao, J. Chen, Graph-FCN for image semantic segmentation, in *Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks*, (2019), 97–105. https://doi.org/10.1007/978-3-030-22796-8_11
25. Y. Huang, Y. Sugano, Y. Sato, Improving action segmentation via graph-based temporal reasoning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 14024–14034. https://doi.org/10.1109/CVPR42600.2020.01404
26. W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, *Data Brief*, **28** (2020), 104863. https://doi.org/10.1016/j.dib.2019.104863
27. Z. Fu, J. Zhang, R. Luo, Y. Sun, D. Deng, L. Xia, TF-Unet: An automatic cardiac MRI image segmentation method, *Math. Biosci. Eng.*, **19** (2022), 5207–5222. https://doi.org/10.3934/mbe.2022244
28. X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, et al., Crack detection and comparison study based on faster R-CNN and mask R-CNN, *Sensors*, **22** (2022), 1215. https://doi.org/10.3390/s22031215
29. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 801–818.
30. R. Huang, M. Lin, H. Dou, Z. Lin, Q. Ying, X. Jia, et al., Boundary-rendering network for breast lesion segmentation in ultrasound images, *Med. Image Anal.*, **80** (2022), 102478. https://doi.org/10.1016/j.media.2022.102478
31. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-net architecture for medical image segmentation, in *Lecture Notes in Computer Science*, (2018), 3–11. http:/dx.doi.org/10.1007/ 978-3-030-00889-51
32. E. Sanderson, B. J. Matuszewski, FCN-Transformer feature fusion for polyp segmentation, in *Medical Image Understanding and Analysis: 26th Annual Conference*, Springer International Publishing, (2022), 892–907. https://doi.org/10.1007/978-3-031-12053-4_65

33. X. Feng, T. Wang, X. Yang, M. Zhang, W. Guo, W. Wang, ConvWin-UNet: UNet-like hierarchical vision transformer combined with convolution for medical image segmentation, *Math. Biosci. Eng.*, **20** (2023), 128–144. https://doi.org/10.3934/mbe.2023007
34. 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, preprint, arXiv:1804.03999. https://doi.org/10.48550/arXiv.1804.03999
35. X. Zhang, K. Liu, K. Zhang, X. Li, Z. Sun, B. Wei, SAMS-Net: Fusion of attention mechanism and multi-scale features network for tumor infiltrating lymphocytes segmentation, *Math. Biosci. Eng.*, **20** (2023), 2964–2979. https://doi.org/10.3934/mbe.2023140
36. R. K. Meleppat, P. Zhang, M. J. Ju, S. K. K. Manna, Y. Jian, E. N. Pugh, et al., Directional optical coherence tomography reveals melanin concentration-dependent scattering properties of retinal pigment epithelium, *J. Biomed. Optics*, **24** (2019), 066011. https://doi.org/10.1117/1.JBO.24.6.066011
37. R. K. Meleppat, C. R. Fortenbach, Y. Jian, E. S. Martinez, K. Wagner, B. S. Modjtahedi, et al., In Vivo imaging of retinal and choroidal morphology and vascular plexuses of vertebrates using swept-source optical coherence tomography, *Trans. Vis. Sci. Tech.*, **11** (2022), 11. https://doi.org/10.1167/tvst.11.8.11
38. R. K. Meleppat, K. E. Ronning, S. J. Karlen, K. K. Kothandath, M. E. Burns, E. N. Pugh, et al., In situ morphologic and spectral characterization of retinal pigment epithelium organelles in mice using multicolor confocal fluorescence imaging, *Invest. Ophthalmol. Vis. Sci.*, **61** (2020), 1. https://doi.org/10.1167/iovs.61.13.1
39. J. He, Q. Zhu, K. Zhang, P. Yu, J. Tang, An evolvable adversarial network with gradient penalty for COVID-19 infection segmentation, *Appl. Soft Comput.*, **113** (2021), 107947. https://doi.org/10.1016/j.asoc.2021.107947
40. X. Liu, D. Zhang, J. Yao, J. Tang, Transformer and convolutional based dual branch network for retinal vessel segmentation in OCTA images, *Biomed. Signal Process. Control*, **83** (2023), 104604. https://doi.org/10.1016/j.bspc.2023.104604
41. C. Zhao, A. Vij, S. Malhotra, J. Tang, H. Tang, D. Pienta, et al., Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms, *Comput. Biol. Med.*, **136** (2021), 104667. https://doi.org/10.1016/j.compbiomed.2021.104667

AIMS Press