Knowledge-Based Systems 267 (2023) 110393

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images

Ahmed Iqbal *, Muhammad Sharif

Department of Computer Science, COMSATS University Islamabad, Wah Campus, Pakistan

ARTICLE INFO

Article history:
Received 15 September 2022
Received in revised form 10 February 2023
Accepted 12 February 2023
Available online 14 February 2023

Keywords:
Dual encoder
Swin transformer
U-Shaped Network
Breast tumor classification
Breast tumor segmentation

ABSTRACT

Breast cancer is considered the most commonly diagnosed cancer globally and falls second to lung cancer. For the early detection of breast tumors in women, breast cancer analysis using Ultrasound, Mammography, and MRI modalities as the initial screening process. Due to the random variation, irregular shapes, and blurred boundaries of tumor regions, the accurate segmentation of breast tumors is still a tricky task. The existing convolutional neural networks (CNNs) inherit their limitation by extracting global context information and, in most cases, proved less efficient in obtaining satisfactory results. As a solution, we proposed the BTS-ST network, a novel solution for breast tumor segmentation and classification that Swin-Transformer (ST) inspires. The BTS-ST network incorporates Swin-Transformer into traditional CNNs-based U-Net to improve global modeling capabilities. To improve the feature representation capability of irregularly shaped tumors, we first introduced a Spatial Interaction block (SIB), encoding spatial knowledge in the Swin Transformer block by developing pixel-level correlation. The segmentation accuracy of small-scale tumor regions is increased by building a Feature Compression block (FCB) to prevent information loss and compress smaller-scale features in patch token down sampling of Swin-Transformer. Finally, a Relationship Aggregation block (RAB) is developed as a bridge between dual encoders to combine global dependencies from Swin-Transformer into the features from CNN hierarchically. Extensive experiments are performed on breast tumor segmentation and classification tasks using multimodality Ultrasound, Mammogram, and MRI-based datasets. The results demonstrate that our proposed solution is comparatively better than other state-of-the-art methods.

1. Introduction

Breast cancer is a worldwide health problem and the second threatening cause of mortality in women all around the world [1,2]. Breast cancer occurs when cancer cells grow uncontrollably in women breast tissues. The tumor region is a mass of abnormal cancerous tissues and is further categorized as benign and malignant. Both types of cancerous tumors have different morphological appearances, shapes, and textures. As a result, by examining the tumor's appearance, clinicians can typically determine the type or presence of breast cancer. According to American cancer statistics, 287,850 breast cases are reported in women, and 2710 cases are reported in men [3,4]. Early diagnosis of a malignant breast tumor and immediate treatment could improve five years survival rate of the patient [5]. Healthcare experts normally recommend a Mammogram, Ultrasound, and MRI imaging modalities for the screening process of breast cancer in a female patient. Due to the high burden of the population, the number of breast cancer patients is increasing, and more screening centers and medical experts are required. The manual screening of breast ultrasound images is costly and time-consuming, even for an expert radiologist [6]. A large number of patient screening centers are a big challenge for healthcare departments. As an alternative, medical experts require a fully automated computer-aided diagnosis (CAD) system for detecting breast cancer in women. In recent years, deep learning-based networks gained significant potential for diagnosing breast cancer with maximum accuracy [7]. However, a researcher from academia and the healthcare industry faces significant challenges with automated medical image processing [8]. Using deep learning techniques to raw ultrasound images alone appears insufficient due to the intricate artifacts in breast images. Moreover, it is a difficult and time-consuming to automatically segment breast tumors from the surrounding normal regions by allocating pixel-wise labels in cancerous breast images. The patient-specific characteristics are may different in breast tumors (*colors, size, texture, location, and shadows*) and challenging for radiologists [9]. In conventional breast tumor segmentation methods, the tumor border area is a key point to differentiate from surrounding background pixels. There are various thresholding, clustering, histogram equalization, and active-contouring methods for breast

* Corresponding author.
E-mail address: ahmedeqbal@gmail.com (A. Iqbal).

https://doi.org/10.1016/j.knosys.2023.110393
0950-7051/

tumor segmentation. Learning medical image attributes based on different convolutional neural network structures is the basis for the proposed deep learning networks. In previous years, different deep-learning CNN networks are introduced to enhance the segmentation results, such as fully connected network (FCN) [10] and U-Net [11]. These networks extract contextual features based on the encoder and decoder architecture, U-Net [11] obtained remarkable results in various medical image segmentation challenges. All these methods normally neglect global contextual information, which is more important for detecting the breast tumor location. The long-range dependencies involve pixel assignment in segmentation and are most important in medical imaging cases, especially the high-definition boundary of pixels. Therefore, long range dependencies are learned to improve the global contextual information of feature maps among pixels in medical images can aid to localize bounders of breast tumor regions and enhance segmentation precision. The innovation of skip connections between encoders and decoders in U-Net [11] have enhanced the performance of medical image segmentation tasks. During the down-sampling process, the location and global contextual knowledge still impede the enactment of segmentation performance, even though the encoder–decoder based structure and skip connections prove helpful for U-Net [11] to affectively extract low-level and high-level feature maps of input data. Moreover, the consecutive upsampling process in the decoders is dependent on the high-level feature maps that neglect the detailed spatial knowledge of low-level feature maps. The best way to increase segmentation performance is to extract more global contextual data and adaptively fuse the feature maps between the encoder–decoder. In recent years, transformer networks [12,13] extract long-range dependencies by utilizing self-attention mechanisms. The benefits have already been shown in the areas of computer vision and natural language processing (NLP). A vision transformer (ViT) network can benefit from parallel self-attention heads to extract longer-range dependencies, in contrast to the non-local attention mechanism [14]. Moreover, a data-efficient image transformer can also use a Feedforward Network (FFN) to enhance its modeling capabilities [15].

The contribution of the proposed study is following:

- Herein, we have proposed a novel segmentation and classification-based network named BTS-ST, a hybrid of parallel Swin-Transformer and U-Net-based architecture.
- The pixel-level feature correlation is focused on a spatial dimension, Spatial Interaction block (SIB) is introduced that can reduce the semantic ambiguity of the background region. Additionally, SIB has developed for Swin-Transformer's window mechanism-related limitations on its ability to perform global modeling.
- The loss of small features is alleviated during the patch token down-sampling process, Feature compression block (FCB) is used in the proposed network. In addition, FCB collects more useful information related to breast tumor regions and decreases the loss of information.
- The details of discriminative features are gathered by the Relational aggregation block (RAB) that captures channel-based information from the auxiliary encoder with guidance to the primary encoder and differentiate tumors from background regions with higher accuracy.

The rest of the work is organized as follows: Section 2 shares the related works close to our proposed work. The detail of the methodology is mentioned in Section 3, with a complete mathematical explanation and graphical representation. Section 4 describes the experimental setup and results. Finally, a detailed discussion and conclusions of the proposed research are given in Section 5 and Section 6.

2. Related work

Breast tumor segmentation and classification have been a hot research trend for many years. The current methods can be categorized into conventional learning methods and deep-learning-based methods. Here, we provide an overview of both methods most relevant to our proposed work, especially state-of-the-art deep learning networks.

In the conventional methods, the Watershed-based method [16], Graph-based method [17], Gradient vector flow snakes method [18], and Thresholding-based methods [19] are proposed for breast tumor segmentation. However, the basic limitation of conventional methods is greatly influenced by image feature definitions, seed points, and various threshold parameters. When segmenting breast tumors, CNN-based methods do not need precise image feature definitions, in contrast to conventional feature-oriented methods.

Arya *et al.* [20] presented the SiGaAtCNN network extracting features from different modalities of breast images. The convoluted feature maps are produced by these CNNs, which also add additional informative features for classifications and embrace the idea of sigmoid-gated attention. Similar, study [21] is proposed with the APSDAE method for selection of useful diagnostic parameters and capture features. The original data information is preserved and combined with auxiliary information to capture better feature representations of breast tumor images. Luo *et al.* [22] presented a novel segmentation-classification network by incorporating segmentation base attention knowledge into a deep convolution neural network to classify breast tumor images accurately. In addition, two parallel models are used to capture features from ultrasound images, and segmentation channel attention-based feature aggregation is utilized to automatically integrate feature capture from two feature models to improve the tumor detection capability of the network. In research [23], a modified variant of FF-UNet is proposed by adjusting the fixed receptive field via a feature-fused module and attention gate mechanism. In previous research, MDA-Net [24] proposed with multiscale fusion block with composed convolution sequences and a dual attention mechanism with channel and lesion-based attention blocks. The proposed MDA-Net [24] was tested against breast ultrasound images and simultaneously validated on an MRI dataset. Segmenting breast cancers in 3-D multi-modal MRI, Chengtao *et al.* [25] present the latest inter-modality information interaction network. The introduced network uses a hierarchical structure to capture local information from smaller tumor regions, enabling accurate tumor boundary segmentation. The small amount of ultrasound datasets and labeling is another challenge, Zhai *et al.* [26] proposed an ASSGAN-based model, which is based on discriminator and generator networks. The synchronization of two networks supervises each other and generates reliable segmentation predicted masks for unlabeled datasets. The model training is effectively improved by using unlabeled datasets. Cheng *et al.* [27] presented the deepest semantically guided multi-scale feature fusion network (DSGMFFN) for the segmentation of breast tumors using ultrasound images. The deepest semantically guided decoder (DSGNet) and a multi-scale feature fusion model (MFFM) are developed to tackle the large variety of breast tumor shapes, textures, and morphology. Zhou *et al.* [28] present a multitasking framework that is based on two sub-networks: encoder and decoders for accurate segmentation and multi-scale network for segmentation tasks. The methodology also employs an iterative training scheme to enhance feature maps with the aid of probability maps collected from earlier rounds in order to address the fuzziness of tumor boundaries in ABUS images. Pan *et al.* [29] proposed a novel breast tumor segmentation SC-FCN-BLSTM by embedding bi-directional long short-term memory and spatial-channel

block into the fully convolutional network. An SC-attention module is meant to incorporate both rich semantic information and finer-grained spatial information in order to reduce performance deterioration brought on by ambiguous boundaries and different tumor sizes. To predict axillary lymph node metastasis status in initial-stage breast cancer, Wang *et al.* [30] present a fusion network based on an encoder and decoder path incorporating a fusion stream path. The basic purpose of a fusion stream path is to aggregate beneficial feature representations from the encoder and decoder paths. They fed the superpixel image to the fusion network together with the original image to enhance boundary information retention and lessen the effects of image noises.

The deep neural network can learn a thorough relationship between the foreground pixels and the background pixels by organizing the global contextual information of the target. These learned relationship patterns help the deep neural network to identify the targeted regions. In biomedical imaging tasks, rich global contextual information is also crucial for the accurate segmentation of breast tumors. In previously proposed methods, global contextual information significance is ignored by various studies. New research directions also focused on object segmentation by introducing collaborative foreground background integration to achieve state-of-the-art results [31]. Chen *et al.* [32] presented a transformer to individually learn patch relationships across four mammograms taken from two-view (CC/MLO) of two-side (right/left) breasts, using local transformer blocks. Two different views are presented by four images of the left and right breast and are concatenated to fed into a global transformer unit to understand the patch relationship among different views and sides. Qin *et al.* [33] proposed a two-step breast cancer image segmentation method. The breast region was first roughly outlined using the U-Net [11] concept. Then a TRIMUnet model was introduced, in which an upgraded dynamic ReLU function was used in place of the ReLU function of the encoder and decoder structural unit, and the MSPCF and Transformer blocks were embedded to the encoder route.

As illustrated in Fig. 1(a), the standard transformer block based on Multi-head Self Attention (MSA), a Multi-Layer Perceptron (MLP), and a Layer Normalization (LN) blocks are shown. Normally, the MSA block performs an essential role for developing global dependencies between input and output sequences. In other research [34], presented a Swin transformer, which includes shifted window method that restricts the computation of MSA to non-intersecting windows while enabling crossed-window knowledge interaction. The proposed ST performs well in various computer vision tasks, including as image segmentation, detection, and classification, with only linear computing complexity. Instead of using the standard Transformer's multiple-head self-attention (MSA), it utilizes Window MSA (W-MSA) and Shifted Window MSA (SW-MSA), as seen in Fig. 1(b).

In studies [35,36] Swin Transformer is used as a backbone, and U-Net [11] encoder–decoder-based networks are proposed for medical image segmentation tasks. Similarly, TransU-Net [37] and TransFuse [38] networks are also proposed for medical image segmentation. However, standard Transformer networks produce results that are below average because they only focus on global modeling and ignore positioning ability. In TransU-Net [37], novel encoder architecture was developed using sequentially stacking CNNs and Transformer, while TransFuse [38] runs both in parallel and tries to fuse the two features.

To extend existing vision transformer (ViT) research, we use the BTS-ST network, a novel solution for breast tumor segmentation and classification that Swin Transformer inspires. Fig. 1(c) shows, we have embedded Spatial Interaction block (SIB) between Swin transformer block W-TB and SW-TB.

3. Methodology

In the following section, we proposed the architecture of BTS-ST, followed by the relational aggregation block (RAB), Spatial interaction block (SIB), and Feature compression block (FCB). All components are introduced in the following sections, and the primary illustration is shown in Fig. 2.

3.1. Network architecture

The architecture of the proposed BTS-ST is shown in Fig. 2. The proposed BTS-ST, a hybrid of parallel U-Net [11] and Swin-transformer (ST), adopts the superior U-Net [11] structure, where skip connection layer is used to connect the encoder and decoder. The BTS-ST constructs a dual encoder which is based on a CNN-based residual network and ST and transfers information using RAB to collect the discriminative features of breast tumor images. Additionally, we developed SIB and FCB blocks to improve the overall performance of the proposed BTS-ST structure. Given image $x \in \mathbb{R}^{h\times w\times c}$, ViT transform into non-overlapping patches to analogize token of sequences. While image patches are the opposite, there is no inherent association between tokens. In a single patch, the pixels of specific objects are normally clustered and have a strong semantic correlation. Therefore, overlapping patch tokens are obtained from each image by convolution operation to prevent loss of linearity of semantic information in initial input phase. The overlap rate is set to 50%, where each patch size is 8×8. These patches are then flattened and projected into dimension $\boldsymbol{C}_1$ via the linear embedding layer. All patch tokens are loaded to the ST block-stacked auxiliary encoder. The auxiliary encoder is based on four feature-capturing phases, and the output of every phase is explained as $\boldsymbol{S}_n$, where $n = 1,2,3, 4..$ The standard ST block is based on two variants, shifted window-based transformer (S-TB), and simple window-based transformer (W-TB). In particular, we proposed the Spatial Interaction block (SIB), which is coupled to ST blocks, to achieve pixel-based information sharing. The window-based self-attention limitation can be efficiently addressed by SIB, which can also aid to resolve the issue of semantic ambiguity brought on by occlusion. Additionally, we construct the feature compression block (FCB) to produce a *4*-stage hierarchical feature encoding mechanism by decreasing the length of the patch token in order to acquire multi-scale features while matching with the feature resolution of the primary encoder. In following phase, the introduced FCB decrease the gaps of small scale object feature map. The output resolution of phase n is $\frac{h}{2^{n+1}} \times \frac{w}{2^{n+1}}$ and the dimension are $2^{n-1}C_1$. To extract the deep features in the primary encoder, the image x is fed to ResNet-18 with half compression on the channels. The residual block nth output feature map can presented as $A_n \in \mathbb{R}\frac{h}{2^{n+1}} \times \frac{w}{2^{n+1}} \times 2^{n-1}C_2$. The A_n and output S_n from corresponding phase of the auxiliary encoder is fed to RAB, and fusion results are returned to the primary encoder block. The RAB creates the link between the channel attention technique and deformable convolution to connect the primary and auxiliary encoders. The above four encoding phases, features $F \in \mathbb{R}\frac{h}{32} \times \frac{w}{32} \times 1024$ fed into decoder block after convolution layer. To expand resolution 2×2 devolution deconvolution is performed on fed input. Like U-Net [11], BTS-ST reduces the channels with 3×3 convolutional operations while concatenating the encoding–decoding features using skip connection layers. Here, a batch normalization and ReLU operations follow each convolutional layer. The following steps are repeated four times, features of F are gradually expanded to $F' \in \mathbb{R}\frac{H}{2} \times \frac{H}{2} \times 64$. Ultimately, implement a 3×3 convolution and linear interpolation up-sample of features F' to achieve the predicted mask.

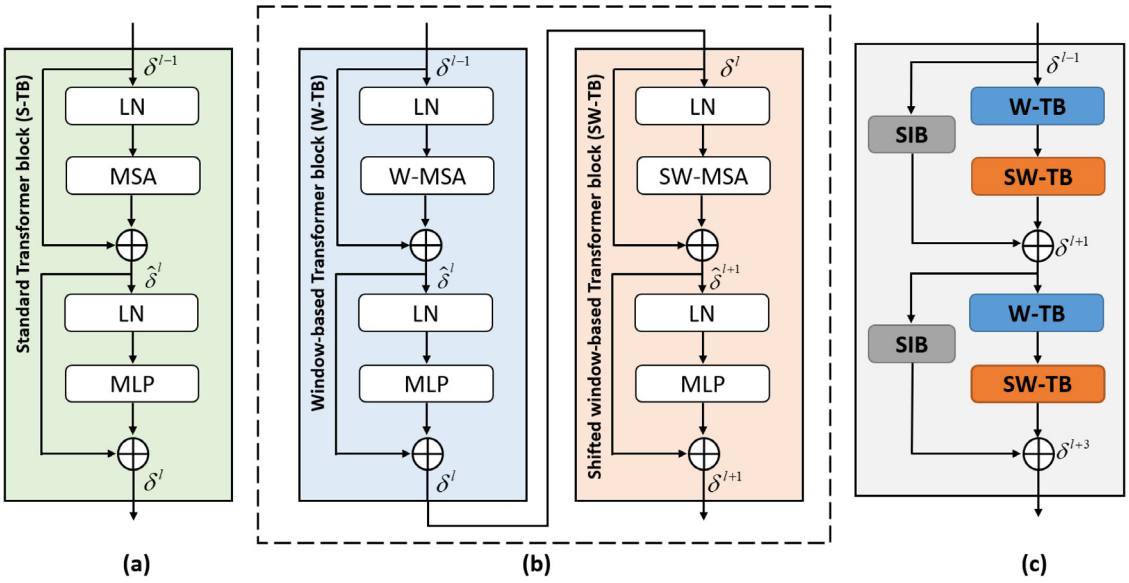

Fig. 1. **(a)** The architecture of the standard transformer block **(b)** Two consecutive Window-based Transformer block (W-TB) and Shifted window-based Transformer block (SW-TB) **(c)** and, Swin Transformer blocks with the novel Spatial Interaction block (SIB).

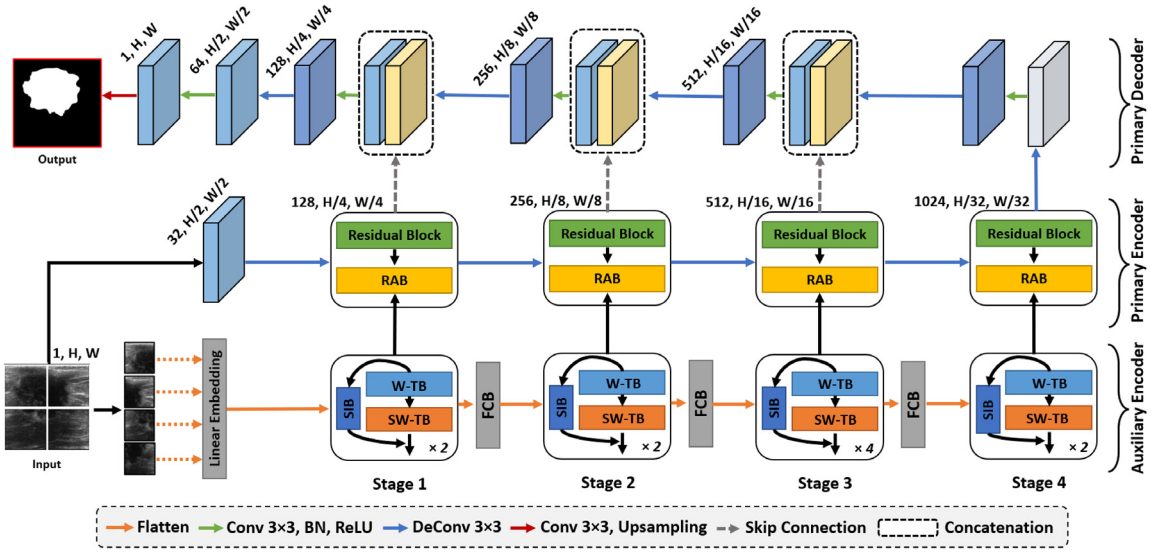

Fig. 2. The structure of the BTS-ST network consists of Residual Block, a Window-based transformer block (W-TB), Shifted window-based transformer block (SW-TB), and three novel blocks: Relational Aggregation Block (RAB), Spatial Interaction Block (SIB), and Feature Compression Block (FCB).

3.2. Swin-Transformer block

The MSA, MLP, and LN components of the conventional transformer block were already stated and visualized in Fig. 1(a). The output of s^l layer l is distributed as:

$$\tilde{\delta}^l = MSA(LN(\delta^{l-1})) + \delta^{l-1},$$

$$\delta^l = MLP(LN(\tilde{\delta}^l)) + \tilde{\delta}^l \quad (1)$$

The MSA is used by the classical transformer block to compute the self-attention globally between all provided tokens, and it causes the quadratic time complexity of a number of tokens along with the restriction of application scope. Normally, for dense predictions or higher resolution image tasks. For more efficiency, the ST network replaced the standard MSA block with two window-based multiple-head self-attention portioning, and both newly introduced blocks are named W-MSA and SW-MSA blocks, respectively.

In windows, where every window covers D × D patches, they engaged in self-attention while ignoring the tokens outside the window. The D value is set to 8 in all experiments for our easiness. As demonstrated in Fig. 1(b) W-MSA and SW-MSA are utilized in components of the ST in succession to improve knowledge connectivity between windows. The ST blocks were renamed to Window-based Transformers block (W-TB) and Shifted Window-based Transformers block (SW-TB) in order to make the distinction clearer. They are expressed as followed:

$$\tilde{\delta}^{l+1} = SW - MSA(LN(\delta^l)) + \delta^l,$$

$$\delta^{l+1} = MLP(LN(\tilde{\delta}^{l+1})) + \tilde{\delta}^{l+1} \quad (2)$$

Here, the output of W-TB is represented as δ^{l}, and the output feature of SW-TB is presented as δ^{l+1}

3.3. Spatial interaction block

Through the effective reduction of memory overhead, the ST blocks creates the relationship between patch tokens inside a limited window. Even though it uses the different execution schemes of the regular and shifted window, this method reduces Transformer's global modeling capabilities. Additionally, the obstacle of ground objects in breast tumor images resulting fuzzy borders that need some spatial information to be removed. Therefore, to further improve knowledge exchange while recording more accurate spatial knowledge, we introduced the spatial interaction block (SIB) across the W-TB and SW-TB. Transformer is more suited for breast tumor image segmentation tasks due to SIB, which proposed attention in two different spatial dimensions to take into account the relation among pixels rather than just patches tokens. The semantic structure of SIB is depicted in Fig. 3. In given phase nth reshaped the input $\delta^{l-1} \in \mathbb{R}(h_i \times w_i) \times c_1$ of W-TB into $z \in \mathbb{R}(h \times w) \times c_1$. Where, $c_1 = 2^{n-1}C_1$, and $h = \frac{h}{2^n+1}$, and $w = \frac{w}{2^n+1}$. In order to rebuild the systemic knowledge of the feature maps over a wide receptive field, the feature z is input into a dilated convolution of 3×3 with dilation size=2. Additionally, the number of channels is decreased to $c_1/2$ in order to lower the computational cost. The feature map statistics in the spatial direction are then obtained by applying the global average pooling method (*horizontal* and *vertical*). The calculating equation for the elements in every path is indicated specifically as:

$$\begin{aligned} v_{h_i}^{k} &= \frac{1}{w}\sum_{j=0}^{w-1}{}^{k}\widetilde{z}\,(i,j), \\ v_{w_i}^{k} &= \frac{1}{h}\sum_{j=0}^{h-1}{}^{k}\widetilde{z}\,(i,j) \end{aligned} \tag{3}$$

Here i, j and k are indices of horizontal and vertical side, and the number of channels. Where $0 \le i < h, 0 \le j < w, 0 \le K < C_1/2$. The convolution operation with dilation rate, batch normalization and GELU activation is represented as $\tilde{z} = f(z)$ and $f(.)$. Using Eq. (3) aggregate tensor in horizontal and vertical directions is denote as $\boldsymbol{v}_h \in \mathbb{R}^{h\times 1\times \frac{c1}{2}}$ and $\boldsymbol{v}_w \in \mathbb{R}^{1\times w\times \frac{c1}{2}}$. To multiply the two to obtained attention map $\boldsymbol{M}$ associated to the position $\boldsymbol{M} \in \mathbb{R}^{h\times w\times \frac{c1}{2}}$, v_h and v_w are converge the pix-level weights of the feature maps in spatial. The output feature map $\mathbf{F}$ of SIB achieved by embedding $\boldsymbol{M}$ and output of SW-Trans block s^{l+1}. Moreover, dimension of $\boldsymbol{M}$ required to be expanded using convolutional operation to match dimensions of feature δ^{l+1}. The $F \in \mathbb{R}^{h\times w\times c1}$ can described as following.

$$F = \delta^{l+1} \oplus \partial(v_h \otimes v_w) \tag{4}$$

Where element-wise addition represented as $\oplus$, and $\otimes$ symbol represent the matrix multiplication, and ∂ represent the convolution operation 1×1 with batch normalize and GLU.

3.4. Feature compression block

In studies like [14,34], the transformer network formed a sequential network by projecting and flattening patches, or merging the features of 2×2 similar image patches and performing linear processing. The segmentation of breast tumor images with small-scale regions and dense layer is difficult using these methods since they frequently result in the loss of significant structure and detailed information. Thus, in order to minimize the aforementioned issues and enhance the segmentation impact of smaller-scale regions, we designed the features compression block (FCB)

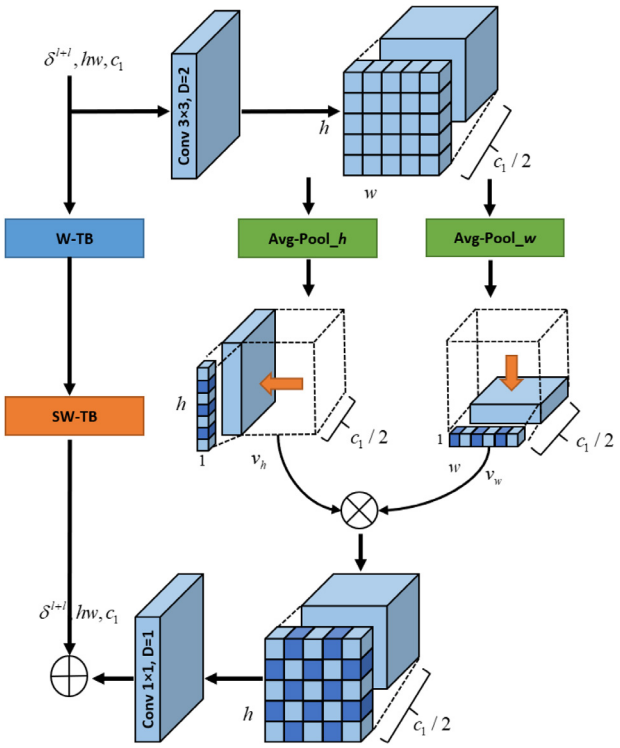

Fig. 3. The structure of spatial interaction block (SIB).

in the image patch token down-sampling of the ST network. FCB has two branches specifically, as seen in Fig. 4. The first is a bottleneck block with dilated convolution, which by enlarging the receptive field of the convolution, extensively collects the characteristics and hierarchical information of smaller regions. In the bottleneck block, 1×1 convolutional is used to improve the dimension, the 3×3 dilated convolution operation is used to obtain vast structural knowledge, and final 1×1 convolutional decrease the feature scale. The provided the output of s of step n, the output of following branch is $F_1 \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 2c1}$. Other branch present soft-pool [39] operation to achieve finer down-sampling. To preserve more specific information, soft-pooling can exponentially weight the pixels in the pooling kernel. For every single pixel in particular kernel neighborhood R, the computation of soft-pooling is illustrated in Eq. (5).

$$\tilde{\delta} = \sum_{i\in R} \frac{e^{\delta i} * \delta i}{\sum j \in R^{e\delta j}} \tag{5}$$

Features after the soft-pooling operation are fed to the convolution operation to achieve the output $F_2 = \mathbb{R}\frac{h}{2} \times \frac{w}{2} \times 2C_1$. The F_2 can be illustrated as Eq.(6)

$$F_2 = \partial(SoftPooling(\delta)) \tag{6}$$

In essence, one branch purpose is to achieve smaller-scale features, and the other branch purpose is to maintain details, both branch's are essential. As a result, the two branches are combined equally to provide the output L of FCB. The operation can be described as Eq. (7).

$$\boldsymbol{L} = F_1 \oplus F_2 \tag{7}$$

Here. $\oplus$ symbol denoted the element-level addition operation.

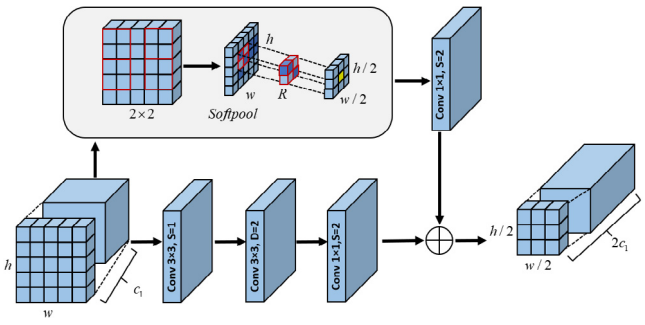

Fig. 4. The structure of feature compression block (FCB).

3.5. Relational aggregation block

In a network, confusion arises when similar distribution patterns are followed in regions but unidentical channels, as the CNNs-based primary encoder capture local knowledge constrained by the spatial dimension convolution kernel but does not explicitly model the relation among the channel dimensions [40]. In studies like [41,42] concluded that encoding the dependence of channel dimension can enhance feature discrimination. Moreover, we introduced relational aggregation block (RAB), whose semantics are represented in Fig. 5. First, channel dependency is used to capture the global features of the auxiliary encoders and embed them into the local features acquired from the primary encoder in order to highlight the significant and more representative channels from the whole feature maps. Additionally, RAB enhances the primary encoder's characteristics and adds the deformable convolution [43] pick to variously shaped object sections. To increase the accuracy of segmentation task of background regions with higher similarity in breast tumor images, more global discriminative features encoded using RAB. The main auxiliary encoder output of stage n, represented as A_n, S_n. To achieve the geometric diversity of breast tumor regions, input A_n is used in deformable convolution as denoted $\tilde{\boldsymbol{A}}_n = \partial(\boldsymbol{S}_n)$. The useful content in the image is the focus of the channel dependence since each channel of the feature maps can be considered as a feature detector. We use three pooling techniques to get a more detailed channel dependence. In order to determine the statistical properties of the feature maps on the channel, we first implement average and max-pooling layer. Then transfer the results to a share the FC layer. $P_{A\&M} \in \mathbb{R}^{1\times 1\frac{c1}{2}}$ is achieved by adding the two. In addition, a soft-pooling with exponential weight is added to a FC layer, denoted as $\boldsymbol{P}_S$, to generate the global weight descriptor at the same time. The entire operation can be explained as follow:

$$P_{A\&M} = \sigma(l_1(A.Pool(\tilde{\boldsymbol{S}}n))) + \sigma(l_1(Max.Pool(\tilde{\boldsymbol{S}}n))),$$

$$\boldsymbol{P}_S = \sigma(l_1(Soft.Pool(\tilde{\boldsymbol{S}}n))) \tag{8}$$

Here, ReLu is denoted as σ, and l_1 show fully connected layer with halved size. Each channel is multiply to optimize the descriptor, using $P_{A\&M}$ and P_S as depicted.

$$P = \delta(l_2(P_{A\&M} \odot P_S)) \tag{9}$$

Here, δ represent the sigmoid activation, l_2 show the FC layer with incremental size, and $\odot$ show element wise multiplication operation. Resulting $\boldsymbol{A}_n$ of deformable convolution achieved refined features, and the channel dependency $\boldsymbol{p}$ is multiplied as a weight. When the refined features are connected to the residual structure, which is represented by the symbol, the output feature $\boldsymbol{T}_n$ of the RAB is created Eq. (10)

$$\boldsymbol{T}_n = \tilde{\boldsymbol{A}}_n \oplus \tilde{\boldsymbol{S}}_n \oplus (\boldsymbol{P} \odot \tilde{\boldsymbol{A}}_n) \tag{10}$$

3.6. Sequential classification block

The feature representation of breast ultrasound, mammogram, and MRI-based images is obtained for the final classification process. The primary aim of the sequential classification block is to automatically classify breast tumor images (*Normal*, *Benign*, and *Malignant* cancer). In the classification task, the primary decoder of the proposed architecture (Fig. 2) is skipped and replaced with a sequential classification block. The sequential classification block is shown in Fig. 6.

Where, $l_{1024}, l_{512}, l_{128}, l_{64}$ and represents the linear layers in a sequential classification block. The batch normalization is represented as β_N Here, two dropouts $D_{0.7}$, and $D_{0.5}$ are also introduced. Where the sigmoid activation function is represented as δ. The final classification results are represented as three different classes of breast tumor images.

4. Experiments and results

This section detailed the experimental results achieved by five different breast tumor ultrasounds, mammograms, and MRI-based datasets to validate the proposed method's efficiency.

4.1. Datasets

Extensive experiments are conducted on three public and one private breast ultrasound segmentation datasets to validate the usefulness of the proposed method.

- The private breast ultrasound data set is downloaded from UltrasoundCases.info,[1] and images are collected by SonoSkills and Hitachi Medical Systems Europe. This dataset includes 811 ultrasound images, 358 benign tumor images, and 453 malignant tumor images. All ground truth images are privately annotated with the help of two senior radiologists for proper training in our method.
- The second dataset is a very small dataset that included 42 breast ultrasound images and was acquired from the Hospital of Shantou University. Because this dataset is small, we use this dataset for cross-dataset validation.
- The third dataset DDSM is based on a digital mammogram was created and maintained by the University of South Florida. The digital mammograms are retrieved 2500 different cases with 43 volumes. Furthermore, digital mammograms include the Mediolateral oblique (MLO) and Cranio-caudal (CC) of each patient. Our work used the mammograms acquired from the optimized and standardized Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset [44].
- The fourth dataset is a collection of MRI scan images that include tumors and healthy MRI scans. These MRI scans are acquired from the National Biomedical Imaging website, which United States National Cancer Institute originally owns [45]. The original dataset includes 1500 images of 5 different patients, and ground truth images are also provided. For our experiments, only 90 tumors images and their corresponding annotated images are used for experimental purposes.
- The fifth dataset consists of MRI images, which are collected from private radiologist. These acquired MRI images are 190, and all images belong to female breast cancer patients. For better training purposes, this dataset is merged with the fourth dataset.

[1] https://www.ultrasoundcases.info

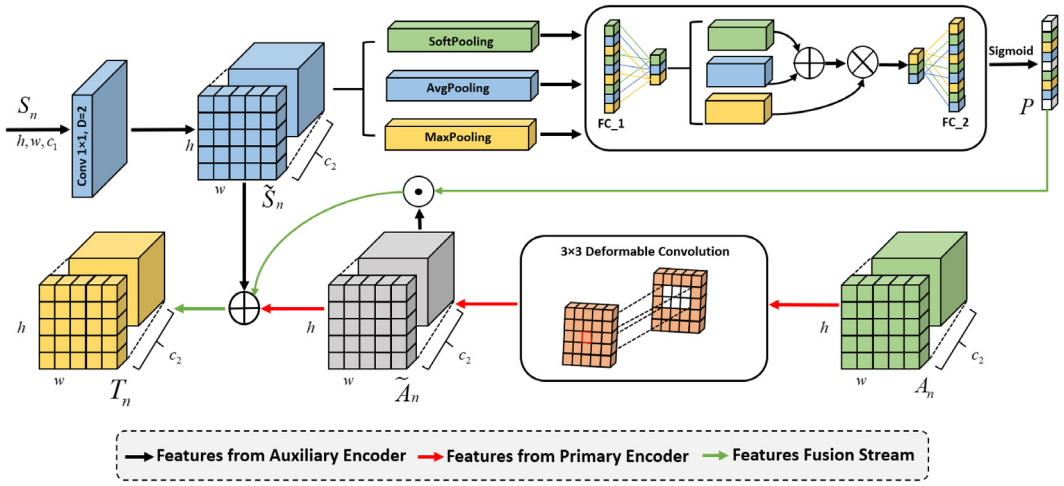

Fig. 5. The structure of relational aggregation block (RAB).

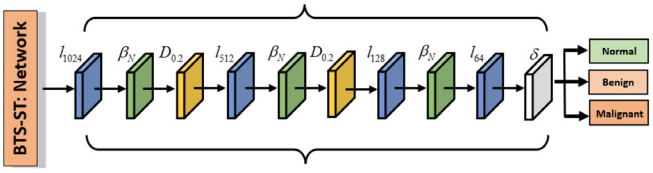

Fig. 6. The structure of sequential classification block.

4.2. Implementation details

The proposed BTS-ST network is implemented in a PyTorch library using NVIDIA RTX 2060 Super GPU. In our experiments, we adopted the SGD optimizer with momentum 0.9, and weight decay 1e-4 to train the model. Furthermore, we use the initial learning rate of 0.01 and adopt Poly decay strategy. The batch-size is fixed to 4 images, and the maximum epochs limit is set to 100. For the learning rate scheduler, we used available in PyTorch ReduceLROnPlateau with patience of 4 and a learning rate decrease factor of 0.2. In all experiments, the resolution of the train, validation, and test dataset were uniformly resized into 224×224. In experiments, 70% of the dataset is used during the training of the network, 15% portion is used for the validation task, and the rest of the 15% dataset is utilized for the testing process. The five-fold cross-validation scheme is employed for experiments to analyze the trained models performance. The training data expansion is achieved by using five different augmentation techniques such as horizontal flipping, vertical flipping, transpose, zoom, and gaussian noise, respectively.

4.3. Loss functions

In Ultrasound, Mammogram and MRI imaging, breast tumor pixels typically have a smaller region of interest than healthy tissues, and resulting class imbalance challenges. The network could be trapped in local minima if trained on the unbalanced dataset. As a result, the prediction results of networks are heavily skewed toward the majority class. The detrimental impact of class imbalance is mitigated by introducing a dice loss function, Eq. (11).

$$\ell_{Dice} = 1 - \frac{1 + 2\sum_{i=1}^{N} \hat{p}_i . \widehat{g_i}}{1 + \sum_{i=1}^{N} \hat{p}_i^2 + \sum_{i=1}^{N} \hat{g}_i^2} \tag{11}$$

Where summation is denoted as $\sum$ and runs N numbers of pixels over a probabilistic prediction map, here, $\hat{p} \in \{0, 1\}_{i=1}^{N}$ and $\hat{g} \in \{\hat{g}\}_{i=1}^{N}$ are the output of final pixel-wise sigmoid and binary masks, respectively.

For the classification task, we adopted Label Smoothing Cross-Entropy (LSCE) loss function to overcome the overfitting problem of the classification network.

$$l_{LSCE} = -\sum_{c=1}^{N} y_c^{LS} \log(g(\hat{y}_c)) \tag{12}$$

In Eq. (12) y_c^{LS} represents the soft targets generated by altering the ground-truth distribution by a uniform distribution sequence of $\frac{\varepsilon}{N}$; N representing the total classes and ε (*here*, 0.1) labeled as the smoothing factor. Where $y_c^{LS} = (1 - \varepsilon)y_c + \frac{\varepsilon}{N}$ and $g(.)$ is representing the softmax operation.

4.4. Evaluation metrics

The efficiency of the proposed segmentation and classification networks evaluated using six different evaluation metrics are utilized to assess the performance, which includes F1score, Jaccard index, Precision, Recall, AUC, and HD^{95}. Precision denoted how many tumor images were correctly classified as the positive output of all positive numbers. The highest the value of Precision, the greater the capability to detect tumors images. The fraction of the sample having a malignant tumor that is appropriately diagnosed as a malignant case is known as Recall. The stronger the capability to detect malignant tumors, the higher the Recall. The F1score is the *harmonic* Mean of the Precision and Recall rates, which is adopted to balance the two metrics for a complete evaluation. The area under the ROC curve, or AUC, is a performance metric that measures the system overall quality. Each of the six metrics has its own significance in assessing the quality of different types of work. The metrics relevant formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{F1Score} = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{15}$$

$$JSC(\widetilde{GT}, \widetilde{PM}) = \frac{|\widetilde{GT} \cap \widetilde{PM}|}{|\widetilde{GT} \cup \widetilde{PM}|} \tag{16}$$

Where $\widetilde{PM}$ and $\widetilde{GT}$ are represents, the region segmented by a network and the ground truth images.

Table 1
BST-ST network basic configuration of Swin-Transformer block.

Parameter	Configuration
Hidden size C1	96
Window Size	8
Number of layers	{2,2,8,2}
Head layers	{3,6,12,24}

Hausdorff distance (HD) is used a as performance measure to compute the distance between two-point sets. Where $P\tilde{M}$ and $G\tilde{T}$ are represents the probability maps and the ground truth. HD mathematical formulation is defined in Eq. (17) and (18). In our research, In HD^{95} is used, where 95-*th* percentile are used rather than using the maximal value.

$$HD'(G\tilde{T}, P\tilde{M}) = \max_{i \in G^t} \min_{j \in P^t} \|i - j\|2 \tag{17}$$

$$HD(G\tilde{T}, P\tilde{M}) = max(HD'(G\tilde{T}, P\tilde{M}), HD'(P\tilde{M}, G\tilde{T})) \tag{18}$$

4.5. Ablation study

The effectiveness of the presented method is demonstrated for breast tumor segmentation and classification, we developed multiple ablation experiments. We observed a significant improvement in the performance of proposed blocks compared with the baseline network. In BTS-ST network, the primary encoder incorporates ResNet-18 (*half-compressed*) as the backbone due to memory limitations. The Swin-Transformer is employed with the following tiny configuration, as shown in Table 1.

4.5.1. Effectiveness of auxiliary encoder

Table 2. the second row shows the segmentation results of AE blocks on the Private BUS dataset and STU-Hospital BUS dataset. In the private BUS dataset case, we have achieved improvement in terms of F1(0.52%), JI(0.72%), Pre(0.2%), Rec(0.83%), and HD^{95}(0.25%). Similarly, we obtained performance improvement of F1(0.76%), Pre(0.72%), and Rec(0.78%) on the STU-Hospital BUS dataset. In Table 3. CBIS-DDSM Mammogram dataset, we have achieved improvement in terms of F1(0.5%), JI(0.05%), Pre(0.44%), and Rec(0.58%). Similarly, we obtained performance improvement of F1(0.46%), JI(0.86%), Pre(0.65%), Rec(0.32%) and HD^{95}(0.08%) on RIDER Breast MRI dataset. Results demonstrate that our adopted auxiliary encoder is the helping network to improve the global context information in the proposed architecture.

4.5.2. Effectiveness of dual encoder block

Table 2. third rows show the segmentation results of DE blocks on the Private BUS dataset and STU-Hospital BUS dataset. In the private BUS dataset case, we have achieved improvement in terms of F1(1.12%), JI(1.56%), Pre(1.0%), Rec(1.18%), and HD^{95}(0.392%). Similarly, we obtained performance improvement of F1(1.16%), Pre(1.37%), Rec(1.0%) and HD^{95}(0.091%) on the STU-Hospital BUS dataset. In Table 3, CBIS-DDSM Mammogram dataset, we have achieved improvement in terms of F1(0.96%), JI(0.1%), Pre(0.91%), and Rec(1.01%). Similarly, we obtained performance improvement of F1(0.98%), JI(1.01%), Pre(0.96%), and Rec(1.0%) on RIDER Breast MRI dataset. The results reflect that after using the DE blocks to incorporate more global contextual knowledge, the segmentation accuracy of tumor segmentation improves in the majority of performance measures. Results demonstrate our adopted auxiliary encoder, which is composed of Swin Transformer blocks, improving global context information of the CNNs-based primary encoder.

4.5.3. Effectiveness of relational aggregation block

Table 2. fourth rows show the segmentation results of RAB blocks on Private BUS dataset and STU-Hospital BUS dataset. In the private BUS dataset case, we have achieved improvement in terms of F1(1.56%), JI(3.14%), Pre(6.75%), and HD^{95}(1.87%). Similarly, we obtained performance improvement of F1(3.83%), JI(8.2%), Pre(1.52%), Rec(5.61%) and HD^{95}(14.793%) on the STU-Hospital BUS dataset. In Table 3 CBIS-DDSM Mammogram dataset, we have achieved improvement in term of JI(0.46%), and Rec(0.11%). Similarly, we obtained performance improvement of F1(1.74%), JI(3.9%), Pre(1.06%), Rec(2.3%), and HD^{95}(0.25%) on RIDER Breast MRI dataset. The results reflect that after using the RAB block to incorporate more global contextual knowledge, the segmentation accuracy of tumor segmentation improves in the majority of performance measures.

4.5.4. Effectiveness of spatial interaction block

Table 2. fifth rows reflect the segmentation results of SIB blocks on the Private BUS dataset and STU-Hospital BUS dataset. In the private bus dataset case, we have obtained improvement in terms of F1(5.29%), JI(9.35%), Pre(5.53%), Rec(5.05%) and HD^{95}(19.37%). Similarly, we achieve better performance F1(5.07%), JI(8.21%), Pre(1.99%), Pre(5.53%), Rec(7.49%) and HD^{95}(15.32%) in the case of the STU-Hospital BUS dataset. In Table 3 CBIS-DDSM Mammogram dataset, we have obtained improvement in terms of F1(0.91%), JI(1.46%), Pre(1.07%), Rec(0.69%) and HD^{95}(0.45%). Similarly, we achieved performance improvement of JI(4.12%), and HD^{95}(0.07%) on RIDER Breast MRI dataset. The results show that using SIB with RAB block improves the segmentation results on various performance measures.

4.5.5. Effectiveness of feature compression block

Table 2. sixth rows show the segmentation results of FCB blocks on the Private BUS dataset and STU-Hospital BUS dataset. In the private bus dataset case, we have achieved improvement in terms of F1(5.89%), JI(10.53%), Pre(6.4%), Rec(5.39%), and HD^{95}(21.49%). Similarly, we obtained performance improvement of F1(5.54%), JI(8.86%), Pre(6.87%), Rec(5.99%) and HD^{95}(19.76%) on the STU-Hospital BUS dataset. In Table 3 CBIS-DDSM Mammogram dataset, we have achieved improvement in terms of F1(1.95%), JI(2.64%), Pre(1.99%), Rec(1.87%), and HD^{95}(1.55%).

Table 2
Ablation experiments of the proposed BTS-ST network with various combinations of BL (Primary encoder and decoder architecture), AE (auxiliary encoder), DE (Primary encoder with composition of auxiliary encoder and decoder architecture), RAB (Relational Aggregation Block), SIB (Spatial Interaction Block), and FCB (Feature Compression Block) on Private BUS dataset and STU-Hospital dataset.

Method	Private BUS dataset					STU-Hospital BUS dataset				
	F1 ↑	JI ↑	Pre ↑	Rec ↑	HD^{95} (mm)↓	F1 ↑	JI ↑	Pre ↑	Rec ↑	HD^{95} (mm)↓
BL	0.8497	0.7238	0.8569	0.8428	28.355	0.8094	0.6798	0.8760	0.7523	28.091
BL+AE	0.8549	0.7310	0.8589	0.8511	28.105	0.8170	0.6741	0.8832	0.7601	28.503
BL+DE	0.8609	0.7394	0.8674	0.8546	27.963	0.8210	0.6758	0.8897	0.7623	28.000
BL+DE+RAB	0.8653	0.7552	**0.9244**	0.8134	30.228	0.8477	0.7618	0.8912	0.8084	13.298
BL+DE+RAB+SIB	0.9026	0.8173	0.9122	0.8933	8.9781	0.8601	0.7619	0.8959	**0.8272**	12.766
BL+DE+RAB+SIB+FCB	**0.9086**	**0.8291**	0.9209	**0.8967**	**6.8596**	**0.8648**	**0.7682**	**0.9247**	0.8122	**8.3310**

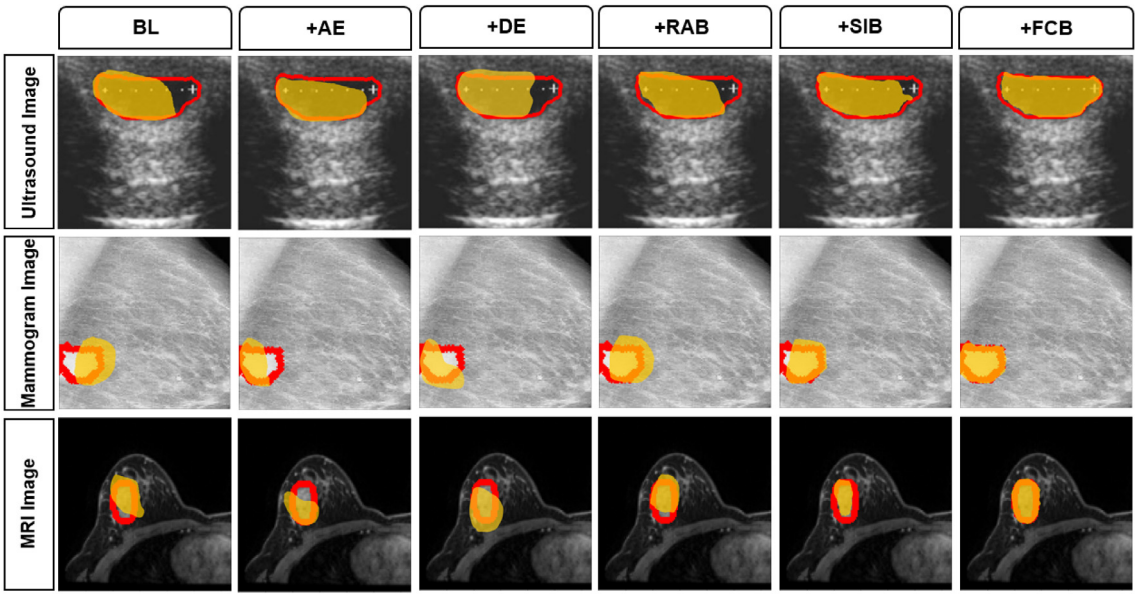

Fig. 7. The visual effectiveness of AE, DE, RB, SIB, and FCB blocks. The red contour represents the actual ground truth, and the yellow region represents the predicted results.

Table 3
Ablation experiments of the proposed BTS-ST network with various combinations of BL (Primary encoder and decoder architecture), AE (auxiliary encoder), DE (Primary encoder with composition of auxiliary encoder and decoder architecture), RAB (Relational Aggregation Block), SIB (Spatial Interaction Block), and FCB (Feature Compression Block) on CBIS-DDSM Mammogram Dataset and RIDER Breast MRI Dataset.

Method	CBIS-DDSM mammogram dataset					RIDER breast MRI dataset				
	F1 ↑	JI ↑	Pre ↑	Rec ↑	HD^{95} (mm)↓	F1 ↑	JI ↑	Pre ↑	Rec ↑	HD^{95} (mm)↓
BL	0.7516	0.5710	0.7089	0.8000	10.110	0.8027	0.6410	0.8505	0.7600	2.2709
BL+AE	0.7566	0.5715	0.7133	0.8058	10.997	0.8073	0.6496	0.8570	0.7632	2.2701
BL+DE	0.7612	0.5720	0.7180	0.8101	10.900	0.8125	0.6511	0.8601	0.7700	2.2901
BL+DE+RAB	0.7507	0.5756	0.7064	0.8011	10.628	0.8201	0.6800	0.8611	0.7830	2.0122
BL+DE+RAB+SIB	0.7607	0.5856	0.7196	0.8069	9.6542	0.7789	0.6822	0.8091	0.7510	2.1950
BL+DE+RAB+SIB+FCB	**0.7711**	**0.5974**	**0.7288**	**0.8187**	**8.5544**	**0.8336**	**0.7401**	**0.8811**	**0.7911**	**1.2358**

Similarly, we obtained performance improvement of F1(3.09%), JI(9.91%), Pre(3.06%), Rec(3.11%), and HD^{95}(0.97%) on the RIDER Breast MRI dataset. The results indicate that using FCB with RAB+SIB block enhanced the segmentation results on different performance measures. Furthermore, the FCB block is originally designed to accurately segment the small-size tumor regions, such kind of tumors exists in Mammogram and MRI-based images.

4.5.6. Ablation study visual results

To analyze the visual significance of each block DE, RB, SIB, and FCB, we presented results in Fig. 7. All three images are acquired from three different modalities Ultrasound, Mammograms, and MRI system. The visual results are projected on three sample images. The results demonstrate that each component proves its significance visually. The BL combination with DE+RAB+SIB, and FCB embeds more global context information capability in our proposed BTS-ST network.

4.6. Results comparison with other state-of-the-art methods

To prove the superiority of the proposed method, we performed three experiments on the Private breast ultrasound dataset, RIDER MRI dataset, and CBIS-DDSM mammogram dataset, respectively. All three datasets are available for both segmentation and classification tasks. For the segmentation task, our proposed method results are compared with U-Net [11], U-Net++ [46], DeepLabv3+ [47] (*backbone: ResNet-101*), SK-U-Net [48], MSU-Net [49], MACUNet [50], TransUNet [37], and TransFuse [38], respectively. Similarly, our method classification results are fairly compared with ResNet-50 [51], DenseNet [52], Inceptionv3 [53], MobileNetv3 [54], EfficientNet [55], Transformer (vit_base_patch16_224) [14], Transformer (vit_small_patch16_224) [14] and Transformer (vit_large_patch16_224) [14] respectively. For a fair comparison with other state-of-the-art methods, we used ReduceLROnPlateau function, which can reduce the learning rate when a metric has stopped improving. We selected the Dice-based loss function because it outperforms all other state-of-the-art methods. The batch size remains 4 images for all state-of-the-art methods and with the early stopping mechanism. The epoch size is set to 100 because most models converge before reaching 100 epochs. We also use the same dataset distribution (train:70%, val:15%, and test:15%) for other state-of-the-art networks. Table 4 shows the experimental results of the Private BUS dataset. The proposed BTS-ST network achieved the highest results of 0.908, 0.978, and 6.8596 in terms of F1, AUC and HD^{95} measures. The second-best results are observed in the case of DeepLabv3+ [47], with 0.896, and 0.977 in terms of F1, and AUC scores. However, SK-U-Net [48] has achieved the highest result term of JI (0.831). The results analysis also shows that U-Net [11] and U-Net++ [46] results are comparatively worse than all other methods. For classification, the highest results are achieved in terms of F1(0.856), Pre(0.792), and AUC(0.813). The second best results are obtained by Transformer (vit_large_patch16_224) [14] in terms of F1(0.739), and AUC(0.654). However, Transformer(vit_base_patch16_224) [14] achieved the second highest performance in Pre(0.714), and DenseNet [52] achieved in terms of Rec(0.977). Table 5 shows the

Table 4
Statistical comparative results of BTS-ST network with other state-of-the-art methods on the Private BUS dataset.

Method	Segmentation				Method	Classification			
	F1 ↑	JI ↑	AUC ↑	HD^{95} (mm)↓		F1 ↑	Pre ↑	Rec ↑	AUC ↑
U-Net [11]	0.699	0.576	0.920	36.204	ResNet-50 [51]	0.708	0.555	**0.980**	0.500
U-Net++ [46]	0.639	0.503	0.898	61.259	DenseNet [52]	0.738	0.594	0.977	0.572
DeepLabv3+ [47]	0.896	0.822	0.977	30.130	Inceptionv3 [53]	0.666	0.611	0.733	0.575
SK-U-Net [48]	0.894	**0.831**	0.976	28.100	MobileNetv3 [54]	0.708	0.555	0.980	0.050
MSU-Net [49]	0.895	0.828	0.977	17.072	EfficientNet [55]	0.728	0.629	0.866	0.613
MACUNet [50]	0.881	0.803	0.973	23.554	Transformer(vit_base_patch16_224) [14]	0.688	0.714	0.665	0.650
TransUNet [37]	0.871	0.792	0.969	20.094	Transformer(vit_small_patch16_224) [14]	0.666	0.663	0.670	0.635
TransFuse [38]	0.881	0.804	0.974	18.854	Transformer(vit_large_patch16_224) [14]	0.739	0.651	0.856	0.654
BTS-ST (our)	**0.908**	0.829	**0.978**	**6.8596**	**BTS-ST (our)**	**0.856**	**0.792**	0.933	**0.813**

Table 5
Statistical comparative results of BTS-ST network with other state-of-the-art methods on the RIDER Breast dataset.

Method	Segmentation				Method	Classification			
	F1 ↑	JI ↑	AUC ↑	HD^{95} (mm)↓		F1 ↑	Pre ↑	Rec ↑	AUC ↑
U-Net [11]	0.741	0.640	0.990	8.5750	ResNet-50 [51]	0.686	0.900	0.555	0.777
U-Net++ [46]	0.810	0.731	0.991	13.830	DenseNet [52]	0.886	0.810	**0.980**	0.900
DeepLabv3+ [47]	0.819	0.722	0.993	7.1188	Inceptionv3 [53]	0.884	0.888	0.881	0.894
SK-U-Net [48]	0.808	0.717	0.989	6.1078	MobileNetv3 [54]	0.809	0.818	0.801	0.900
MSU-Net [49]	0.817	0.728	0.993	4.3082	EfficientNet [55]	0.848	0.901	0.802	0.941
MACUNet [50]	0.782	0.685	0.997	8.6964	Transformer(vit_base_patch16_224) [14]	0.895	0.900	0.891	0.943
TransUNet [37]	0.778	0.672	0.981	7.7843	Transformer(vit_small_patch16_224) [14]	0.765	0.902	0.665	0.877
TransFuse [38]	0.793	0.695	0.991	8.9965	Transformer(vit_large_patch16_224) [14]	0.779	0.900	0.687	**0.951**
BTS-ST (our)	**0.833**	**0.741**	**0.999**	**1.5808**	**BTS-ST (our)**	**0.937**	**0.980**	0.898	0.944

experimental results of the RIDER Breast dataset. The proposed BTS-ST network achieved the highest results of 0.833, 0.741, 0.999, and 1.5808 in terms of F1, JI, AUC, and HD^{95} measures. The second best results are observed in the case of U-Net++ [46], DeepLabv3+ [47], MSU-Net [49] in term of performance measures JI(0.731), F1(0.819), and HD^{95}(4.3082). For classification, the highest results are achieved in terms of Pre(0.937), and Rec(0.980). Similarly, Transformer(vit_base_patch16_224) [14] and Transformer (vit_small_patch16_224) [14] achieved the highest results in terms of F1(0.895) and Pre(0.902). However, BTS-ST also achieved a second better performance in terms of Rec(0.898) and AUC(0.944). Table 6 shows the experimental results of the CBIS-DDSM Mammogram dataset. The proposed BTS-ST network achieved the highest results of 0.771, 0.597, 0.996, and 8.5544 in terms of F1, JI, AUC, and HD^{95} measures. The second best results are observed in the case of MSU-Net [49] in terms of performance measures JI(0.594) and AUC(0.994). We have also achieved second-best results in F1(0.737), HD^{95}(9.7780) in case of SK-U-Net [48] and DeepLabv3+ [47]. For classification, the highest results are achieved in terms of F1(0.6838), Pre(0.6463), and AUC(0.7054). Similarly, the second highest results achieved by Transformer(vit_small_patch16_224) [14], ResNet-50 [51], EfficientNet [55] in term of F1(0.6801), Pre(0.6400), and AUC(0.6900). In segmentation results, MACUNet [50] remains second-best performer on the RIDER Breast dataset and CBIS-DDSM Mammogram dataset. The MACUNet [50] architecture is originally based on multiple convolution sequences that are used to extract more semantic features from the images. Second, the convolution kernel with different receptive fields is used to make features more diverse. Similarly, DeepLabv3+ (*backbone: ResNet-101*) [47] architecture remains second-best performer because the network incorporates depth-wise separable convolution to both Atrous Spatial Pyramid Pooling and decoder modules, resulting in a faster and stronger encoder–decoder network.

4.7. Computational efficiency comparison with state-of-the-art methods

Table 7 is presented to compare the proposed BTS-ST computational comparison with other state-of-the-art methods in a similar environment. The parameters, model size, training times, inference times, and mean FPS are used as primary indicators to compute model efficiency. The proposed method parameters size is comparatively better than DeepLabv3+ [47], MSU-Net [49], and TransUNet [37]. However, SK-U-Net [48] is trainable parameters are lesser than all other methods. Similarly, BTS-ST requires a memory size of 158 MB, which is comparatively superior to DeepLabv3+ [47], MSU-Net [49], and TransUNet [37]. But comparative results indicate that MACUNet [50] require lesser model size as compared to other methods. The proposed method is required 65m to train the model, which is better training time compare with U-Net++ [46], DeepLabv3+ [47], SK-U-Net [48], MSU-Net [49], TransUNet [37], and TransFuse [38]. The proposed BTS-ST inference time is 42.602 ms which is comparatively better than DeepLabv3+ [47], TransUNet [37], TransFuse [38]. However, U-Net [11] inference time is 17.431 ms which is comparatively better than all methods. In Mean FPSs, our method achieved 24.862 FPSs, which is comparatively superior to DeepLabv3+ [47], SK-U-Net [48], and TransUNet [37]. However, U-Net [11] achieved 57.900 Mean FPSs which is better than all other networks.

4.8. Visual comparison with other state-of-the-art methods

The visual comparison of the proposed BTS-ST method is compared with other state-of-the-art methods in Fig. 8. We have selected two challenging sample images from three different datasets. The solid red (ground truth) and aqua blue (prediction) contours are drawn against tumor regions. The first two rows of images belong to a private ultrasound dataset, and eight different results are presented. The visuals demonstrate that U-Net [11], U-Net++ [46], SK-U-Net [48], and MACUNet [50] achieved the worst results, and the prediction contour totally missed the ground truth region. However, DeepLabv3+ [47], TransUNet [37], and TransFuse [38] results are satisfactory in both cases. The proposed BTS-ST achieved the highest performance in both cases and achieved overlapping results with the highest precision. The third and fourth rows of images belong to CBIS-DDSM mammogram images. Here, MACUNet [50], SK-U-Net [48] and TransFuse [38] results are worse as compared with other methods. However, our proposed BTS-ST method achieved higher performance on

Table 6
Statistical comparative results of BTS-ST network with other state-of-the-art methods on the CBIS-DDSM Mammogram dataset.

Method	Segmentation				Method	Classification			
	F1 ↑	JI ↑	AUC ↑	HD^{95} (mm)↓		F1 ↑	Pre ↑	Rec ↑	AUC ↑
U-Net [11]	0.708	0.568	0.984	15.834	ResNet-50 [51]	0.6486	0.6400	0.6575	0.6820
U-Net++ [46]	0.714	0.557	0.950	12.876	DenseNet [52]	0.6394	0.6351	0.6438	0.6751
DeepLabv3+ [47]	0.729	0.584	0.990	9.7780	Inceptionv3 [53]	0.6315	0.6075	0.6575	0.6602
SK-U-Net [48]	0.737	0.590	0.980	11.765	MobileNetv3 [54]	0.6580	0.6219	0.6986	0.6808
MSU-Net [49]	0.700	0.594	0.994	10.789	EfficientNet [55]	0.6670	0.6317	0.7086	0.6900
MACUNet [50]	0.701	0.561	0.989	9.8456	Transformer(vit_base_patch16_224) [14]	0.5479	0.5966	0.5067	0.6174
TransUNet [37]	0.729	0.584	0.990	9.8531	Transformer(vit_small_patch16_224) [14]	0.6801	0.5564	**0.8768**	0.6610
TransFuse [38]	0.731	0.586	0.992	9.8654	Transformer(vit_large_patch16_224) [14]	0.6595	0.5337	0.8629	0.6324
BTS-ST (our)	**0.771**	**0.597**	**0.996**	**8.5544**	**BTS-ST (our)**	**0.6838**	**0.6463**	0.7260	**0.7054**

Table 7
The computational comparison of presented BTS-ST network with previous state-of-the-art methods.

Method	Parameters	Model size	Training times	Inference times	Mean FPS
U-Net [11]	7.74×10^6	30 MB	35 m 26 s	17.431 ms	57.900
U-Net++ [46]	9.03×10^6	35 MB	70 m 20 s	23.711 ms	43.543
DeepLabv3+ [47]	5.83×10^7	227 MB	80 m 28 s	50.143 ms	19.683
SK-U-Net [48]	3.92×10^6	46 MB	67 m 30 s	41.102 ms	22.761
MSU-Net [49]	4.70×10^7	179 MB	128 m 43 s	26.867 ms	37.608
MACUNet [50]	5.16×10^6	21 MB	55 m 26 s	36.500 ms	27.665
TransUNet [37]	9.62×10^7	646 MB	226 m 10 s	71.100 ms	14.411
TransFuse [38]	2.61×10^7	84.2 MB	78 m 20 s	46.110 ms	22.904
BTS-ST (our)	4.68×10^7	158 MB	65 m 23 s	42.602 ms	24.862

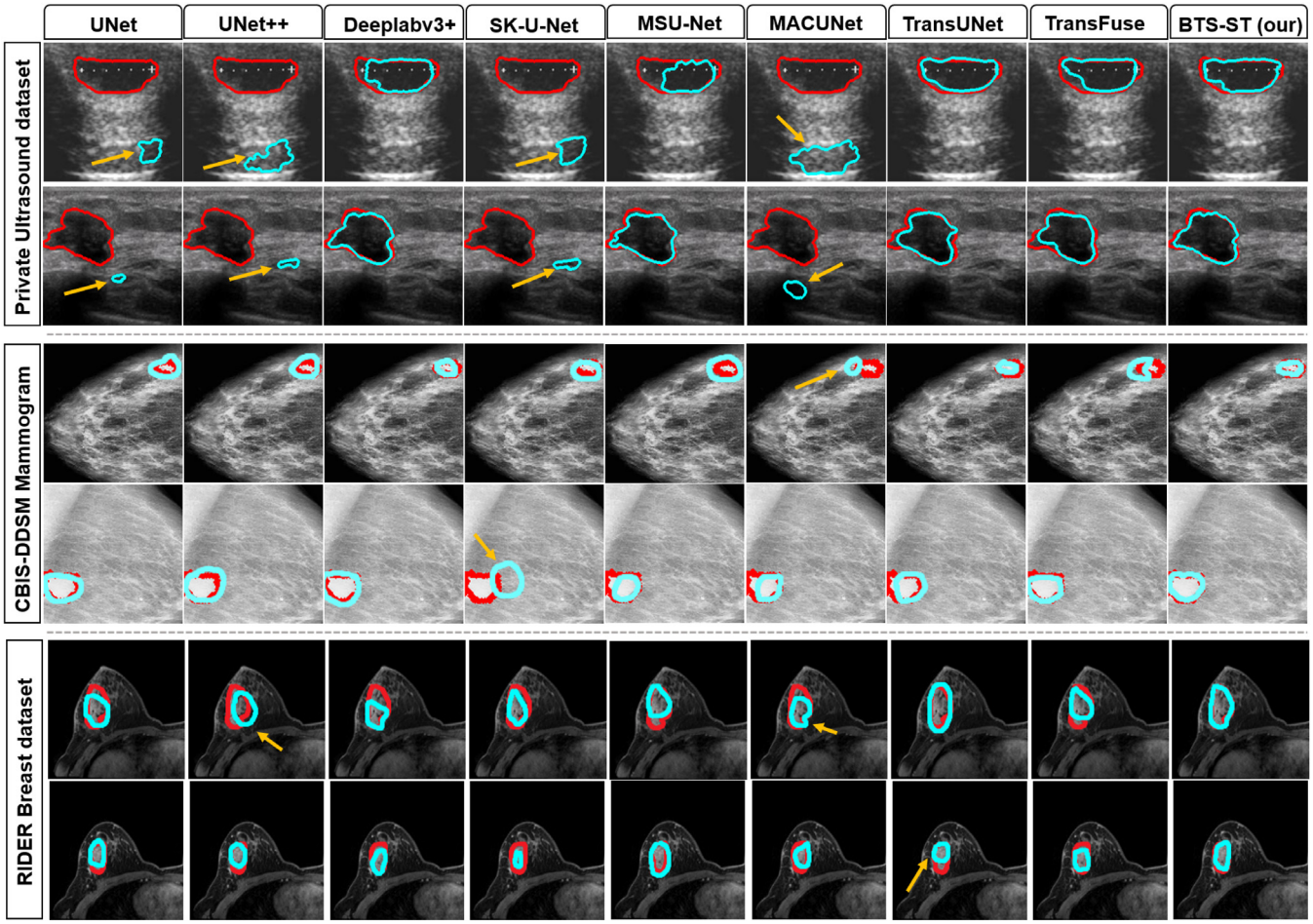

Fig. 8. The visualization of the segmentation results on three different Private BUS datasets, CBIS-DDSM Mammogram dataset, and RIDER Breast MRI dataset. The solid red contour represents the ground truth regions, and the aqua blue contour draws the predicted segmentation result.

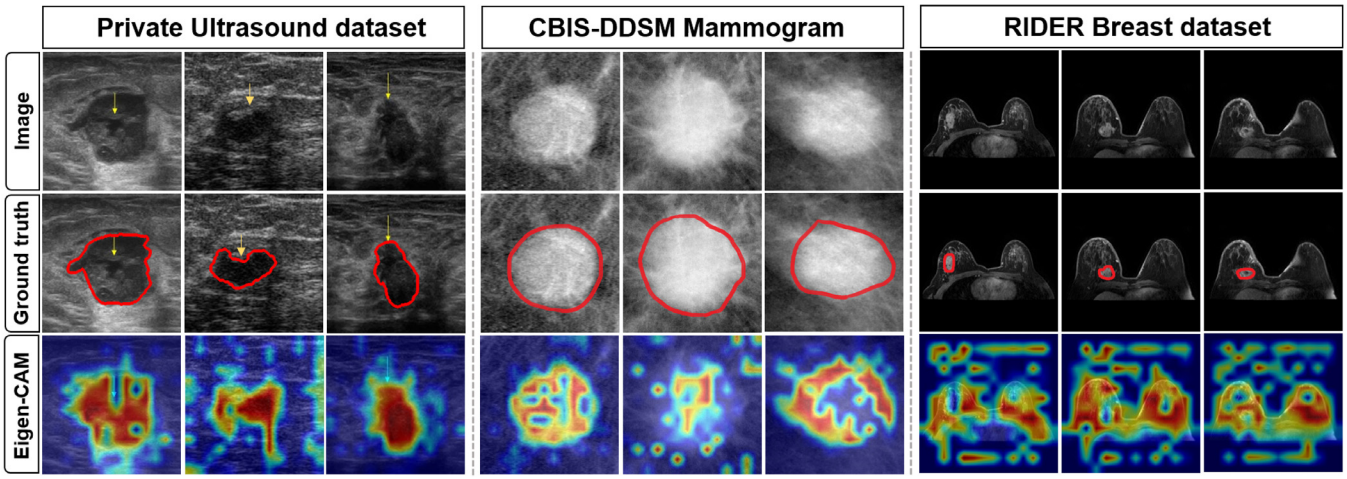

Fig. 9. Visualization of 2D Private ultrasound images, mammogram images, and MRI images. Here, the first row represents the original input images, the second row represents the ground truth images, and the third row represents the Eigen-CAM [56] image.

both images with outstanding accuracy. Similarly, the fifth and sixth rows represent sample images of the RIDER MRI dataset. Where SK-U-Net [48] and TransUNet [37] did not perform well as compared with other methods. However, our presented BTS-ST achieved supervisor results, and the prediction contour is fully overlapped on ground truth images. In literature, a class activation map (CAM) is used to visualize what a deep learning model learns from the image data or why it behaves poorly in different circumstances. In our research, we used Eigen-CAM [56] to visualize the breast tumor region of interest to display how features learned from the convolutional layers of BTS-ST. As we can see in Fig. 9, the first row represents the original input images of Private BUS dataset, RIDER MRI dataset, and CBIS-DDSM mammogram dataset. Similarly, the second row represents the ground truth images of the corresponding dataset. The final third row represents the Eigen-CAM [56] results of the proposed classification network. The third row of Private BUS dataset images shows that our method learns properly discriminative features from the region of interest. Similarly, the third row of mammogram images also shows discriminative patterns which are learned by our proposed network. However, the third row of MRI images is more complex to understand, but it shows that the network properly learns the pattern from two breast regions and makes a decision based on specific patterns.

5. Discussion and limitations

The most prevalent kind of cancer in women is breast cancer, which affects 1 in 8 of them over the course of their lifetimes. Finding early signs of breast cancer in women is crucial in clinical practice, and it could be beneficial in reducing the mortality rate. In the breast cancer screening process, Ultrasound, MRI, and Mammogram modalities are most commonly used during the screening process. Despite their exceptional representational capability, CNN-based approaches frequently have limitations when it comes to capturing explicit long-range interactions. These structures thus frequently yield subpar results, particularly significant size of tumors, shape, and textural variation between breast tumor images. To overcome such limitations, previous studies were introduced to establish self-attention mechanisms based on CNNs features. With recent advancements, Transformers proposed for sequence-to-sequence prediction, have appeared as an alternative solution that used dispensed convolution operators solely and entirely based on attention-based mechanisms instead. In this research, we proposed a BTS-ST network for accurate segmentation and classification of breast tumor images. The dual encoder mechanism incorporates the U-Net [11] encoder with the Swin-Transformer encoder to improve the encoder mechanism to use more global features and improve the discriminative features. To be more precise, the primary encoder is directed at obtaining more discriminative features using the proposed relationship aggregation block (RAB) using global features. Additionally, the spatial interaction block (SIB) and the feature compression block (FCB) were created to further enhance Swin Transformer's capacity for global modeling. The SIB block creates pixel-level information sharing, doing away with Swin Transformer's window restriction and solving the issue of semantic ambiguity brought on by occlusion. The patch token down-sampling for small-scale tumor regions uses the feature compression block (FCB) to keep as many precise features as possible. The ablation experiments are performed in Tables 2, 3, and Private BUS Dataset, STU-Hospital BUS Dataset, CBIS-DDSM Mammogram Dataset, and RIDER Breast MRI Dataset are used for the experiments. The ablation experiments demonstrate that each block improves the overall performance of segmentation on most of the performance measures. It is proved that dual encoders can aggregate more information that is conducive to tumor prediction by cascading hierarchically. In addition, results have shown that after using RAB to embed more global context information, the segmentation accuracy effectively improved. Furthermore, results indicate that FCB is beneficial in improving the segmentation accuracy of small-scale tumor regions. It can be observed from a visual demonstration in Fig. 7 that the introduction of the SIB block effectively diminishes the negative impact of the mutual occlusion of foreground and background pixels. However, in the ablation study, we also observed failure cases on RIDER Breast MRI Dataset in Table 3. Where SIB block is poorly performed compared with other four datasets. As compared with Ultrasound and Mammogram, MRI breast tumor images region of interest is too small. The BTS-ST network is compatible with 224×224 image resolutions, which is also compatible with original architecture of ResNet-18 and Swin-Transformer. The structure of SIB is based on an average-pooling operation followed by a convolutional operation. The average-pooling operation is lossy and does not preserve the spatial

information well by reduction of spatial resolution. Resulting in the SIB block losing its performance due to tiny and smaller regions of tumors. However, when we embedded SIB with FCB block, overall results improved. Tables 4, 5, and 6 are presented to compare the proposed method results with other state-of-the-art methods on the same parameter settings. The results demonstrate that our proposed method outperforms as compared with all other methods on both segmentation and classification tasks. Similarly, Table 5 results also show that our classification results of Recall and AUC are lower than DenseNet [47] and Transformer(vit_large_patch16_224) [10]. Table 6 shows the proposed BTS-ST computational comparison with other state-of-the-art methods. The computational comparison shows that our method has the advantage in most of the parameters. However, U-Net [11] is more lightweight and less computationally expensive. But our proposed method BTS-ST achieved better performance and the highest Dice score on most of the datasets (Private BUS Dataset, STU-Hospital BUS Dataset, CBIS-DDSM Mammogram Dataset, and RIDER Breast MRI). It is comparatively better to more accurate in terms of segmentation and classification results. In visual performance comparison, Fig. 8 is illustrated to show how our proposed method is better for drawing prediction contour accurately on ground truth image. The visual results show that our proposed method contour is properly overlapping in most of the cases (Ultrasound dataset, MRI dataset, and Mammogram dataset). Furthermore, we visualized the gradient-weighted class activation mapping (*using* Eigen-CAM [56]) obtained by our network, as shown in Fig. 9. It shows how much attention our network pay to the breast tumor regions, and we can clearly see that the breast tumor regions are always in the network attention area. During the experiments on MRI and Mammogram based images remain challenging for the BTS-ST network due limited availability of annotated dataset. A sufficient dataset can also benefit our SIB block to improve its performance. Similarly, 512×512 can also be used to reduce information loss during average-pooling operation. The higher-resolution images can also benefit from cropping into multiple images and enlarging the training dataset size.

6. Conclusion

In this paper, we have proposed a BTS-ST vision transformer based network for segmenting and classifying breast tumor images. The primary structure of BTS-ST is based on a dual encoder using U-Net and Swin-Transformer, Spatial Interaction Block (SIB), Feature Compression Block (FCB), and Relationship Aggregation Block (RAB), respectively. We have performed multiple experiments on Ultrasound, MRI, and Mammogram images which prove the model effectiveness on multimodalities images. The proposed network achieved F1(0.908), F1(0.833), and F1(0.771) on the Private BUS dataset, RIDER Breast dataset, and CBIS-DDSM Mammogram dataset. We believe that our proposed network can be used as a reliable, computer-aided breast tumor screening system for the early prediction of breast tumors in women. Normally, vision Transformer's computing time is spent on the attention mechanism. In the future, we can probably address these challenges by introducing lightweight architecture, such as knowledge distillation, tensor decomposition, and deep separable convolution, to achieve model compression with minimal loss of accuracy.

CRediT authorship contribution statement

Ahmed Iqbal: Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Muhammad Sharif:** Supervision, proofreading, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author gratefully acknowledges UltrasoundCases.info, standardized Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset [44], and National Biomedical Imaging center [45] for providing access to Ultrasounds Mammograms, and MRI datasets.

References

[1] W. Chen, K. Sun, R. Zheng, H. Zeng, S. Zhang, C. Xia, Z. Yang, H. Li, X. Zou, J. He, Cancer incidence and mortality in China, 2014, Chin. J. Cancer Res. 30 (2018) 1–12, http://dx.doi.org/10.21147/j.issn.1000-9604.2018.01.01.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (2018) 394–424, http://dx.doi.org/10.3322/caac.21492.

[3] American Cancer Society, American Cancer Society: Cancer Facts and Figures 2022, American Cancer Society, Atlanta, Ga, 2022.

[4] A.Z. Dag, Z. Akcam, E. Kibis, S. Simsek, D. Delen, A probabilistic data analytics methodology based on Bayesian belief network for predicting and understanding breast cancer survival, Knowl. Based Syst. 242 (2022) 108407, http://dx.doi.org/10.1016/j.knosys.2022.108407.

[5] J.R. Eisenbrey, J.K. Dave, F. Forsberg, Recent technological advancements in breast ultrasound, Ultrasonics 70 (2016) 183–190, http://dx.doi.org/10.1016/j.ultras.2016.04.021.

[6] A. Iqbal, M. Sharif, M. Yasmin, M. Raza, S. Aftab, Generative adversarial networks and its applications in the biomedical image segmentation: A comprehensive survey, Int. J. Multimed. Inf. Retr. 11 (2022) 333–368, http://dx.doi.org/10.1007/s13735-022-00240-x.

[7] E.D. Pisano, AI shows promise for breast cancer screening, Nature 577 (2020) 35–36, http://dx.doi.org/10.1038/d41586-019-03822-8.

[8] I. Scholl, T. Aach, T.M. Deserno, T. Kuhlen, Challenges of medical image processing, Comput. Sci.-Res. Dev. 26 (2011) 5–13, http://dx.doi.org/10.1007/s00450-010-0146-9.

[9] X. Meng, J. Fan, H. Yu, J. Mu, Z. Li, A. Yang, B. Liu, K. Lv, D. Ai, Y. Lin, H. Song, T. Fu, D. Xiao, G. Ma, J. Yang, Y. Gu, Volume-awareness and outlier-suppression co-training for weakly-supervised MRI breast mass segmentation with partial annotations, Knowl. Based Syst. 258 (2022) 109988, http://dx.doi.org/10.1016/j.knosys.2022.109988.

[10] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 640–651, http://dx.doi.org/10.1109/TPAMI.2016.2572683.

[11] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9351, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017.

[13] A. Iqbal, M. Usman, Z. Ahmed, An efficient deep learning-based framework for tuberculosis detection using chest X-ray images, Tuberculosis 136 (2022) 102234, http://dx.doi.org/10.1016/j.tube.2022.102234.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, 2020.

[16] W.-Y. Hsu, Improved watershed transform for tumor segmentation: Application to mammogram image compression, Expert. Syst. Appl. 39 (2012) 3950–3955, http://dx.doi.org/10.1016/j.eswa.2011.08.148.

[17] Q.-H. Huang, S.-Y. Lee, L.-Z. Liu, M.-H. Lu, L.-W. Jin, A.-H. Li, A robust graph-based segmentation method for breast tumors in ultrasound images, Ultrasonics 52 (2012) 266–275, http://dx.doi.org/10.1016/j.ultras.2011.08.011.

[18] A. Rodtook, S.S. Makhanov, Multi-feature gradient vector flow snakes for adaptive segmentation of the ultrasound images of breast cancer, J. Vis. Commun. Image Represent. 24 (2013) 1414–1430, http://dx.doi.org/10.1016/j.jvcir.2013.09.009.

[19] D.K. Patra, T. Si, S. Mondal, P. Mukherjee, Breast DCE-MRI segmentation for lesion detection by multi-level thresholding using student psychological based optimization, Biomed. Signal Process. Control 69 (2021) 102925, http://dx.doi.org/10.1016/j.bspc.2021.102925.
[20] N. Arya, S. Saha, Multi-modal advanced deep learning architectures for breast cancer survival prediction, Knowl. Based Syst. 221 (2021) 106965, http://dx.doi.org/10.1016/j.knosys.2021.106965.
[21] G. Yu, Z. Chen, J. Wu, Y. Tan, A diagnostic prediction framework on auxiliary medical system for breast cancer in developing countries, Knowl. Based Syst. 232 (2021) 107459, http://dx.doi.org/10.1016/j.knosys.2021.107459.
[22] Y. Luo, Q. Huang, X. Li, Segmentation information with attention integration for classification of breast tumor in ultrasound image, Pattern Recognit. 124 (2022) 108427, http://dx.doi.org/10.1016/j.patcog.2021.108427.
[23] A. Iqbal, M. Sharif, M.A. Khan, W. Nisar, M. Alhaisoni, FF-UNet: A U-shaped deep convolutional neural network for multimodal biomedical image segmentation, Cognit. Comput. (2022) http://dx.doi.org/10.1007/s12559-022-10038-y.
[24] 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.
[25] C. Peng, Y. Zhang, J. Zheng, B. Li, J. Shen, M. Li, L. Liu, B. Qiu, D.Z. Chen, IMIIN: An inter-modality information interaction network for 3D multi-modal breast tumor segmentation, Comput. Med. Imaging Graph. 95 (2022) 102021, http://dx.doi.org/10.1016/j.compmedimag.2021.102021.
[26] D. Zhai, B. Hu, X. Gong, H. Zou, J. Luo, ASS-GAN: Asymmetric semi-supervised GAN for breast ultrasound image segmentation, Neurocomputing 493 (2022) 204–216, http://dx.doi.org/10.1016/j.neucom.2022.04.021.
[27] Z. Cheng, Y. Li, H. Chen, Z. Zhang, P. Pan, L. Cheng, DSGMFFN: Deepest semantically guided multi-scale feature fusion network for automated lesion segmentation in ABUS images, Comput. Methods Programs Biomed. 221 (2022) 106891, http://dx.doi.org/10.1016/j.cmpb.2022.106891.
[28] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, D. Shen, Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images, Med. Image Anal. 70 (2021) 101918, http://dx.doi.org/10.1016/j.media.2020.101918.
[29] P. Pan, H. Chen, Y. Li, N. Cai, L. Cheng, S. Wang, Tumor segmentation in automated whole breast ultrasound using bidirectional LSTM neural network and attention mechanism, Ultrasonics 110 (2021) 106271, http://dx.doi.org/10.1016/j.ultras.2020.106271.
[30] K. Wang, S. Liang, S. Zhong, Q. Feng, Z. Ning, Y. Zhang, Breast ultrasound image segmentation: A coarse-to-fine fusion convolutional neural network, Med. Phys. 48 (2021) 4262–4278, http://dx.doi.org/10.1002/mp.15006.
[31] Z. Yang, Y. Wei, Y. Yang, Collaborative video object segmentation by multi-scale foreground-background integration, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1, http://dx.doi.org/10.1109/TPAMI.2021.3081597.
[32] X. Chen, K. Zhang, N. Abdoli, P.W. Gilley, X. Wang, H. Liu, B. Zheng, Y. Qiu, Transformers improve breast cancer diagnosis from unregistered multi-view mammograms, Diagnostics 12 (2022) 1549, http://dx.doi.org/10.3390/diagnostics12071549.
[33] C. Qin, Y. Wu, J. Zeng, L. Tian, Y. Zhai, F. Li, X. Zhang, Joint transformer and multi-scale CNN for DCE-MRI breast cancer segmentation, Soft. Comput. (2022) http://dx.doi.org/10.1007/s00500-022-07235-0.
[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: ICCV, 2021.
[35] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: Unet-like pure transformer for medical image segmentation, 2021, CoRR.
[36] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, DS-TransUNet:Dual swin transformer U-Net for medical image segmentation, 2021, CoRR.
[37] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, http://arxiv.org/abs/2102.04306.
[38] Y. Zhang, H. Liu, Q. Hu, TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, 2021, pp. 14–24, http://dx.doi.org/10.1007/978-3-030-87193-2_2.
[39] A. Stergiou, R. Poppe, G. Kalliatakis, Refining activation downsampling with SoftPool, 2021,
[40] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745.
[41] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2019, pp. 3141–3149, http://dx.doi.org/10.1109/CVPR.2019.00326.
[42] Y. Huang, D. Kang, W. Jia, X. He, L. Liu, Channelized axial attention for semantic segmentation – Considering channel relation within spatial attention for semantic segmentation, 2021, CoRR.
[43] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets V2: More deformable, better results, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2019, pp. 9300–9308, http://dx.doi.org/10.1109/CVPR.2019.00953.
[44] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (2017) 170177, http://dx.doi.org/10.1038/sdata.2017.177.
[45] T. cancer imaging Archive, Data from RIDER-breast-MRI, 2015, https://wiki.cancerimagingarchive.net/display/Public/RIDER+Breast+MRI.
[46] 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 (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11045 LNCS, 2018, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-00889-5_1.
[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, Pertanika J. Trop. Agric. Sci. 34 (2018) 137–143.
[48] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, M. Andre, Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network, Biomed. Signal Process. Control 61 (2020) 102027, http://dx.doi.org/10.1016/j.bspc.2020.102027.
[49] R. Su, D. Zhang, J. Liu, C. Cheng, MSU-Net: Multi-scale U-Net for 2D medical image segmentation, Front. Genet. 12 (2021) 1–14, http://dx.doi.org/10.3389/fgene.2021.639930.
[50] R. Li, C. Duan, S. Zheng, C. Zhang, P.M. Atkinson, MACU-Net for semantic segmentation of fine-resolution remotely sensed images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5, http://dx.doi.org/10.1109/LGRS.2021.3052886.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-Decem, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[52] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2017, pp. 2261–2269, http://dx.doi.org/10.1109/CVPR.2017.243.
[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, 2015.
[54] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q.V. Le, H. Adam, Searching for MobileNetV3, 2019.
[55] M. Tan, Q.v. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, 2019, pp. 6105–6114, arXiv:1905.11946.
[56] M.B. Muhammad, M. Yeasin, Eigen-CAM: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–7, http://dx.doi.org/10.1109/IJCNN48605.2020.9206626.