



## OPEN A hybrid attention network for accurate breast tumor segmentation in ultrasound images

Muhammad Azeem Aslam<sup>1✉</sup>, Asim Naveed<sup>2</sup>, Nisar Ahmed<sup>3</sup> & Zhang Ke<sup>4</sup>

Breast ultrasound (BUS) imaging is widely recognized as a non-invasive and cost-effective modality for the timely diagnosis of breast cancer. Despite its clinical importance, automatic tumor segmentation remains a highly challenging task because of speckle noise, varying lesion scale, and inherently indistinct boundaries between malignant and healthy tissue. To address these challenges, we introduce a novel hybrid attention-based segmentation framework, named HA-Net, tailored for BUS images. The proposed HA-Net uses a pre-trained DenseNet-121 backbone in the encoder to extract discriminative features, ensuring robustness against imaging artifacts. At the bottleneck, three complementary modules, Global Spatial Attention (GSA), Position Encoding (PE), and Scaled Dot-Product Attention (SDPA), are incorporated to capture long-range dependencies, preserve structural relationships, and model contextual interactions among features. Moreover, a Spatial Feature Enhancement Block (SFEB) is incorporated within the skip connections to refine spatial detail and emphasize tumor-relevant regions, thereby strengthening the decoder's reconstruction capability. To further improve segmentation reliability, a composite loss function is employed by combining Binary Cross-Entropy (BCE) with Jaccard Index loss, ensuring balanced optimization across pixel-level classification and region-level overlap. In comparison to current state-of-the-art (SOTA) approaches, extensive experiments on publicly available BUS datasets show that the proposed HA-Net achieves competent performance, highlighting its potential as an efficient decision-support tool for radiologists.

Breast cancer is a significant health concern for women worldwide, as it is one of the most commonly diagnosed malignancies and the leading cause of deaths<sup>1</sup>. Timely detection of breast cancer is crucial for improving patient prognosis, and medical imaging serves as a key tool in screening, diagnosis, and treatment planning<sup>2,3</sup>. Breast ultrasound imaging serves as a commonly adopted supplementary technique to mammography, owing to its non-invasive qualities, low expense, real-time diagnostic capability, and effectiveness in identifying tumors in dense breast tissue<sup>4,5</sup>. Unlike mammograms, which often struggle with overlapping tissue structures, ultrasound offers superior contrast for distinguishing solid masses from cystic structures and is especially beneficial for younger women with denser breast composition<sup>5,6</sup>.

Despite its advantages, automated breast lesion segmentation presents substantial challenges. Ultrasound images are inherently characterized by low signal-to-noise ratio, low contrast boundaries, speckle noise, and operator-dependent variability, which collectively hinder the reliable delineation of tumor margins<sup>7</sup>. Furthermore, intra-class variability and inter-class similarity between malignant and benign masses exacerbate the difficulty of precise segmentation, particularly in small datasets commonly encountered in medical imaging<sup>7</sup>. These challenges are less pronounced in mammographic images, where tissue structures are generally more consistent and the signal quality is higher<sup>8–10</sup>.

Medical image segmentation has recently seen considerable progress with deep learning approaches, especially convolutional neural networks (CNNs)<sup>11–14</sup>. Encoder-decoder frameworks, such as U-Net<sup>15</sup> and its variants, have become widely used for biomedical applications due to their efficacy in capturing both fine-grained texture data and high-level semantic information<sup>16</sup>. Despite their popularity, conventional CNN-based models often struggle to model contextual relationships and long-range spatial dependencies that are necessary for precise segmentation in complex modalities such as breast ultrasound<sup>17</sup>. Furthermore, when dealing with substantial speckle noise and indistinct tumor boundaries, relying solely on basic skip connections, as adopted

<sup>1</sup>School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi 710071, China. <sup>2</sup>Department of Computer Science, University of Engineering and Technology Lahore, Faisalabad Campus, Faisalabad 37630, Pakistan. <sup>3</sup>Department of Informatics and Systems, University of Management and Technology, Lahore, Punjab 54000, Pakistan. <sup>4</sup>School of Astronautics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China. ✉email: azeem@eurasia.edu

in many encoder–decoder architectures, may be insufficient for transmitting detailed spatial information from the encoder to the decoder, which can ultimately hinder segmentation accuracy<sup>17</sup>.

To overcome these limitations, we introduce a novel hybrid attention-based network for the segmentation of breast ultrasound lesions. It combines a decoder with a pre-trained DenseNet-121 encoder<sup>18</sup> for reliable feature extraction with multiple attention mechanisms. At the bottleneck, the model integrates Global Spatial Attention (GSA)<sup>19</sup>, Position Encoding (PE)<sup>20</sup>, and Scaled Dot-Product Attention (SDPA)<sup>21</sup> to capture global context, spatial dependencies, and relative positional information. Furthermore, the Spatial Feature Enhancement Block (SFEB) is incorporated at the skip connections to refine spatial representations, allowing the model to concentrate more effectively on relevant regions. This architecture improves the precise localization of lesions and sharpens boundary definition, both of which are essential for reliable clinical use<sup>18,19</sup>. To optimize, a composite loss function is employed that combines jaccard index loss and binary cross-entropy (BCE)<sup>18</sup>, balancing pixel-level classification with region-level overlap. This strategy improves robustness against class imbalance and accommodates irregular tumor shapes, resulting in more accurate and reliable segmentation outcomes<sup>18,19,22</sup>.

The key contributions of the proposed HA-Net are outlined below:

- A hybrid attention network is proposed using an encoder composed of DenseNet-121, pre-trained on ImageNet, specifically designed for precise segmentation of breast ultrasound images (BUSI).
- A transformer-based attention mechanism is introduced to incorporate spatial, positional, and semantic cues, improving segmentation precision.
- Spatial Feature Enhancement Block (SFEB) is incorporated in skip connections to refine feature propagation and enhance focus on tumor regions.
- A combined loss that integrates BCE and Jaccard Index loss is employed to optimize both pixel-level classification and region-level overlap, effectively tackling speckle noise and ambiguous tumor margins.
- Extensive tests on publicly accessible breast ultrasound datasets show that our proposed HA-Net performs well when evaluated against competent approaches, indicating its potential to help radiologists diagnose breast cancer early and accurately.

The remaining parts of this manuscript are classified as follows. The related work section reviews existing studies with a focus on the limitations of U-Net and conventional CNN-based architectures, the emergence of attention mechanisms, and the recent adoption of Transformer and Mamba-based models in medical imaging. The methodology section presents the proposed framework in detail, including the transformer attention module, transformer self-attention, global self-attention, SFEB, and the hybrid loss functions employed for optimization. The experiments section describes the datasets utilized, preprocessing strategies, implementation details, ablation studies, and evaluation metrics performed to validate the model design. The results section reports both numerical and visual outcomes, supported by ablation results, while the discussion provides statistical analysis and interprets the significance of findings in the context of clinical application. Finally, the conclusion highlights the main contributions, summarizes key insights, acknowledges limitations, and outlines potential ideas for future exploration.

## Related work

The segmentation of breast tumors has become a major focus in recent research due to its importance in early detection and treatment planning<sup>23</sup>. Compared to other modalities like mammography, ultrasound imaging offers a number of benefits, such as real-time imaging, reduced ionizing radiation, cost-effectiveness, and improved visibility in dense breast tissue. However, ultrasound images pose unique challenges for automated analysis because of speckle noise, low contrast, operator variability, and anatomical ambiguities<sup>23,24</sup>. These difficulties have driven the development of advanced deep learning methods capable of extracting robust features and leveraging contextual information to improve segmentation accuracy<sup>23,25</sup>.

## Limitations of U-Net and CNNs

Initial attempts at breast tumor segmentation relied on classical computer vision techniques such as filtering, active contours, and clustering methods<sup>26</sup>. For example, threshold-based segmentation and graph-based approaches were used in early studies to delineate lesions in ultrasound images<sup>27</sup>. However, these methods required extensive domain knowledge and struggled with noise sensitivity and over-segmentation<sup>23,24</sup>.

The field was revolutionized by CNNs, particularly the U-Net design, which enabled the learning of hierarchical features in an end-to-end manner<sup>23,25</sup>. Recent studies demonstrate that densely connected U-Net variants with attention mechanisms achieve dice scores exceeding 0.83, outperforming traditional methods<sup>23</sup>. For instance, ACL-DUNet<sup>23</sup> integrates channel attention modules and spatial attention gates to suppress irrelevant regions while enhancing tumor features. Similarly, SK-U-Net<sup>24</sup> employs selective kernels with dilated convolutions to adapt receptive fields, gaining a mean dice score of 0.826 in comparison to 0.778 for the standard U-Net.

To address limited contextual awareness, multi-branch architectures have emerged. One approach<sup>25</sup> combines classification and segmentation branches, achieving an AUC of 0.991 for normal/abnormal classification and a dice score of 0.898 for segmentation. These models reduce false positives in normal images while maintaining sensitivity advancement for clinical screening<sup>25</sup>. Hybrid designs like DeepCardinal-50<sup>28</sup> further optimize computational efficiency, achieving 97% accuracy in tumor detection with real-time processing capabilities.

However, challenges persist in modeling long-range dependencies for lesions with irregular morphology. While attention mechanisms in ACL-DUNet improve spatial focus<sup>23</sup>, and scale attention modules enhance multi-level feature integration<sup>23</sup>, fuzzy boundaries in low-contrast ultrasound images remain difficult<sup>24</sup>. These

constraints are being addressed by ongoing advancements in adaptive kernel selection and boundary-guided networks<sup>23,24</sup>.

### Rise of attention mechanisms

Recent advancements in breast tumor segmentation in ultrasound imaging have been driven by the incorporation of attention mechanisms and hybrid network architectures. Early strategies focused on spatial-channel attention to address challenges such as fuzzy lesion boundaries and variable tumor sizes. For example, SC-FCN-BLSTM<sup>29</sup> combined bi-directional LSTM with spatial-channel attention to exploit inter-slice contextual information in 3D automated breast ultrasound. Abraham et al.<sup>30</sup> presented hybrid attention mechanisms that adaptively reweigh feature maps based on contextual saliency, improving segmentation performance in noisy ultrasound images. Similarly, adaptive attention modules such as HAAM<sup>31</sup> replaced standard convolutions in U-Net variants, allowing dynamic adjustment of the receptive field across spatial and channel dimensions for more robust segmentation.

Further improvements were achieved with CBAM-RIUnet<sup>4</sup>, which combined convolutional block attention modules with residual inception blocks, yielding intersection-over-union (IoU) and dice scores of 88.71% and 89.38%, respectively, by effectively suppressing irrelevant features. The authors<sup>32</sup> presented ESKNet, which integrates particular kernel networks into the U-Net to dynamically modulate receptive fields using attention, enhancing segmentation accuracy across diverse lesion types. Although attention-based models have improved segmentation accuracy, many approaches are still limited in adequately representing long-range spatial relationships, specifically when relying on a single attention strategy. This has led to the exploration of hybrid models that combine multiple attention mechanisms to provide a richer representation of both local and global features.

ARF-Net<sup>33</sup> was introduced for breast mass segmentation in both mammographic and ultrasound images, leveraging an encoder-decoder backbone integrated with a Selective Receptive Field Module (SRFM) to adaptively regulate receptive field sizes based on lesion scale, thereby balancing global context and local detail for improved accuracy. In<sup>34</sup>, the authors presented a lightweight CNN-based model for mammogram segmentation, incorporating feature strengthening modules for enhanced representation, a parallel dilated convolution block for multi-scale context and boundary refinement, and a mutual information loss to maximize consistency with ground truth. These innovations collectively enable accurate and efficient segmentation with low computational cost. ATFE-Net<sup>35</sup> employed an Axial-Trans module to efficiently capture long-range dependencies and a Trans-FE module to enhance multi-level feature representations.

### Transformer and Mamba-based architectures in medical imaging

Inspired by the breakthroughs of Transformer architectures in natural language processing, Vision Transformers (ViTs) and their variants have gained significant traction in medical image analysis, demonstrating strong capability in modeling global context and capturing long-range dependencies<sup>36</sup>. Transformers overcome CNNs' local constraints by enabling global context modeling through self-attention processes. Several studies have successfully incorporated transformers into segmentation pipelines, either as standalone modules or in combination with CNN backbones<sup>37,38</sup>.

To integrate local convolutional features with long-range contextual information, a hybrid CNN-transformer architecture was presented by He et al.<sup>39</sup> and Ma et al.<sup>35</sup>. While these models demonstrate strong performance, these architectures often face challenges in retaining fine-grained boundary details, which are essential for precise segmentation of medical images. Swin Transformer-based networks address this limitation by employing hierarchical attention and shifted windows to capture features at multiple scales. For instance, DS-TransUNet<sup>40</sup> leverages these mechanisms to simultaneously extract coarse and fine features, enhancing segmentation precision. Similarly, Swin-Net<sup>41</sup> combines a Swin Transformer encoder with feature refinement and hierarchical multi-scale feature fusion modules to achieve more accurate lesion delineation. SwinHR<sup>42</sup> further enhances performance by adopting hierarchical re-parameterization with large kernel convolutions, capturing long-range dependencies efficiently while maintaining high accuracy through shifted window-based self-attention. Cao et al.<sup>43</sup> took this further by developing a pixel-wise neighbor representation learning approach (NeighborNet), allowing each pixel to adaptively select its context based on local complexity. This approach is particularly suitable for ultrasound segmentation, where lesion boundaries may be fragmented or ambiguous.

In breast cancer segmentation, a critical research gap lies in the integration of transformer-based models with CNNs, where semantic mismatches between locally extracted CNN features and globally contextualized transformer representations often lead to suboptimal fusion<sup>44,45</sup>. Inflexible or disjointed fusion strategies, such as rigidly inserting transformer blocks into CNN architectures without addressing feature consistency, result in redundant or insufficient hierarchical representations<sup>45</sup>. This challenge is exacerbated in noisy or irregular data, such as breast ultrasound images, where speckle artifacts, shadowing, and blurred lesion boundaries create discordance between local texture details and global anatomical structures<sup>45,46</sup>. Current approaches frequently fail to link the semantic gap between CNNs' localized feature extraction and transformers' long-range dependency modeling, particularly in decoder stages where misaligned feature maps reduce segmentation precision for small lesions and complex margins<sup>39</sup>. Furthermore, the lack of adaptive cross-attention mechanisms to harmonize multi-scale features often diminishes model robustness against ultrasound-specific noise patterns<sup>45</sup>, highlighting the need for more sophisticated hybrid architectures that enable synergistic local-global feature interaction while maintaining computational efficiency<sup>39</sup>.

Accurate medical image segmentation is essential for clinical decision-making, but existing CNN-Transformer hybrid models often depend heavily on skip connections, which limit the extraction of contextual features. To address this, MRCTransUNet combines a lightweight MR-ViT with a reciprocal attention module to close the semantic gap and retain fine details. The MR-ViT and RPA modules enhance long-range contextual

learning in deeper layers, but skip connections are only utilized in the first layer, in contrast to conventional U-Net variations. Tests on breast, brain, and lung datasets show that MRCTransUNet exceeds the performance of current leading methods on Dice and Hausdorff metrics, demonstrating its potential for reliable clinical use applications<sup>47</sup>.

The authors proposed HCMNet<sup>48</sup>, a hybrid CNN–Mamba network that integrates CNN’s strength in local feature extraction with Mamba’s capability for efficient global representation. A wavelet feature extraction module enriches feature learning by combining low- and high-frequency components, reducing spatial information loss during downsampling. Furthermore, an adaptive feature fusion module enhances skip connections by dynamically merging encoder and wavelet features, thereby preserving critical details and suppressing redundancy. The authors introduced AttnNet<sup>49</sup>, a novel multiscale attention-mamba (MAM) module in a U-shaped model. Using a Mamba unit that combines self-attention and Mamba processes, the MAM block combines multi-level convolutional layers to extract features across various spatial scales. With this design, the model can retain fine structural features.

## Methodology

The proposed HA-Net consists of four key components: an encoder, a decoder, a transformer-based attention module, and a spatial feature enhancement block. For the encoder backbone, DenseNet-121<sup>50</sup> is used to effectively capture both complex and fine-grained representations. DenseNet’s used feature–direct connections between all layers within a dense block (DB) encourages feature reuse, improves gradient flow, and supports efficient information propagation. These characteristics are especially valuable in medical image segmentation, where subtle anatomical variations and boundary precision are critical for reliable lesion delineation. In the encoding path, four hierarchical encoding stages are constructed following the standard DenseNet-121 design. Each stage comprises multiple dense blocks interleaved with transition layers (TLs), as illustrated in Fig. 1. This hierarchical organization enables the model to progressively learn low-level texture features alongside high-level semantic information while maintaining spatial continuity. The dense connectivity within DBs strengthens feature propagation, while TLs serve to reduce dimensionality and regulate complexity without discarding critical details. Together, these mechanisms ensure that the encoder produces a rich, multi-scale, and highly discriminative representation suitable for subsequent decoding and attention operations. To further refine extracted features, we append a convolutional block—consisting of a  $3 \times 3$  convolution, a ReLU activation, and batch normalization (BN) after the pre-trained DenseNet-121.

The decoding path follows a simplified U-Net<sup>15</sup> inspired design, optimized to maintain strong representational power while reducing the number of parameters for improved computational efficiency. Rather than relying on transposed convolutions, which are prone to introducing checkerboard artifacts and can substantially increase computational complexity, our approach utilizes bilinear upsampling followed by convolutional layers. This combination preserves spatial resolution and fine-grained feature details while minimizing parameter count and inference time. By preserving detailed feature reconstruction and precise boundary delineation, the proposed decoding pathway delivers accurate segmentation while keeping computational demands low.

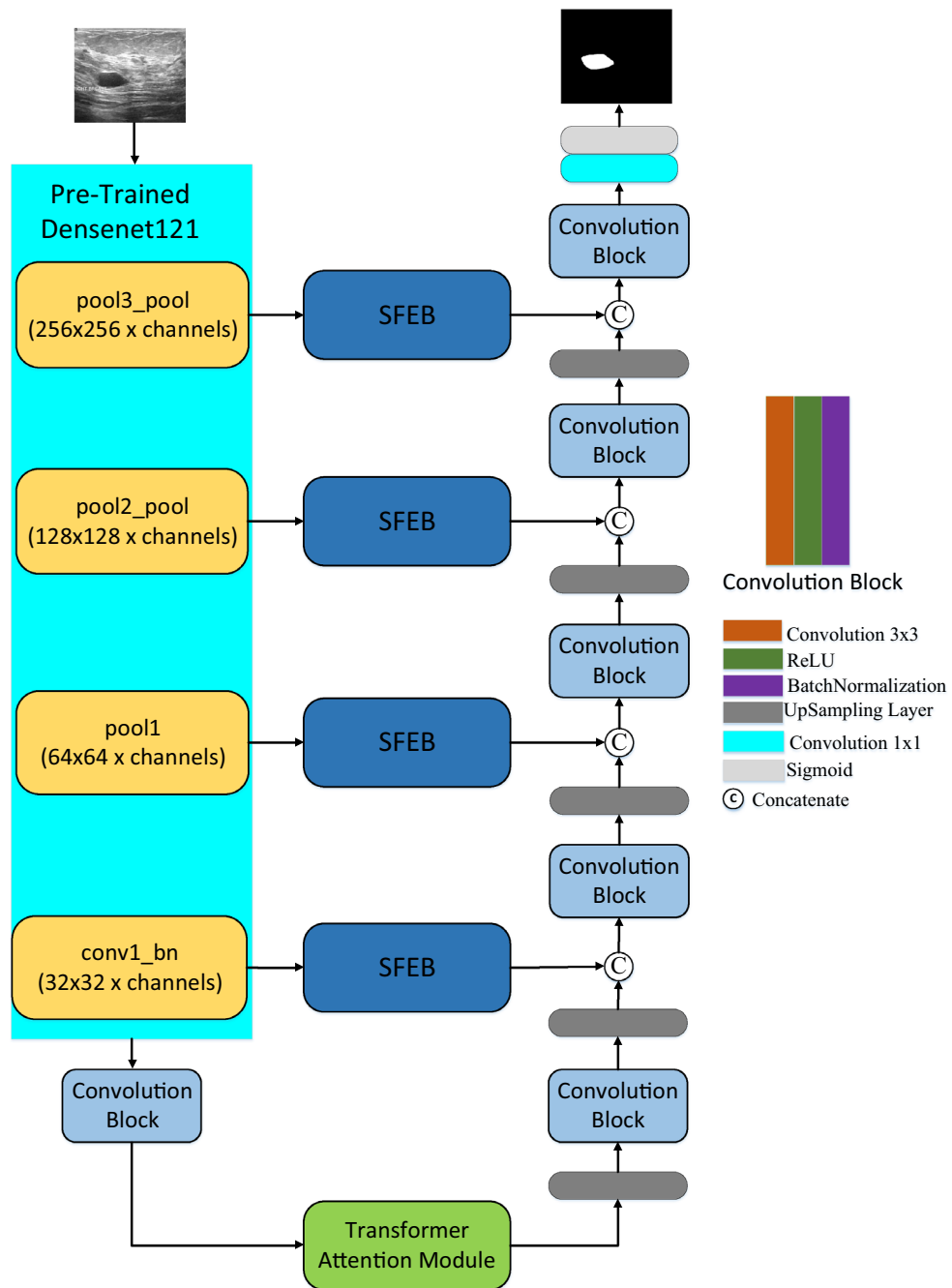
The proposed HA-Net employs five sequential convolutional blocks in the decoder path as shown in Fig. 1, to progressively extract hierarchical features. Each convolutional block is composed of a  $3 \times 3$  convolution layer, BN, and a ReLU activation function to stabilize training and enhance feature representation. This design ensures stable and efficient training while enabling the network to simultaneously capture high-level semantic representations and fine-grained textural details. BN reduces internal covariate shift, speeds up convergence, and enhances generalization, while ReLU activation adds non-linearity to efficiently represent intricate patterns in breast ultrasound images.

Additionally, skip connections from the encoder are employed to preserve spatial information and facilitate multi-scale feature fusion across different resolution levels. The integration of SFEB within the decoding path further refines the feature representations by selectively emphasizing tumor-relevant regions, thereby improving segmentation accuracy while maintaining a reduced parameter count relative to a conventional U-Net. This optimized architecture not only enables efficient processing of high-resolution medical images but also ensures precise delineation of fine structural details, making the model highly suitable for practical clinical deployment and real-time applications.

## Transformer Attention Module (TAM)

To strengthen the method’s capacity to capture and fuse contextual information, we incorporate a self-aware attention module<sup>51</sup>. There are two main components to this module. Initially, contextual information is captured by the Transformer Self-Attention (TSA) block by taking into account relative positions within the input data. It integrates positional information by concatenating input features with positional embeddings paths to allow the model to understand spatial relationships within the input data. Secondly, the Global Spatial Attention (GSA) block refines local contextual information by aggregating it with global features. By incorporating a broader perspective, this design enhances the model’s ability to retain fine structural details while simultaneously maintaining a holistic understanding of the lesion’s overall morphology. Collectively, these attention mechanisms improve feature representation, helping the model effectively balance local and global information for more precise segmentation.

Figure 2 depicts the Transformer Attention Module (TAM) architecture. The input feature map  $F_{in}$  is first enriched with positional encoding and passed to two parallel branches. In the top branch (TSA), the encoded features are projected into  $Q$ ,  $K$ , and  $V$  for calculation of scaled dot-product attention, capturing long-range contextual dependencies. In the bottom branch (GSA), the features are embedded into two complementary representations whose dot product produces a spatial attention map, highlighting global positional relationships.



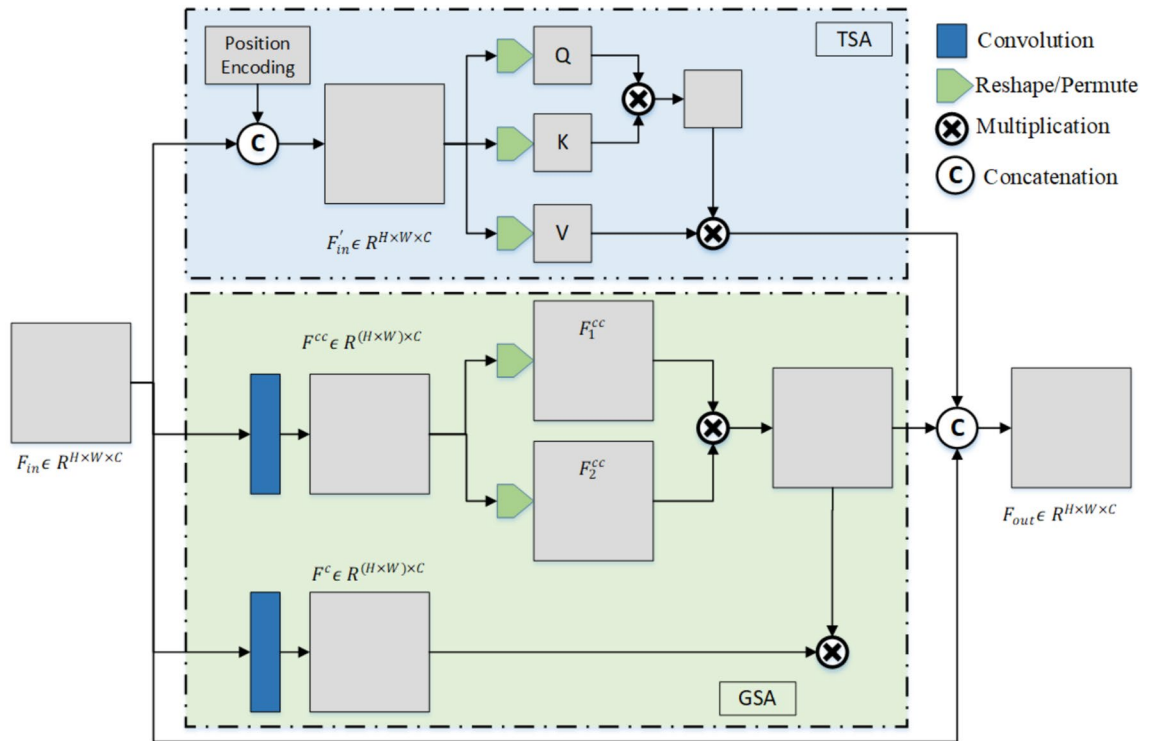
**Fig. 1.** The specifics of the HA-Net. The proposed HA-Net includes a transformer attention module, a spatial features augmentation block, a pre-trained encoder, and a particular decoder.

Finally, the outputs of TSA, GSA, and the PE-enriched input are concatenated to generate  $F_{out}$ , which jointly preserves local details, global context, and spatial correlations.

#### Transformer Self-Attention (TSA)

Since multi-head attention effectively captures self-correlation but cannot learn spatial relationships, a common strategy is to introduce positional encoding before applying attention. Specifically, the incoming feature representation  $F \in \mathbb{R}^{h \times w \times c}$  is first enriched with positional information, producing a representation, which is then fed into the multi-head attention block (Fig. 2).  $F$  is first reshaped into a two-dimensional representation  $F' \in \mathbb{R}^{c \times (h \times w)}$ . Using learnable weight matrices,  $F'$  is then projected into three distinct spaces: queries  $Q \in \mathbb{R}^{c \times (h \times w)}$ , keys  $K \in \mathbb{R}^{c \times (h \times w)}$ , and values  $V \in \mathbb{R}^{c \times (h \times w)}$ , defined as

$$Q = W_q F', \quad K = W_k F', \quad V = W_v F', \quad (1)$$



**Fig. 2.** The components of the transformer attention module. The top block illustrates the transformer self-attention, while the bottom block displays the global self-attention block.

where  $W_q, W_k, W_v \in \mathbb{R}^{c \times c}$  are learnable projection matrices.

The scaled dot-product attention mechanism computes the similarity between different channels by applying the Softmax-normalized dot-product of  $Q$  and the transposed version of  $K$ . This matrix represents the contextual attention map  $A \in \mathbb{R}^{c \times c}$ . Finally, the contextual attention map  $A$  is applied to the value matrix  $V$  to produce attention-weighted feature representations. This mechanism allows the multi-head attention module to selectively aggregate relevant features while preserving essential contextual dependencies across spatial positions. Mathematically, the Transformer Self-Attention (TSA) operation can be expressed as:

$$A_{TSA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where  $d_k$  denotes the dimensionality of the key vectors, ensuring proper scaling of the dot-product attention. This formulation allows the TSA block to model long-range dependencies and refine feature aggregation while maintaining spatial coherence in the output representations.

#### Global Spatial Attention (GSA)

To further enhance contextual learning, the TAM incorporates the Global Spatial Attention (GSA) block, which captures correlations among different spatial positions across the feature map. The initial feature representation  $F \in \mathbb{R}^{h \times w \times c}$  is embedded in  $F^c \in \mathbb{R}^{h \times w \times c}$  and  $F^{cc} \in \mathbb{R}^{h \times w \times c'}$  where  $c' = c/2$ . The latter is reshaped to  $F_1^{cc} \in \mathbb{R}^{h \times w \times c'}$  and  $F_2^{cc} \in \mathbb{R}^{h \times w \times c'}$ . The scaled dot product of these matrices is computed and subsequently passed through a Softmax normalization layer, resulting in an attention map  $GSA \in \mathbb{R}^{(h \times w) \times (h \times w)}$  that encodes the pairwise correlations between different spatial positions. The multi-head attention mechanism is then formulated as:

$$A_{GSA} = \text{softmax}(F_1^{cc} \cdot F_2^{cc}) \quad (3)$$

The outputs from TSA, GSA, and the original input are then concatenated to create the output feature map ( $F_{out} \in \mathbb{R}^{h \times w \times c}$ ) of the self-aware attention module. The model's capacity to extract significant features for precise segmentation is improved by this method, which guarantees that both local spatial relationships and global context are well recorded.

#### Spatial Feature Enhancement Block (SFEB)

Pooling operations play a critical role in deep learning by reducing the spatial dimensions of feature maps, accelerating computation, and enhancing feature robustness. In lesion segmentation, it is crucial to simultaneously capture fine-grained structural details and global contextual cues, since tumors often exhibit low contrast, small

spatial extent, and heterogeneous textural patterns. To address these challenges, we incorporate an SFEB within the skip connections of our network, which strengthens feature fusion, spatial awareness, and residual learning, ultimately improving segmentation accuracy and the preservation of lesion boundaries.

To improve discriminative characteristics and refine spatial features before fusion, the SFEB is integrated into skip connections. Global max-pooling highlights sharp lesion boundaries, while global average-pooling preserves contextual information, and their combination ensures a balance between local detail and global context. The attention pathway further reweights channels to emphasize lesion-relevant features and suppress background noise. Finally, residual fusion preserves fine spatial details, making the SFEB particularly effective for refining skip connection features in noisy ultrasound images with irregular tumor boundaries.

The input tensor is first passed through a  $3 \times 3$  convolutional layer, BN, and a ReLU activation, resulting in an intermediate feature map  $I_1$ .

$$I_1 = \text{ReLU}(\mu(f^{3 \times 3}(I))), \quad (4)$$

where  $I = \mathbb{R}^{H \times W \times C}$  represents the input tensor with height H, width W, and channel depth C. To extract global contextual information, the intermediate feature map  $I_1$  is subjected to both global max-pooling ( $G_m$ ) and global average-pooling ( $G_a$ ), producing the pooled feature maps  $GP_m$  and  $GP_a$ , respectively:

$$GP_m = G_m(I_1), \quad GP_a = G_a(I_1) \quad (5)$$

These pooled features are concatenated to form a complementary representation  $P_o$ :

$$P_o = GP_m \oplus GP_a. \quad (6)$$

The concatenated pooled features  $P_o$  are then refined by applying a  $3 \times 3$  convolution, BN, and ReLU activation:

$$F_c = \text{ReLU}(\mu(f^{3 \times 3}(P_o))). \quad (7)$$

In parallel,  $G_a$  is applied to the original input  $I$ , followed by a  $1 \times 1$  convolution, BN, and sigmoid activation, producing an attention map  $F_{cc}$ :

$$F_G = GAP(I), \quad F_{cc} = \sigma(\mu(f^{1 \times 1}(F_G))), \quad (8)$$

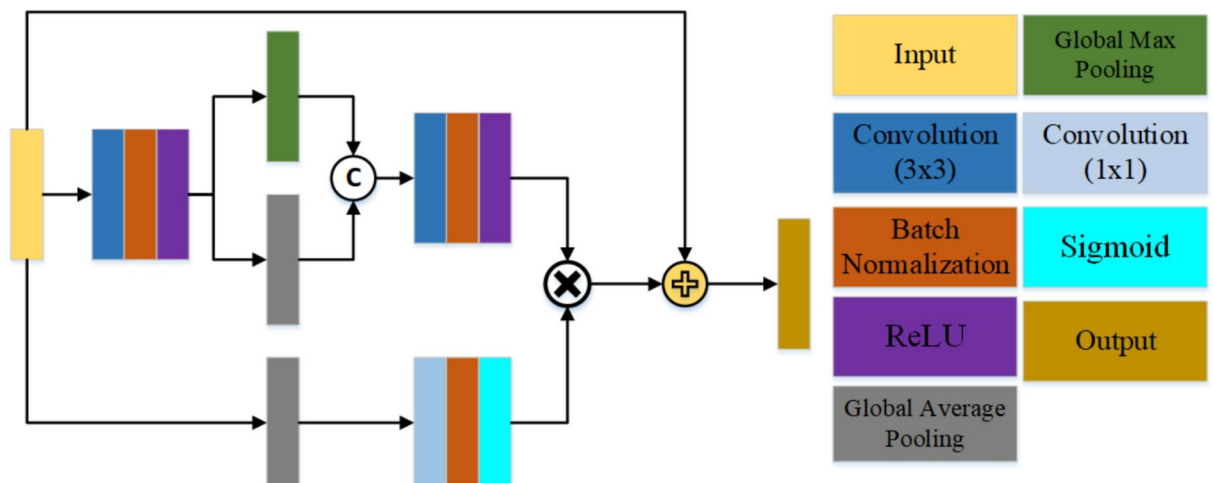
where  $\sigma$  represents the sigmoid activation. The attention-enhanced features  $F_{em}$  are computed by element-wise multiplication of the refined feature map  $F_c$  and the attention coefficients  $F_{cc}$ :

$$F_{em} = F_c \otimes F_{cc}. \quad (9)$$

Finally, to preserve the original spatial information and maintain residual learning, the input tensor  $I$  is added element-wise to the attention-enhanced features:

$$F = F_{em} \oplus I. \quad (10)$$

By maintaining fine structural details, this architecture effectively balances local feature intricacy with global contextual information, enabling the model to focus on relevant regions. The SFEB's architecture is shown in Fig. 3.



**Fig. 3.** Spatial Feature Enhancement Block.

## Loss functions

An appropriate choice of a loss function is crucial to train the model because it directly influences the convergence behavior, stability, and the balance between pixel-wise accuracy and region-level consistency in segmentation tasks<sup>52</sup>. BCE loss quantifies the pixel-wise difference between the predicted probability map and the ground truth mask. By computing the negative log-likelihood of the predicted probabilities, it penalizes incorrect classifications and enforces accurate pixel-level segmentation. Formally, for  $N$  pixels, BCE loss is defined as:

$$\text{Loss}_{\text{bce}} = - \sum_{i=1}^N \left( y_i, \log, \hat{y}_i + (1 - y_i), \log(1 - \hat{y}_i) \right), \quad (11)$$

where  $y_i \in 0, 1$  represents the ground truth label of the  $i$ -th pixel, and  $\hat{y}_i \in [0, 1]$  denotes the predicted probability. This formulation ensures that confident misclassifications are penalized more heavily, guiding the model toward robust pixel-level discrimination.

Jaccard loss, also known as Intersection-over-Union (IoU) loss, is a region-level metric that evaluates the degree of overlap between the predicted segmentation mask and the ground truth, emphasizing accurate delineation of target regions. Because Jaccard loss prioritizes structural similarity over pixel-wise loss, it works especially well with highly unbalanced datasets in which the lesion or region of interest only takes up a small portion of the image. It has the following mathematical definition:

$$\text{Loss}_{\text{jaccard}} = 1 - \frac{\sum_{i=1}^N (P_i G_i)}{\sum_{i=1}^N (P_i + G_i - P_i G_i)}, \quad (12)$$

where  $P_i \in [0, 1]$  is the predicted probability for the  $i$ -th pixel, and  $G_i \in 0, 1$  is the corresponding ground truth label.

To leverage both pixel-level accuracy (captured by BCE loss) and region-level similarity (captured by Jaccard loss), a hybrid objective is formulated. The final training objective is defined as:

$$\text{Loss}_{\text{combined}} = \text{Loss}_{\text{bce}} + \text{Loss}_{\text{jaccard}}, \quad (13)$$

where  $\text{Loss}_{\text{bce}}$  ensures fine-grained classification at each pixel, and  $\text{Loss}_{\text{jaccard}}$  enforces global shape and boundary consistency. This joint formulation stabilizes convergence and improves segmentation performance across varying lesion sizes and shapes.

## Code availability

The source code implementing the proposed Hybrid Attention Network (HA-Net) for breast tumor segmentation in ultrasound images is openly available at GitHub Repository Link: <https://github.com/nisarahmedrana/HA-Net>. A DOI has been generated via Zenodo to ensure long-term accessibility: <https://doi.org/10.5281/zenodo.17190194>. The repository includes processed dataset, Jupyter Notebook describing architecture, preprocessing pipelines, training and evaluation scripts and usage instructions required to reproduce the results presented in this study. The code is released for research purposes only under the specified license.

## Experiments

### Datasets for breast ultrasound image segmentation

To rigorously evaluate the effectiveness of the HA-Net, we conducted extensive experiments on two publicly available breast ultrasound datasets, BUSI and UDIAT. Both datasets consist of grayscale ultrasound images with corresponding pixel-level annotations provided by clinical experts, serving as reliable benchmarks for tumor segmentation. The BUSI dataset contains ultrasonograms of multiple patients with varying lesion types (benign, malignant, and normal), thereby reflecting the heterogeneity of real clinical scenarios. The UDIAT dataset, on the other hand, offers high-quality ultrasound scans with consistent acquisition settings, enabling controlled evaluation. Together, these datasets provide complementary characteristics, ensuring that the proposed method is validated across diverse imaging conditions and lesion appearances.

**BUSI Dataset:** The BUSI dataset<sup>53</sup> comprises 780 grayscale breast ultrasound images obtained from 600 female patients within the age range of 25 to 75 years. Each image has an approximate spatial resolution of  $500 \times 500$  pixels and is annotated into three diagnostic categories: normal, benign, and malignant. For tumor segmentation, only the benign and malignant categories were retained, as these are accompanied by expert-annotated binary masks delineating tumor regions. Images belonging to the normal class were excluded since they lack lesion annotations. To ensure uniformity in model input, all images and their corresponding masks were resized to  $256 \times 256$ . This preprocessing step not only standardizes input dimensions across the dataset but also reduces computational overhead during training and evaluation.

**UDIAT Dataset:** The UDIAT dataset was introduced by<sup>54</sup> and consists of 163 breast ultrasound images. These images are divided into benign and malignant classifications and have a resolution of  $760 \times 570$  pixels. Pixel-wise segmentation masks with expert annotations that identify tumor locations are included with every image. Before training, all images and their corresponding masks were resized to  $256 \times 256$  pixels to ensure consistency. Table 1 provides details of the BUSI and UDIAT datasets' separation into training and test sets.

Dataset	Training	Test	Image Resolution
BUSI	700	80	256 × 256
UDIAT	133	33	256 × 256

**Table 1.** A summary of the datasets employed in this study, including the total number of images, diagnostic categories, and image resolution, is presented.

Metric	Formula	Interpretation
Jaccard Index (IoU)	$\frac{TP}{TP + FP + FN}$	Measures overlap between predicted and ground-truth masks. Higher values indicate stronger agreement.
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Represents the overall proportion of correctly classified pixels.
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Evaluates the ability to correctly identify lesion pixels. High recall reduces the risk of missed detections.
Precision	$\frac{TP}{TP + FP}$	Reflects the proportion of correctly predicted lesion pixels among all pixels labeled as lesion.
Dice Coefficient (DSC)	$\frac{2TP}{2TP + FP + FN}$	Balances recall and precision, providing an intuitive measure of segmentation overlap.
Specificity	$\frac{TN}{TN + FP}$	Assesses the ability to correctly classify background pixels, minimizing false alarms.

**Table 2.** Summary of evaluation metrics used for segmentation performance assessment.

Implementation details

To ensure robust training and reliable performance evaluation, 20% of the training set was withheld for validation, enabling effective monitoring of learning progress and guiding hyperparameter adjustments. Model optimization was performed using the Adam optimizer<sup>55</sup> with an initial learning rate of 0.001. To promote stable convergence and mitigate the risk of stagnation, the learning rate was reduced by a factor of 0.25 when the validation loss plateaued for four consecutive epochs. In addition, early stopping was employed to prevent overfitting and automatically terminate training once no further improvements were observed.

A hybrid loss function combining Binary Cross-Entropy (BCE) and Jaccard loss was utilized, allowing simultaneous optimization at both the pixel level and the region overlap level. Training was conducted with a batch size of 10, and the model achieved competitive performance without the application of explicit data augmentation strategies. The proposed framework was implemented in Keras with TensorFlow as the backend. All experiments were executed on a workstation equipped with an NVIDIA Tesla K80 GPU, an Intel Xeon 2.20 GHz CPU, 13 GB of system RAM, and 12 GB of dedicated GPU memory

Evaluation metrics

The segmentation performance of the proposed HA-Net was quantitatively assessed using a set of widely adopted evaluation metrics in medical image analysis. These metrics capture both pixel-level accuracy and region-level overlap, providing a comprehensive view of model performance. The definitions and interpretations of all metrics are summarized in Table 2.

Ablation studies

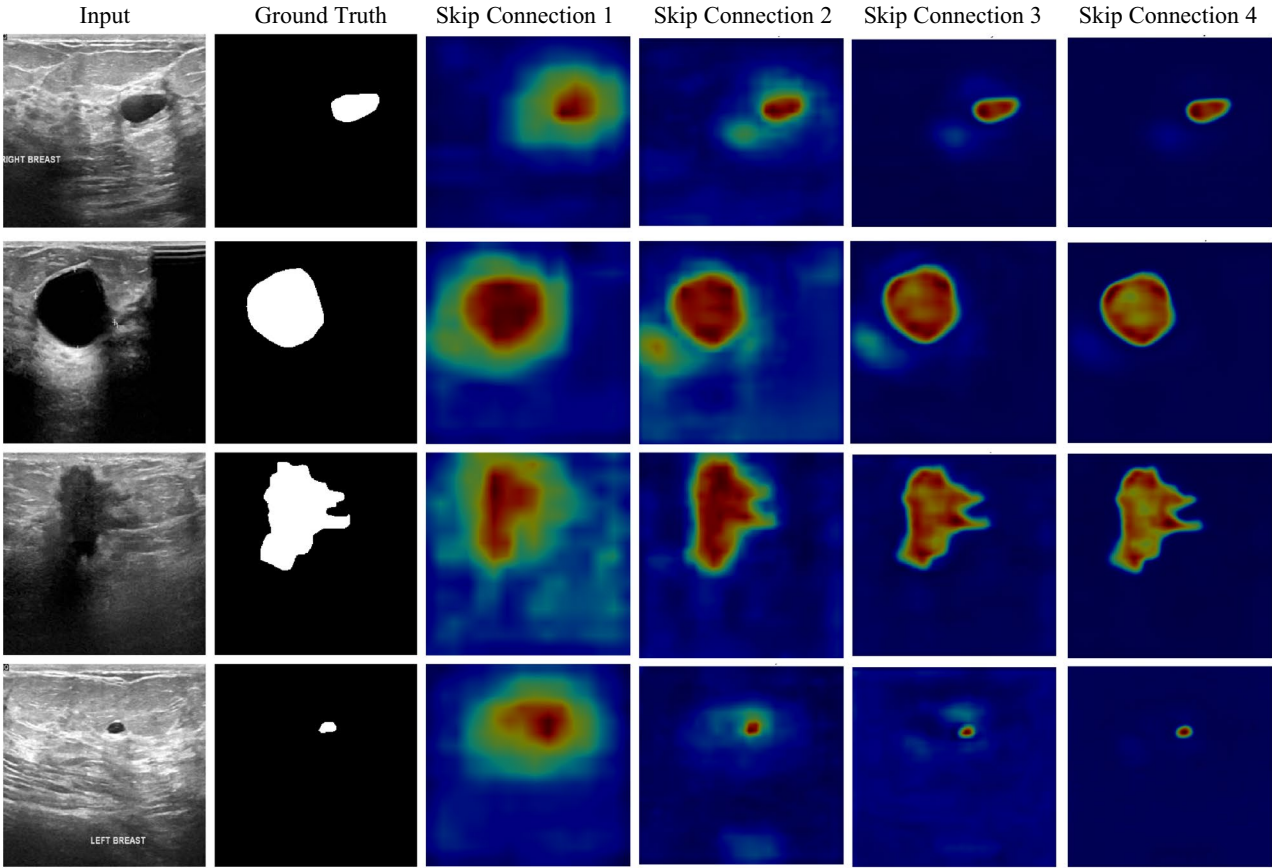
A series of ablation studies were performed on the BUSI dataset to systematically assess the individual contributions of each component within the proposed HA-Net. The pre-trained DenseNet-121 encoder used in the backbone model was chosen for its strong feature extraction and multi-scale representation capabilities. To progressively enhance spatial and contextual understanding, we incrementally integrated three key modules into the baseline: a convolutional block, the SFEB, and TAM.

The results in Table 3 clearly demonstrate the impact of each module, with sequential integration consistently improving performance across all evaluation metrics. In particular, the incorporation of SFEB and TAM yields significant gains in performance, underscoring their effectiveness in refining feature representation and enhancing lesion localization. These findings highlight the critical role of both spatial feature refinement and attention-based contextual modeling in enabling precise delineation of tumor regions, validating the design choices of the proposed architecture.

To further explain the interpretability of the HA-Net, heatmaps of the SEFB are visualized using Grad-CAM<sup>56</sup> on the BUSI dataset. The SEFB module is integrated into skip connections across four hierarchical levels of the network, enabling progressive refinement of feature representations. The visualization demonstrates how SEFB adaptively emphasizes salient lesion regions while suppressing background noise throughout the encoding-decoding process. In the presented results in Fig. 4, the first column corresponds to the original ultrasound image, while the second column provides the ground truth segmentation mask. The subsequent columns depict the SEFB attention responses at the four skip-connection stages. These stage-wise heatmaps highlight

Method	Parameters	Jaccard	Dice	Sensitivity	Accuracy	Precision	Specificity
U-Net <sup>15</sup>	34,513,345	76.54	83.13	82.83	97.91	83.94	98.81
Baseline (BL)	10,498,945	90.44	94.37	94.33	99.46	94.43	99.69
BL + ConvBlock	11,089,537	92.89	96.27	96.33	99.60	96.44	99.77
BL + ConvBlock + SEFB	15,339,329	93.51	96.80	96.58	99.69	96.66	99.82
BL + ConvBlock + SEFB + TAM	15,427,713	<b>94.75</b>	<b>97.28</b>	<b>97.15</b>	<b>99.74</b>	<b>97.42</b>	<b>99.84</b>

**Table 3.** Outcomes of ablation studies on the BUSI dataset.



**Fig. 4.** Heatmaps of SEFB at four skip-connection stages on the BUSI dataset. The first column shows the original image, the second column displays the ground truth, and the remaining columns present SEFB responses at successive stages, highlighting progressively focused lesion regions.

the evolving focus of the network, where shallow layers capture broad structural context and deeper layers progressively concentrate on more discriminative lesion boundaries. This stage-wise visualization confirms that SEFB effectively guides the network toward lesion-relevant regions, thereby improving the reliability of feature propagation through skip connections and contributing to more accurate segmentation outcomes.

**Results and discussion**  
**Comparison with SOTA methods on the BUSI dataset**

To comprehensively assess the efficiency of HA-Net, the outcomes of SOTA approaches on the BUSI breast ultrasound dataset are compared with the proposed HA-Net. The selected benchmark models encompass a range of architectures and design strategies, including classical encoder-decoder variants (U-Net, UNet++, Attention U-Net), transformer-based networks (Swin-UNet, Eh-Former, BGRD-TransUNet), and specialized attention-guided frameworks (BGRA-GSA, AAU-Net, MCRNet, DDRA-Net). These models represent the current landscape of approaches for the segmentation task and provide a rigorous reference for evaluating the HA-Net.

The quantitative outcomes are summarized in Table 4. The HA-Net consistently achieves competent performance across all quantitative metrics, including DSC, IoU, sensitivity, precision, specificity, and accuracy. The combined use of SFEB and TAM equips the model with the ability to emphasize detailed boundary

Method	Parameters (Million)	$J_i$	$D_c$	$S_n$	$A_{cc}$	$P_r$	$S_p$
BGRA-GSA <sup>57</sup>	101.34 M	68.75	81.43	84.14	96.34	79.01	97.63
AAU-Net <sup>31</sup>	29.2 M	69.26	78.18	86.06	-	81.17	99.17
MCRNet <sup>58</sup>	26.63 M	69.94	82.31	81.65	96.78	-	-
Swin-unet <sup>59</sup>	27.3 M	74.16	79.45	83.16	96.55	-	97.34
Eh-former <sup>60</sup>	184.2 M	76.37	84.6	87.74	-	-	98.17
U-Net <sup>15</sup>	34.51 M	76.54	83.13	82.83	97.91	83.94	98.81
BGRD-TransUNet <sup>61</sup>	109.65 M	76.77	85.08	87.62	97.14	85.89	-
Attention U-Net <sup>62</sup>	8.14 M	77.89	85.96	85.80	97.85	86.65	98.54
Unet++ <sup>63</sup>	9.04 M	81.09	88.11	87.29	98.57	89.53	99.18
DDRA-Net <sup>64</sup>	5.46 M	89.23	75.32	92.32	-	95.02	-
HA-Net (Proposed)	15.43 M	<b>94.75</b>	<b>97.28</b>	<b>97.15</b>	<b>99.74</b>	<b>97.42</b>	<b>99.84</b>

**Table 4.** Comparison with cutting-edge segmentation methods on the BUSI dataset.

Method	Parameters (Million)	$J_i$	$D_c$	$S_n$	$A_{cc}$	$P_r$	$S_p$
Unet++ <sup>63</sup>	9.04 M	67.59	78.64	79.52	96.48	78.43	97.52
U-Net <sup>15</sup>	34.51 M	69.60	79.96	81.85	97.03	78.80	97.83
AAU-Net <sup>31</sup>	29.2 M	72.08	80.45	81.62	-	80.67	97.81
Attention U-Net <sup>62</sup>	8.14 M	72.34	81.55	79.93	96.70	84.43	98.14
BGRA-GSA <sup>57</sup>	101.34 M	78.80	88.01	89.60	98.80	86.46	99.27
MCRNet <sup>58</sup>	26.63 M	81.90	90.05	89.18	98.96	-	-
Eh-former <sup>60</sup>	184.2 M	84.48	91.22	91.92	-	-	99.39
BGRD-TransUNet <sup>61</sup>	109.65 M	86.61	92.47	<b>92.78</b>	<b>99.15</b>	93.01	-
Proposed method	15.43 M	<b>86.71</b>	<b>92.38</b>	92.14	99.00	92.74	<b>99.51</b>

**Table 5.** Comparison with cutting-edge segmentation approaches on the UDIAT dataset.

information while retaining a broader contextual understanding. This architectural design enables the network to effectively handle common challenges in BUSI.

The improvements are particularly notable in metrics that emphasize overlap and boundary accuracy (DSC and IoU), highlighting the method’s ability to precisely delineate tumor regions. High sensitivity and precision scores further indicate that the model reliably identifies tumor pixels with lower false positives, which is critical in clinical practice for accurate diagnosis and reducing unnecessary interventions. Furthermore, the model maintains high specificity, demonstrating its ability to correctly classify normal tissue and avoid mislabeling background regions as lesions.

By effectively combining dense feature extraction, contextual information based on attention features, and spatial features refinement, the framework consistently outperforms existing SOTA methods, providing reliable and accurate segmentation results that could assist radiologists in early breast cancer detection.

*Statistical significance analysis*

To validate the observed performance improvements of HA-Net over other segmentation models on the BUSI dataset, we applied the Wilcoxon signed-rank test. Compared to Attention U-Net, HA-Net achieved a p-value of  $1.55 \times 10^{-14}$ , and against U-Net, the p-value was  $2.71 \times 10^{-14}$ . These highly significant results confirm that the superior performance of HA-Net is statistically robust, highlighting its reliability and effectiveness.

**Comparison with SOTA methods on the UDIAT dataset**

The generalization capability of HA-Net was further examined through comparative experiments on the UDIAT breast ultrasound dataset. The benchmarked models encompass a wide range of recent SOTA approaches, including BGRA-GSA, AAU-Net, MCRNet, Swin-UNet, Eh-Former, U-Net, BGRD-TransUNet, Attention U-Net, and Unet++. These models collectively represent diverse architectural strategies, from encoder-decoder networks to attention-guided and transformer-based frameworks, providing a robust reference for performance assessment.

As presented in Table 5, HA-Net demonstrates strong and consistent performance across all evaluated metrics. It attains the highest scores in Jaccard Index, Dice coefficient, and specificity, which are critical indicators of precise tumor localization and accurate segmentation boundaries. Although BGRD-TransUNet exhibits slightly higher sensitivity and overall accuracy, our model demonstrates a more balanced performance profile, with notable advantages in overlap-based metrics that are particularly relevant for assessing segmentation quality in medical imaging.

These findings highlight the robustness and adaptability of the model across datasets with diverse imaging conditions and tumor characteristics, thereby demonstrating its strong potential for reliable integration into real-world clinical breast cancer diagnosis and screening workflows. Consistent results on UDIAT further demonstrate the suitability of the proposed HA-Net for clinical deployment, supporting its role in accurate tumor segmentation for early diagnosis and effective treatment planning.

#### Statistical significance analysis

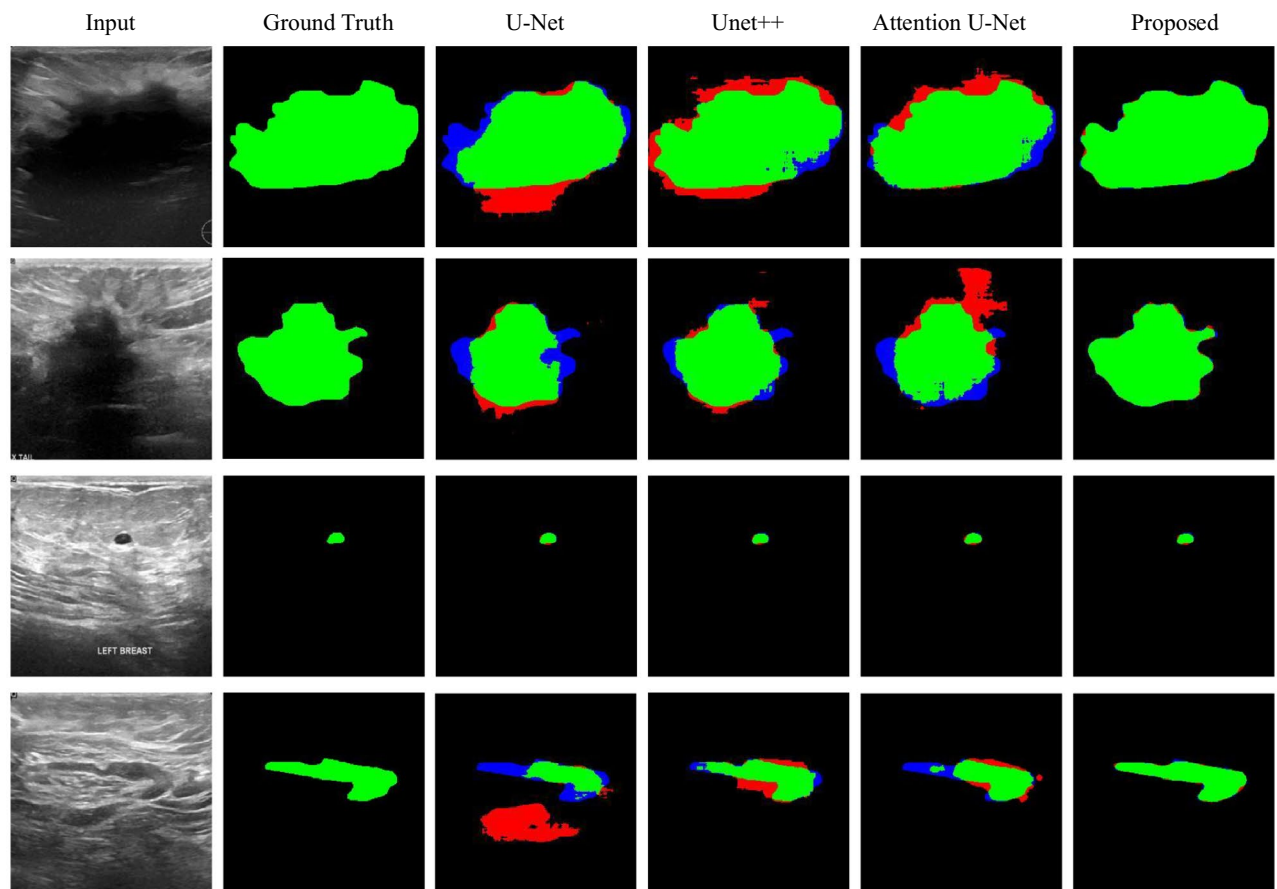
To evaluate the statistical significance of HA-Net's performance on the UDIAT dataset, a Wilcoxon signed-rank test was conducted, comparing the proposed model against Attention U-Net and U-Net. The results indicate a p-value of  $1.76 \times 10^{-6}$  when compared to Attention U-Net and  $1.47 \times 10^{-5}$  against U-Net. These highly significant values demonstrate that HA-Net's superior segmentation performance is statistically robust, confirming its effectiveness and reliability in accurately delineating breast tumors in ultrasound images.

#### Qualitative visualization results

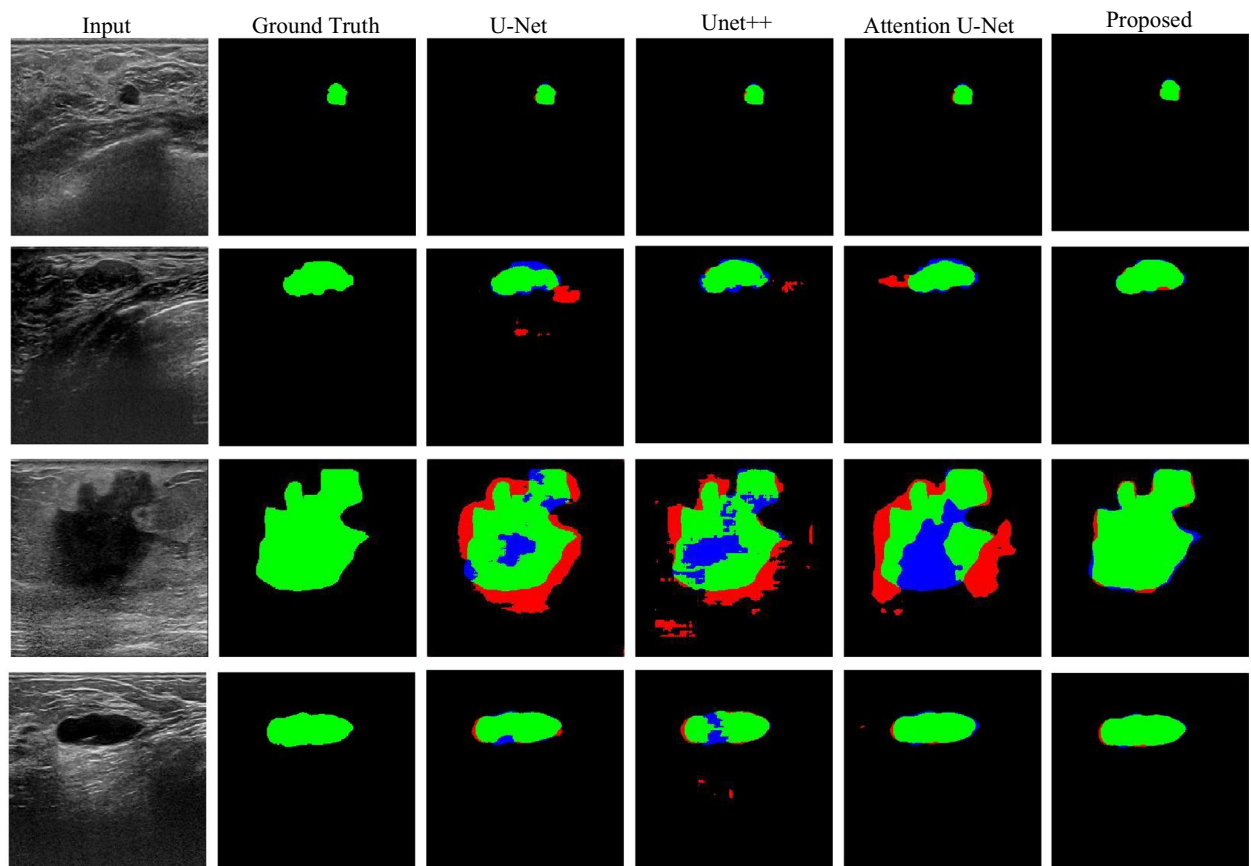
To complement the quantitative results, we also provide qualitative segmentation examples from both the BUSI and UDIAT datasets. Figure 5 presents side-by-side comparisons between HA-Net and representative SOTA models, including U-Net<sup>15</sup>, UNet++<sup>63</sup>, and Attention U-Net<sup>62</sup>, on the BUSI dataset. These visual results emphasize how different methods perform on challenging conditions, such as poor contrast, heterogeneous lesion boundaries, and varying tumor sizes.

The proposed HA-Net consistently produces more precise segmentation boundaries with higher spatial alignment to the ground truth. It effectively suppresses false positives (highlighted in red) and recovers missed tumor regions (highlighted in blue), resulting in cleaner and more reliable segmentation maps. The SFEB and TAM contribute to these improvements by enhancing both local detail and global contextual understanding.

Similarly, Fig. 6 shows qualitative outcomes on the UDIAT dataset, comparing the proposed HA-Net against U-Net, UNet++, and Attention U-Net. The outcomes indicate that the model maintains high segmentation fidelity even in the presence of noise, low-intensity contrast, and irregular tumor morphology. These visualizations reinforce the quantitative findings reported earlier, particularly improvements in Dice coefficient, Jaccard Index, and specificity, emphasizing the model's robustness.



**Fig. 5.** Qualitative comparisons with state-of-the-art methods on the BUSI dataset are illustrated, where green highlights correctly segmented tumor regions, red marks false-positive detections, and blue indicates missed tumor areas.



**Fig. 6.** Visual results comparison with SOTA approaches on the UDIAT dataset.

## Discussion

The HA-Net consistently demonstrates competent performance, outperforming recent SOTA models in critical metrics such as Dice coefficient, Jaccard index, and specificity. These advancements underscore the significance of the proposed HA-Net, which synergistically combines dense feature extraction, a spatial feature enhancement block, and a Transformer-based attention module. The model exhibits a strong ability to accurately delineate lesions even in challenging imaging conditions characterized by low contrast, speckle noise, and irregular tumor morphology, highlighting its robustness and generalizability.

Statistical analyses further validate the significance of these performance gains, particularly when compared against leading segmentation frameworks such as UNet++, Attention U-Net, and BGRD-TransUNet. These findings reinforce the potential clinical utility of HA-Net as a reliable tool for automated breast cancer detection and decision support in real-world scenarios.

Despite these competent results, limitations should be acknowledged. First, the model was trained and evaluated on two datasets of moderate size. While the outcomes are encouraging, further validation on larger, multi-center, or multi-device ultrasound datasets is essential to fully assess generalizability. Second, the proposed method can exhibit reduced performance on images with extremely low contrast or poorly defined lesion boundaries, which may challenge accurate feature extraction and segmentation. Future work could address this limitation by incorporating advanced contrast enhancement techniques, adaptive preprocessing, or specialized attention mechanisms to better handle such challenging cases.

General-purpose backbones such as OverLoCK<sup>65</sup>, SparX<sup>66</sup>, TransXNet<sup>67</sup>, and SegMAN<sup>68</sup> have advanced visual recognition through novel architectural strategies, but their validation largely relies on natural image benchmarks. In contrast, HA-Net is explicitly designed for breast ultrasound segmentation, addressing domain-specific challenges including speckle noise, scale variation, and indistinct tumor boundaries. HA-Net combines a DenseNet-121 encoder with hybrid attention modules GSA, PE, and SDPA to model long-range dependencies and contextual feature interactions. The inclusion of SFEB in skip connections strengthens lesion-specific spatial details, while a composite BCE and Jaccard loss ensures balanced optimization across pixel- and region-level accuracy. Compared with OverLoCK's biologically inspired attention, SparX's cross-layer aggregation, and TransXNet's dynamic token mixing, HA-Net adapts these concepts more effectively to the clinical setting by prioritizing feature clarity, spatial refinement, and noise robustness. Unlike SegMAN, which targets large-scale semantic segmentation, HA-Net demonstrates how domain-adaptive design can substantially improve segmentation reliability under the complex conditions of BUS imaging.

Method	Parameters (Million)	$J_i$	FLOPs(G)	MemorySize(MB)
BGRA-GSA <sup>57</sup>	101.34 M	68.75	334.76	-
AAU-Net <sup>31</sup>	29.2 M	69.26	370.00	-
MCRNet <sup>58</sup>	26.63 M	69.94	13.29	-
U-Net <sup>15</sup>	34.51 M	76.54	46.02	134.82
Attention U-Net <sup>62</sup>	8.14 M	77.89	45.62	31.81
Unet++ <sup>63</sup>	9.04 M	81.09	28.45	35.321
HA-Net (Proposed)	15.43 M	<b>94.75</b>	8.01	60.62

**Table 6.** Comparison of computational complexity of different models on the BUSI dataset.

Computational complexity

To provide a rigorous evaluation of the proposed HA-Net against SOTA approaches, a computational complexity analysis was conducted on the BUSI dataset. The primary goal of this analysis is to establish a trade-off between segmentation performance and computational complexity, where resource-constrained environments such as portable ultrasound scanners or clinical workstations are common. The comparison considers several complementary metrics. First, the number of trainable parameters is reported, which directly reflects the capacity of the model and its tendency toward overfitting or generalization. Models with fewer parameters generally require less storage and faster inference but may sacrifice representational power if overly simplified. Second, the IoU is adopted as the main performance metric, as it provides a reliable measure of region overlap between the predicted segmentation mask and the ground truth annotation. This metric is especially suitable for medical segmentation tasks, where precise boundary delineation is critical. Alongside segmentation accuracy, we also report the floating-point operations (FLOPs), which represent the theoretical computational cost of a single forward pass through the network. A lower FLOP count reflects reduced arithmetic complexity, thereby improving the model's suitability for real-time clinical deployment. Finally, the memory footprint is reported, capturing the storage and runtime memory requirements. This measure is crucial in scenarios where computational resources are limited, such as edge devices or cloud-based telemedicine applications. By integrating these four metrics, parameters, IoU score, FLOPs, and memory consumption provide a comprehensive view of model performance that extends beyond accuracy alone. The results, summarized in Table 6, enable a fair comparison between methods and highlight the balance between predictive reliability and computational feasibility, thereby guiding the choice of models for practical medical imaging applications.

Conclusion

This study introduces HA-Net, a hybrid attention-based architecture specifically designed for the automated segmentation of breast tumors in ultrasound images. The architecture leverages a pre-trained DenseNet-121 encoder combined with an attention mechanism incorporating Global Spatial Attention (GSA), Position Encoding (PE), and Scaled Dot-Product Attention (SDPA), thereby allowing the model to effectively capture global contextual relationships while preserving fine-grained spatial details that are critical for precise tumor delineation. The Spatial Feature Enhancement Block was integrated into skip connections to preserve high-resolution information and refine focus on tumor regions. The segmentation process is guided by a combined loss function, thereby effectively mitigating challenges arising from class imbalance and the heterogeneous morphologies of breast lesions. Comprehensive experiments conducted on the BUSI and UDIAT datasets show that HA-Net consistently surpasses existing SOTA segmentation methods across multiple evaluation metrics. Both quantitative and qualitative assessments validate its robustness, high performance, and generalizability, highlighting its potential utility as a clinical tool for facilitating early and precise breast cancer diagnosis.

Future work will aim to improve cross-device and multi-center generalization via domain adaptation, incorporate lesion classification to create a comprehensive diagnostic framework, and enable real-time deployment in clinical workflows to enhance diagnostic efficiency and patient care.

Data availability

The data used in this research is publicly available for research and development purposes at our GitHub repository: <https://github.com/nisarahmedrana/HA-Net>.

Received: 4 June 2025; Accepted: 6 October 2025

Published online: 12 November 2025

References

1. Zhang, S., Jin, Z., Bao, L. & Shu, P. The global burden of breast cancer in women from 1990 to 2030: assessment and projection based on the global burden of disease study 2019. *Front. Oncol.* **14**, 1364397 (2024).
2. Zheng, D., He, X. & Jing, J. Overview of artificial intelligence in breast cancer medical imaging. *J. clinical medicine* **12**, 419 (2023).
3. Karellas, A. & Vedantham, S. Breast cancer imaging: a perspective for the next decade. *Med. physics* **35**, 4878–4897 (2008).
4. Benson, S., Blue, J., Judd, K. & Harman, J. Ultrasound is now better than mammography for the detection of invasive breast cancer. *The Am. journal surgery* **188**, 381–385 (2004).
5. Madjar, H. Role of breast ultrasound for the detection and differentiation of breast lesions. *Breast Care* **5**, 109–114 (2010).
6. Gonzaga, M. A. How accurate is ultrasound in evaluating palpable breast masses? *Pan Afr. Med. J.* **7** (2010).

7. Liu, L. et al. Automated breast tumor detection and segmentation with a novel computational framework of whole ultrasound images. *Med. & biological engineering & computing* **56**, 183–199 (2018).
8. Ahmed, N., Asif, H. M. S. & Khalid, H. Piqu: perceptual image quality index based on ensemble of gaussian process regression. *Multimed. Tools Appl.* **80**, 15677–15700 (2021).
9. Khalid, H., Ali, M. & Ahmed, N. Gaussian process-based feature-enriched blind image quality assessment. *J. Vis. Commun. Image Represent.* **77**, (2021).
10. Ahmed, N. & Asif, H. M. S. Perceptual quality assessment of digital images using deep features. *Comput. Informatics* **39**, 385–409 (2020).
11. Ahmed, N., Shahzad Asif, H., Bhatti, A. R. & Khan, A. Deep ensembling for perceptual image quality assessment. *Soft Comput.* **26**, 7601–7622 (2022).
12. Aslam, M. A. et al. Vrl-iqu: Visual representation learning for image quality assessment. *IEEE Access* **12**, 2458–2473 (2023).
13. Aslam, M. A. et al. Qualitynet: A multi-stream fusion framework with spatial and channel attention for blind image quality assessment. *Sci. Reports* **14**, 26039 (2024).
14. Aslam, M. A. et al. Tqp: An efficient video quality assessment framework for adaptive bitrate video streaming. *IEEE Access* (2024).
15. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* **18**, 234–241 (Springer, 2015).
16. Song, K., Feng, J. & Chen, D. A survey on deep learning in medical ultrasound imaging. *Front. Phys.* **12**, 1398393 (2024).
17. Huang, Z., Wang, L. & Xu, L. Dra-net: Medical image segmentation based on adaptive feature extraction and region-level information fusion. *Sci. Reports* **14**, 9714 (2024).
18. Anari, S., Sadeghi, S., Sheikhi, G., Ranjbarzadeh, R. & Bendechache, M. Explainable attention based breast tumor segmentation using a combination of unet, resnet, densenet, and efficientnet models. *Scientific Reports* **15**, 1027 (2025).
19. Fang, W. & Han, X.-h. Spatial and channel attention modulated network for medical image segmentation. In *Proceedings of the Asian conference on computer vision* (2020).
20. Murase, R., Suganuma, M. & Okatani, T. How can cnns use image position for segmentation? arXiv preprint [arXiv:2005.03463](https://arxiv.org/abs/2005.03463) (2020).
21. Shen, X. et al. Dilated transformer: residual axial attention for breast ultrasound image segmentation. *Quant. Imaging Medicine Surg.* **12**, 4512 (2022).
22. He, J. et al. Sab-net: Self-attention backward network for gastric tumor segmentation in ct images. *Comput. Biol. Medicine* **169**, (2024).
23. Zhang, H. et al. Acl-dunet: A tumor segmentation method based on multiple attention and densely connected breast ultrasound images. *PloS one* **19**, e0307916 (2024).
24. Byra, M. et al. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomed. Signal Process. Control.* **61**, (2020).
25. Zhang, S. et al. Fully automatic tumor segmentation of breast ultrasound images with deep learning. *J. Appl. Clin. Med. Phys.* **24**, (2023).
26. Michael, E., Ma, H., Li, H., Kulwa, F. & Li, J. Breast cancer segmentation methods: current status and future potentials. *BioMed research international* **2021**, 9962109 (2021).
27. Xu, Y., Liu, F., Xu, W. & Quan, R. Overview of graph theoretical approaches in medical image segmentation. In *International Conference on Computational & Experimental Engineering and Sciences*, 819–835 (Springer, 2024).
28. Li, L., Niu, Y., Tian, F. & Huang, B. An efficient deep learning strategy for accurate and automated detection of breast tumors in ultrasound image datasets. *Front. Oncol.* **14**, 1461542 (2025).
29. Pan, P. et al. Tumor segmentation in automated whole breast ultrasound using bidirectional lstm neural network and attention mechanism. *Ultrasonics* **110**, 106271 (2021).
30. Abraham, N. & Khan, N. M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, 683–687 (IEEE, 2019).
31. Chen, G., Li, L., Dai, Y., Zhang, J. & Yap, M. H. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging* **42**, 1289–1300 (2022).
32. Chen, G. et al. Esknet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation. *Expert. Syst. with Appl.* **246**, (2024).
33. Xu, C. et al. Arf-net: An adaptive receptive field network for breast mass segmentation in whole mammograms and ultrasound images. *Biomed. Signal Process. Control.* **71**, (2022).
34. Pi, J. et al. Fs-unet: Mass segmentation in mammograms using an encoder-decoder architecture with feature strengthening. *Comput. Biol. Medicine* **137**, (2021).
35. Ma, Z. et al. Atfe-net: axial transformer and feature enhancement-based cnn for ultrasound breast mass segmentation. *Comput. Biol. Medicine* **153**, (2023).
36. Xiao, H., Li, L., Liu, Q., Zhu, X. & Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control.* **84**, (2023).
37. Liu, Q. et al. Optimizing vision transformers for medical image segmentation. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 1–5 (IEEE, 2023).
38. Zhang, J., Li, F., Zhang, X., Wang, H. & Hei, X. Automatic medical image segmentation with vision transformer. *Appl. Sci.* **14**, 2741 (2024).
39. He, Q., Yang, Q. & Xie, M. Hctnet: A hybrid cnn-transformer network for breast ultrasound image segmentation. *Comput. Biol. Medicine* **155**, (2023).
40. Lin, A. et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation Meas.* **71**, 1–15 (2022).
41. Zhu, C. et al. Swin-net: a swin-transformer-based network combing with multi-scale features for segmentation of breast tumor ultrasound images. *Diagnostics* **14**, 269 (2024).
42. Zhao, Z. et al. Swinhr: Hemodynamic-powered hierarchical vision transformer for breast tumor segmentation. *Comput. biology medicine* **169**, (2024).
43. Cao, W. et al. Neighbornet: Learning intra-and inter-image pixel neighbor representation for breast lesion segmentation. *IEEE J. Biomed. Heal. Informatics* (2024).
44. Zhang, H. et al. Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation. *Biomed. Signal Process. Control.* **87**, (2024).
45. Wu, R., Lu, X., Yao, Z. & Ma, Y. Mfmsnet: A multi-frequency and multi-scale interactive cnn-transformer hybrid network for breast ultrasound image segmentation. *Comput. Biol. Medicine* **177**, (2024).
46. Taghnamas, J., Ramadan, H., Yahyaouy, A. & Tairi, H. Multi-task approach based on combined cnn-transformer for efficient segmentation and classification of breast tumors in ultrasound images. *Vis. Comput. for Ind. Biomed. Art* **7**, 2 (2024).
47. Zhang, Z. et al. A novel deep learning model for medical image segmentation with convolutional neural network and transformer. *Interdiscip. Sci. Comput. Life Sci.* **15**, 663–677 (2023).
48. Xiong, Y., Shu, X., Liu, Q. & Yuan, D. Hcmnet: A hybrid cnn-mamba network for breast ultrasound segmentation for consumer assisted diagnosis. *IEEE Transactions on Consumer Electron.* 1–1, <https://doi.org/10.1109/TCE.2025.3593784> (2025).

49. Zhu, H. et al. Attnmnet: a hybrid transformer integrating self-attention, mamba, and multi-layer convolution for enhanced lesion segmentation. *Quant. Imaging Medicine Surg.* **15**, 4296–4310, <https://doi.org/10.21037/qims-2024-2561> (2025).
50. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
51. Chen, B., Liu, Y., Zhang, Z., Lu, G. & Kong, A. W. K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerg. Top. Comput. Intell.* (2023).
52. Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 1–7 (IEEE, 2020).
53. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in brief* **28**, 104863 (2020).
54. Yap, M. H. et al. Breast ultrasound region of interest detection and lesion localisation. *Artif. intelligence medicine* **107**, (2020).
55. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
56. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).
57. Hu, K. et al. Boundary-guided and region-aware network with global scale-adaptive for accurate segmentation of breast tumors in ultrasound images. *IEEE J. Biomed. Heal. Informatics* **27**, 4421–4432 (2023).
58. Lou, M., Meng, J., Qi, Y., Li, X. & Ma, Y. Mcrnet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging. *Neurocomputing* **470**, 154–169. <https://doi.org/10.1016/j.neucom.2021.10.102> (2022).
59. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision—ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 205–218 (Springer, 2023).
60. Qu, X. et al. Eh-former: Regional easy-hard-aware transformer for breast lesion segmentation in ultrasound images. *Inf. Fusion* **109**, (2024).
61. Ji, Z. et al. Bgrd-transunet: A novel transunet-based model for ultrasound breast lesion segmentation. *IEEE Access* (2024).
62. Oktay, O. et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999) (2018).
63. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
64. Sun, J. et al. Ddra-net: Dual-channel deep residual attention upernet for breast lesions segmentation in ultrasound images. *IEEE Access* (2024).
65. Lou, M. & Yu, Y. Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 128–138 (2025).
66. Lou, M., Fu, Y. & Yu, Y. Sparx: A sparse cross-layer connection mechanism for hierarchical vision mamba and transformer networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* **39**, 19104–19114 (2025).
67. Lou, M. et al. Transxnet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks Learn. Syst.* (2025).
68. Fu, Y., Lou, M. & Yu, Y. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19077–19087 (2025).

## Acknowledgements

All authors thank the School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi, China, for their Financial Support and Funding.

## Author contributions

M.A.A. contributed to the interpretation of results and participated in manuscript review and editing. A.N. conceived the research idea, designed the methodology, conducted the experiments, performed data analysis, and led the manuscript drafting. N.A. contributed to methodology development, assisted in data analysis, and supported manuscript preparation. Z.K. assisted with experiments, contributed to data analysis, and supported manuscript review and editing. All authors reviewed and approved the final version of the manuscript.

## Funding

The study is supported by the School of Information Engineering, Xi'an Eurasia University, Xi'an, Shanxi, China. The authors declare no conflict of interest.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.A.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025