Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

SWINUNET: A HYBRID SWIN TRANSFORMER-CNN ARCHITECTURE WITH ADAPTIVE FEATURE FUSION FOR BREAST ULTRASOUND LESION SEGMENTATION

Saeeda Naz[1], Saddam Hussain Khan[*2], Rashid Iqbal[3], Muhammad Safiullah[4]

[1,*2,3]*Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences (UEAS), Swat 19060, Pakistan,*
[4]*Department of International Graduate School of Artificial Intelligence, University of National Yunlin University of Science and Technology*

[1]saeedanaz396@gmail.com, [*2]hengrshkhan822@gmail.com, [3]rashidibms1@gmail.com, [4]m11363013@yuntech.edu.tw

DOI: https://doi.org/10.5281/zenodo.18383193

Keywords
Breast Cancer, Ultrasound Imaging, Image Segmentation, CNN, ViT, Swin Transformer

Article History
Received: 18 November 2025
Accepted: 25 December 2025
Published: 27 January 2026

Corresponding Author: *
Saddam Hussain Khan

Abstract
Breast ultrasound imaging is a very safe and cost-effective lesion detection technique; however, exact lesion segmentation is a difficult task because of its low contrast, speckle noise, and unclear boundaries. To overcome these limitations, this work presents SwinUNet, a novel hybrid model that leverages the strong global modeling ability of a Swin-transformer(encoder) and the strong local modeling ability of a convolutional decoder. The main idea of this work is the utilization of a novel Adaptive Feature Fusion (AFF) module inside the skip connections, aiming to adaptively weigh channel-wise features, thus highlighting lesion-related feature maps while eliminating background noise inherent to ultrasound images. SwinUNet was tested for its efficacy on the publicly available BUSI breast ultrasound imaging dataset following a five-fold cross-validation setup. The results clearly show that SwinUNet can effectively achieve a Mean Intersection over Union of 89.93%, with a Global Accuracy of 97.68%, thus outperforming traditional CNN models as well as baseline transformer models. Furthermore, ablation tests have confirmed that the AFF module plays a crucial role in achieving these improvements over traditional techniques. These findings clearly demonstrate SwinUNet's ability for accurate lesion segmentation, thus making it a potential candidate for inclusion within a computer-aided diagnostic environment. This confirms that the suggested deep network presents accurate lesion delineation capability, emphasizing its use to assist medical specialists through computer-aided tools.

INTRODUCTION

Breast cancer continues to be a leading cause of cancer-related deaths in the world; hence, an efficient diagnosis is critical in ensuring that treatment strategies are effective [1]. Ultrasound is one of the widely employed non-invasive techniques used in breast lesion analysis [2], [3]. However, despite the efficiency of the technique in breast lesion analysis, particularly when breast tissue is dense, the analysis is often followed by human interpretation involving the manual demarcation of lesions from breast ultrasound images.

The incorporation of deep learning techniques in the analysis of medical images has opened up

hopeful directions for the automation and standardization of breast lesion analysis in ultrasound images[4]–[6] The application of Convolutional Neural Networks (CNNs), most prominently the U-Net [7] network, has become the most popular approach in the analysis of medical images for the extraction of specific features [8]. The localized filter response of the convolution operation makes it inherently less capable of processing long-range information, which is essential in the analysis of breast ultrasound images.

Vision Transformers (ViTs) mitigate the above disadvantage by using self-attention mechanisms for the global relations between the pixels of an image [9]. However, traditional Vision Transformers remain plagued by a computational cost proportional to the square of the image resolution, thereby lacking the necessary inductive bias for accurate localization at the pixel-level [10]. However, with the advent of the Swin Transformer [11], which proposed a hierarchical model of a shifted window-based attention mechanism [12], there has been efficient global context learning for high-resolution images [13]. This resulted in the rise of new architectures that combined the strengths of transformers regarding global learning with the spatial acuity of convolutional neural networks, thereby appreciably enhancing the localization abilities of Vision Transformers for accurate segmentation of objects at the pixel-level [14], [15]. One of the key difficulties encountered while designing such hybrid models is appropriate feature fusion between the rich global features of the encoder and the fine features of the decoder, as simple feature concatenation leads to the spread of redundant information, eventually causing a deterioration of the segmentation quality.

For overcoming these challenges, the proposed study introduces a hybrid encoder and decoder architecture for lesion segmentation in breast ultrasound images, named SwinUNet. The proposed network combines the advantages of the Swin Transformer (SwinT) encoder and the CNN decoder. The most important part of the proposed network is the Adaptive Feature Fusion (AFF) component in the skip connection, wherein channel attention is used to weigh the encoder feature map based on the importance of the lesions and eliminate the ultrasound artifact and background noise. The contributions of this study are outlined below:

Developing SwinUNet, which is a hybrid model that integrates SwinTs for global perspective modeling and a CNN for local boundary description.

Adaptive Feature Fusion module: focused on the improvement of the robustness of multi-resolution skip connections against speckle noise and low-contrast boundaries of the BUS image through attention-weighted feature gating.

State-of-the-art experimental assessment on two public breast ultrasound image datasets, including BUSI, to prove that our method has a competitive performance level compared to the latest state-of-the-art CNN and Transformer segmentation approaches.

This paper is divided into sections 2 and 3, containing earlier studies and methodology. The experimental setup is presented in Section 4. Section 5 contains a discussion on the results, whereas Section 6 illustrates the conclusion.

2. Related Work

2.1. CNN-Based Medical Image Segmentation

The rise of CNNs has significantly transformed the field of medical image segmentation.[16], [17] The U-Net model, with its symmetric encoder-decoder structure and skip connections, had set the standard by achieving a proper blend of contextual and spatial information [18]. This approach opened avenues to a plethora of variants and improvements for greater accuracy in image segmentations. The U-Net++ model added nested and dense skip connections to bridge the semantic gap between features of the encoder and decoder layers, and the Attention U-Net used attention gates to draw more highlights on the skip connections [19]. Others, such as ResUNet++ and DeepLabV3, had advanced the technology with the addition of ideas of residual learning and atrous spatial pyramid pooling, respectively [20]–[22]

Although such improvements have been achieved, there is a major drawback: Convolutions involve the use of receptive fields. This directly impacts the capacity of CNN models to incorporate global contextual information. This can be a major concern when breast ultrasound images are being processed, as the pattern of breast cancer may be quite ambiguous and require global contextual information.

2.2. Vision Transformers for Global Context Modeling

Transformers utilized self-attention modules for capturing the global relations and were first proposed in the NLP tasks and later extended to computer vision as ViTs in [20]. The use of ViTs in the context of the medical image segmentation task, as in TransUNet, validated the improvement in bringing the benefit of capturing the global context information for dense prediction tasks [18]. Moreover, the standard self-attention modules are quadratically scaled with the resolution, and thus the standard ViTs are computationally costly for processing large-sized resolution inputs in the case of the medical imaging domain. Moreover, [23] the standard ViTs also lack the inbuilt strong inductive biases for the spatial contexts, as in CNNs, which might be a hindrance for the fine-grained pinpointing task in the case of ultrasound modalities with unclear boundaries.

2.3. Efficient and Hybrid CNN-Transformer Models

Intending to overcome the efficiency issues associated with the standard ViTs' self-attention layers, [12] the Swin Transformer brought a novel hierarchical shifted window self-attention strategy that supports linear computational complexity while promoting the global receptive field [24], [25]. This marked the beginning of the development of CNN-Transformer-based architectures that utilize the benefits of both global reasoning and spatial precision with CNN and Swin Transformers as the encoder and CNN as the decoder [26]–[29]. The concept was proven successful for applications like medical image segmentation, as shown in the case of the Swin-UNet.TransFuse and SwinBTS++ employed novel ways for the combination of CNN and transformer operations based on either parallel and series configurations [30], [31]. And all the related work can be seen from **Table 1.**

Table 1: review of selected papers

Ref.	Author's Name and Year	Dataset_ Used	Models / Methods	Performance (Reported Metrics)
[32]	Eisemann et al. (25)	Nationwide mammography screening data	Clinical AI CAD system	AUC ≈ 0.90 (real-world study)
[33]	Mehmood et al. (25)	BUS datasets	CNN-Transformer with boundary learning	Accuracy ≈ 95%
[34]	Chen et al. (24)	Medical segmentation datasets	TransUNet	Dice ≈ 0.87
[35]	Zhang et al. (24)	BUS datasets	Hybrid CNN-Transformer	Dice ≈ 0.90
[36]	Zhou et al. (23)	BUS datasets	Transformer U-Net	Dice ≈ 0.91
[36	Luo et al. (22)	BUS datasets	Attention-based CNN	Accuracy ≈ 94%
[37]	Cao et al. (22)	Medical segmentation datasets	Swin-UNet	Dice ≈ 0.88
[38]	Zhou et al. (21)	3D ABUS	Multi-task CNN	Dice ≈ 0.82
[39]	Qin et al. (20)	Salient object datasets	U²-Net	Dice ≈ 0.88
[40]	Yala et al. (19)	Large screening mammography cohort	DL risk prediction CNN	AUC = 0.76
[41]	Zhou et al. (18)	Medical segmentation datasets	U-Net++	Dice ≈ 0.89
[42]	Oktay et al. (18)	Medical CT/MRI	Attention U-Net	Dice improvement over U-Net
[43],	Ronneberger et al. (15)	ISBI biomedical datasets	U-Net	Dice ≈ 0.92

Yet, the fusion aspect in the feature fusion mechanism in the hybrid architecture remains a challenge. The concatenation of the encoder and the decoder features in the skip connection might result in the transmission of redundant information and ultrasound-specific artifacts, such as speckle patterns, hindering the definition of the boundary for lesions in the model. The proposed solution for this problem in the SwinUNet is the incorporation of the Adaptive Feature Fusion technique in the skip connection.

3. Methodology

3.1. General Architecture

The SwinUNet proposed herein is a hybrid encoder–decoder network tailored for accurate segmentation of breast ultrasound lesions [44]. As shown in Figure 1, its overall topology is in a U-shape. The encoder relies on a Swin Transformer backbone, constructing representations with hierarchically multi-scale features, while the decoder is a CNN that gradually recovers spatial resolution for dense pixel-wise prediction [45]. The encoder and decoder are interconnected with multi-resolution skip connections, each of which is enhanced with an Adaptive Feature Fusion (AFF) module. Such a structure leverages the strengths brought about by the Swin Transformer in capturing long-range contextual dependencies, along with the local spatial feature extraction capabilities provided by CNN. The AFF modules perform scale-aware feature selection and integration.

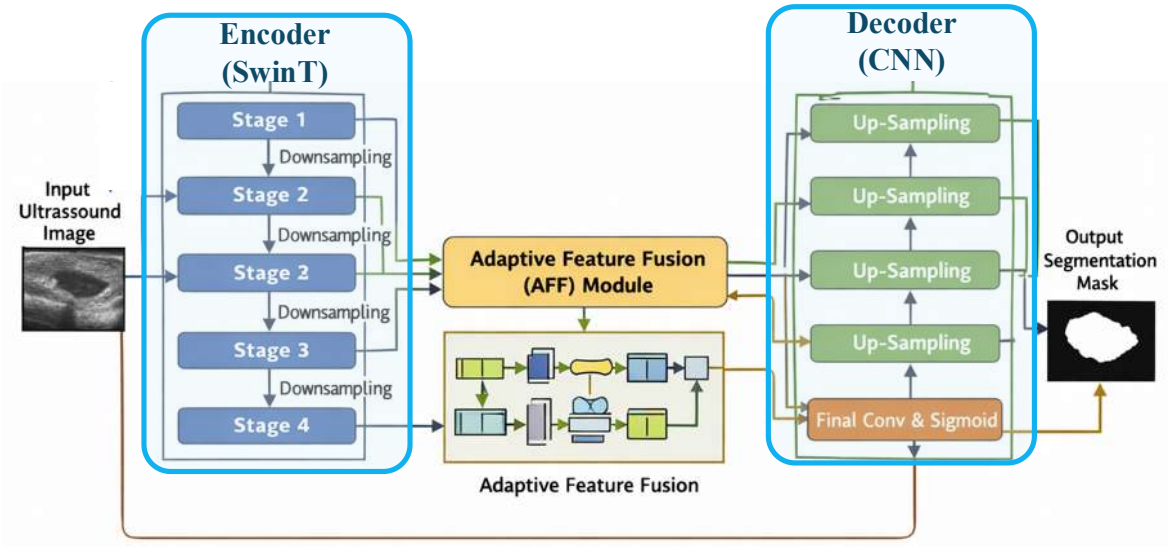

Figure 1: The proposed SwinUNet framework for BUSI lesion segmentation.

3.2 Swin Transformer Encoder

Therefore, it adopts an effective Swin Transformer architecture to capture global contextual information efficiently. Unlike standard Vision Transformers that compute self-attention across all image patches and incur quadratic computational complexity, the Swin Transformer adopts a hierarchical shifted-window attention mechanism, enhancing the computational efficiency for high-resolution images considerably.

[44] In detail, feature maps are divided into non-overlapping local windows, inside which Window-based Multi-Head Self-Attention (W-MSA) is computed. In subsequent layers, the Shifted Window-based MSA (SW-MSA) is performed by shifting the partitioning of windows, allowing exchange of information across neighboring windows and expanding the effective receptive field progressively, as illustrated in Figure 2.

It forms a multiscale feature pyramid. Let the outputs of the four encoder stages be $\mathbf{E}_i, i \in \{1,2,3,4\}$, where $\mathbf{E}_1$ has the highest spatial

resolution. Successive stages contain several Swin Transformer blocks with residual connections and layer normalization added to ensure stable optimization and effective gradient propagation.

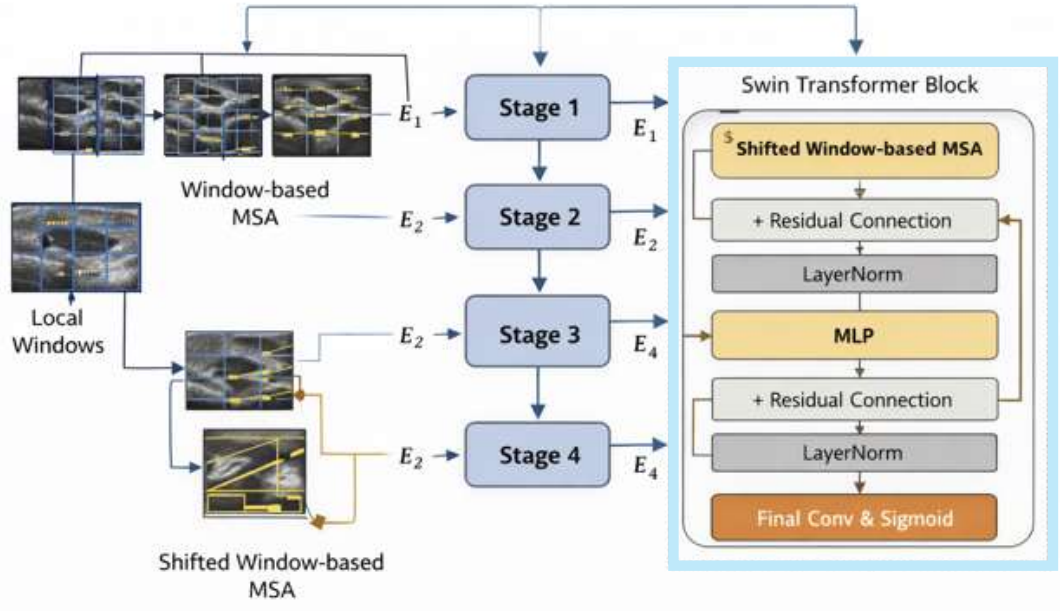

Figure 2: The proposed Swin Transformer (Encoder) block.

3.3. CNN Decoder with Adaptive Feature Fusion

In fact, [46] the CNN-based decoder is used for upsampling $\mathbf{E_4}$ symmetrically to the original size of the input feature map $\mathbf{E_0}$. Each of the stages of the decoder upsamples features, followed by convolutional blocks [47]. What makes this framework novel is its use of the Adaptive Feature Fusion (AFF) module, which is an alternative solution in place of direct feature concatenation in skip connections, and whose functionality is shown in Figure 3, where the channel-wise feature recalibration process occurs after every encoder feature map $\mathbf{E}_i$ and its corresponding upsampled decoder feature map $\mathbf{D}_{i+1}^{up}$. Firstly, the statistics for each channel are calculated through global average pooling from $\mathbf{E}_i$: $= \text{GAP}(\mathbf{E}_i)$, where $\mathbf{z} \in \mathbb{R}^C$. A gating function with a sigmoid activation function produces the attention vector as represented in equation 1:

$$\mathbf{a} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})) \quad (1)$$

Where $\mathbf{W}_1 \in \mathbb{R}^{C/r \times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times C/r}$, δ denotes the ReLU activation function, σ is the sigmoid activation function, and r is the reduction ratio (fixed at 16). The recalibrated encoder features are obtained through element-wise multiplication: $\tilde{\mathbf{E}}_i = \mathbf{a} \otimes \mathbf{E}_i$. The gated features are concatenated with the upsampled features of the decoder: $\mathbf{D}_i = \text{Conv}([\tilde{\mathbf{E}}_i, \mathbf{D}_{i+1}^{up}])$. This step removes the irrelevant background information present in the encoder features and allows the decoder to focus on the context.

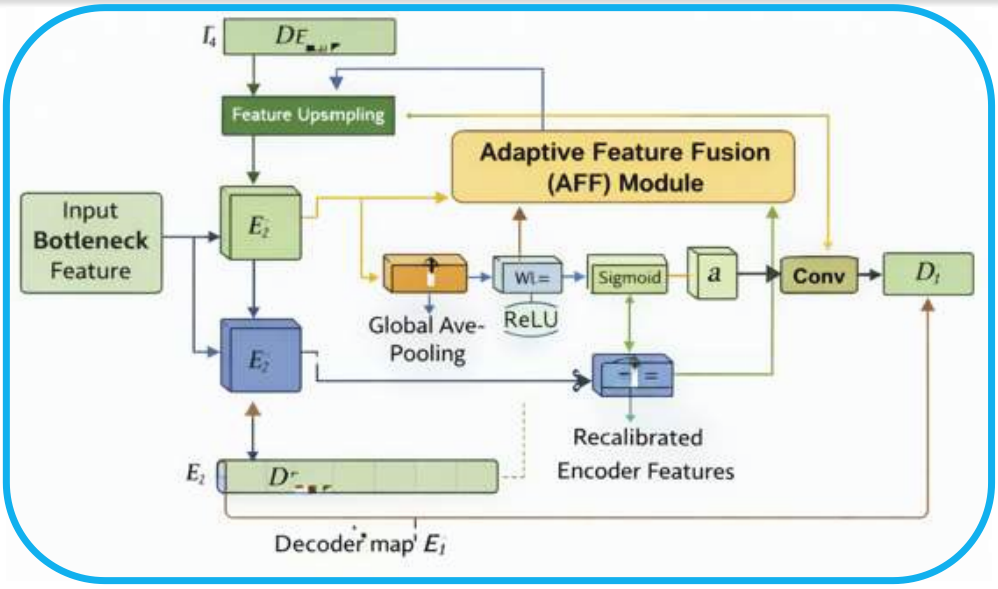

Figure 3 CNN Encoder with Adaptive Feature Fusion

Moreover, the Symmetric CNN Decoder is a method that progressively up-samples the feature maps of the CNN encoder from the bottleneck to the original size of the images. It is comprised of stages that involve upsampling layers followed by convolutional layers. However, the key breakthrough is in the use of skip connections. Instead of using a straightforward concatenation method for skip connections, an Adaptive Feature Fusion module is used to control information flow in the encoder skip connections E_i to the corresponding decoder stage.
The AFF module uses channel attention to weight the feature maps. It obtains an attention map through a squeeze and excitation process done on the feature maps of the encoder. This helps to emphasize the relevant information of the lesion and suppress the redundant and noisy information of the background tissue. This serves the purpose of adaptive spatial weighting of the attention map mentioned in the abstract. The gated feature maps of the encoder are then concatenated to the upsampled feature maps from the previous stage.

3.5. Loss Function

To treat the strong foreground and background imbalance existing in foreground and background pixels in breast ultrasound images, the loss function of the neural network is optimized by combining loss function $\mathcal{L}$ combining the Dice loss $\mathcal{L}_{Dice}$ and binary cross-entropy loss $\mathcal{L}_{BCE}$ is used in equation 2:

$$\mathcal{L} = \mathcal{L}_{Dice} + \lambda \mathcal{L}_{BCE} \quad (2)$$

The Dice loss encourages maximal overlap between the predicted mask **P** and the ground truth mask **G** and is defined in equation 3:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum \mathbf{P} \odot \mathbf{G}}{\sum \mathbf{P} + \sum \mathbf{G}} \quad (3)$$

The BCE loss provides stable per-pixel probability calibration, complementing the region-based optimization of the Dice loss [48]. The weighting parameter λ is set to 1.0 in all experiments.

4. Experimental Setup

4.1. Datasets and Preprocessing

To test its performance, experiments were carried out on the public Breast Ultrasound Images (BUSI) dataset, which was first proposed by [49]. This involves a total of 780 ultrasound images, which were acquired from 600 patients and belong to 437 benign, 210 malignant, and 133 normal images. Since this proposed research is focused on lesion segmentation, the normal images were not used, and in total, 647 lesion images were obtained. Resizing was done to have a uniform resolution of 256 × 256 pixels for all images, while standardization of pixel intensities ranging from 0 to 1 ensured successful optimization in training.

4.2. Evaluation Protocol

A five-fold cross-validation approach was taken to properly and objectively assess the performance. The data were divided based on the patient ID to make sure that the images of a patient only went into a particular folder and no data leakage occurred. The four remaining folds were employed for the development of models through an internal 80:20 split during each cross-validation cycle. Performance metrics were determined through the averaging of the results of the five testing folds.

4.3. Implementation and Training Details

Implementation of SwinUNet: PyTorch was used to implement the proposed SwinUNet. In this network, the Swin-Tiny transformer encoder was initialized with ImageNet-1K pre-trained weights. The CNN decoder was trained from scratch. In the optimization process, the AdamW optimizer is utilized, whose initial learning rate is set to 1×10^{-4}, and weight decay is set to 1×10^{-2}. For decaying the learning rate polynomially, the power is 0.9. The number of epochs used in this process for training is 100, while the batch size is 16.

The composite Dice and Binary Cross-Entropy loss function described in Section 3.4, with $\lambda = 1.0$, was used. Online data augmentation, including random horizontal and vertical flipping, with random rotation within ±15°, was applied to enhance generalization and simulate realistic variations in lesion appearance.

4.4. Evaluation Metrics

Quantitative analysis was performed using Dice Similarity Coefficient (DSC), Intersection over Union (IoU), pixel-wise Accuracy (Acc), and the 95th percentile Hausdorff Distance (HD95). The DSC and IoU quantify the spatial overlap of the predicted segmentation mask P and the ground truth G, while Accuracy is calculated for the correctness of the pixel classification. HD95 quantifies boundary delineation accuracy by measuring the distance between predicted and reference contours, with lower values indicating superior boundary alignment. Moreover, The Hyperparameter configuration for SwinUNet in **Table 2. The evaluation metrics are defined as follows in equations 4-6:**

$$DSC = (2\times TP)/(2\times TP+FP+FN) \quad (4)$$

$$Acc = (TP+TN)/(TP+TN+FP+FN) \quad (5)$$

$$IoU = TP/(TP+FP+FN) \quad (6)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative pixels, respectively. Additional metrics, including Mean IoU, Weighted IoU, and Mean Boundary F-score (MeanBFScore), are reported in the Results section for comprehensive comparative analysis.

Table 2: Hyperparameter configuration for SwinUNet.

Hyperparameter	Value
Input Size	256 × 256
Batch Size	16
Epochs	100
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Weight Decay	1×10^{-2}

Hyperparameter	Value
Loss Function	$\mathcal{L}_{Dice} + \mathcal{L}_{BCE}$
Augmentation	Flip, Rotate (±15°)

5. Results and Discussion

5.1 Quantitative Results

Numerical segmentation performance comparisons of SwinUNet, as well as the benchmark models on the BUSI dataset, are provided in Table 3 and Figure 4, which are calculated with five-fold cross-validation. The performance of SwinUNet is better than that of the benchmark models in terms of Global Accuracy of 97.68%, Mean Accuracy of 94.88%, and Mean IoU of 89.93%. In contrast to conventional CNN models such as U-Net++ and ResUNet++, SwinUNet shows a significant improvement in terms of overlap measures, emphasizing the use of global contextual reasoning capability. In terms of hybrid methods that utilize CNN and Transformer architectures like TransFuse and SwinUNet, it is clear that our method has acquired a better Weighted IoU and Mean Boundary F-score. Despite SwinBTS++ showing a minor gain in Mean IoU, SwinUNet outperforms SwinBTS++ in terms of Global Accuracy, Weighted IoU, and Boundary F-score, which indicates a better compromise between region detection and border precision. These experiments thus verify that the proposed SwinUNet is capable of providing state-of-the-art region segmentation accuracy in breast ultrasound images.

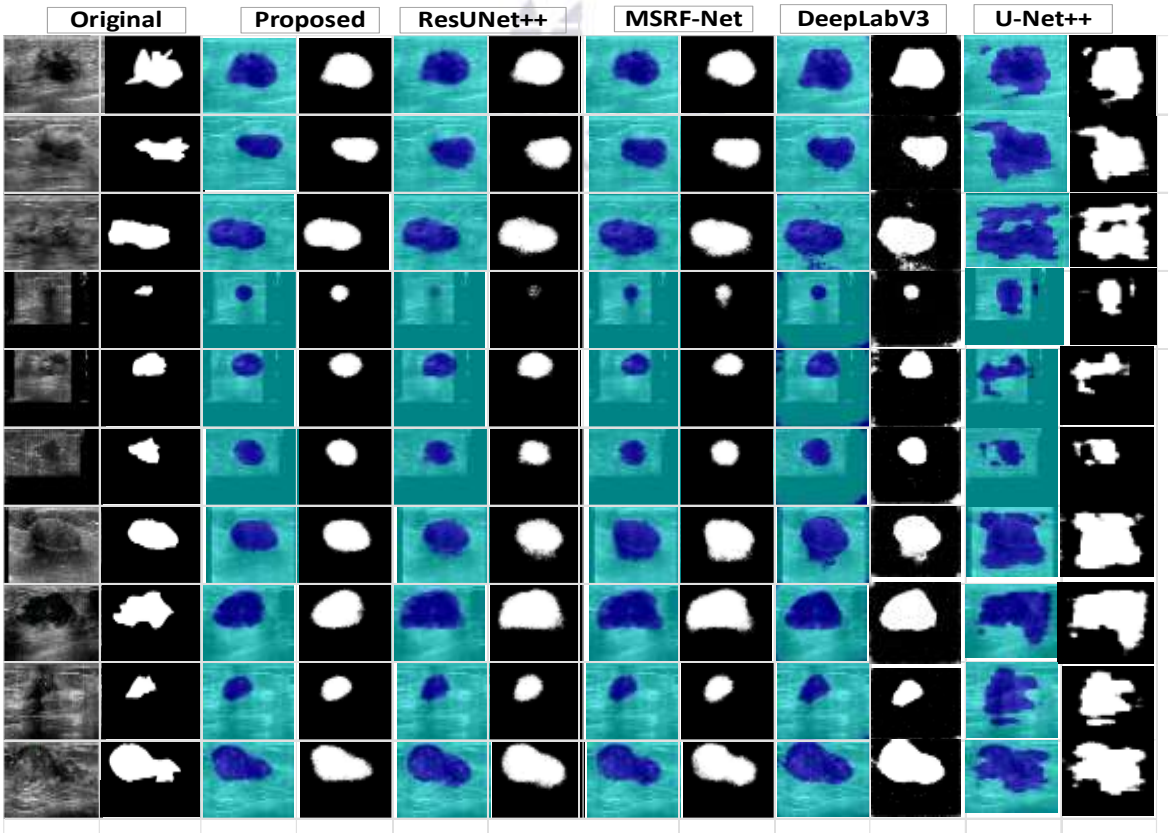

Figure 4: BUSI lesion segmentation analysis of the proposed SwinUNet framework with existing CNNs/ViTs.

Table 3. Quantitative comparison in terms of segmentation metrics.

Models	Regions.	DSC.	Acc.	IoU.	BF.Score.
Proposed SwinUNet					
SwinUNet	Lesion	80.15	91.205	82.465	71.937
	background	97.80	98.56	97.397	86.534
Existing ViTs/CNNs					
Inf_Net++	Lesion	69.43	83.205	69.465	71.937
	background	96.11	98.56	94.397	86.534
Res_UNet++	Lesion	78.95	90.26	81.857	73.733
	background	98.07	97.112	97.015	82.241
MSRF_Net	Lesion	75.08	88.302	74.524	62.671
	background	97.01	97.124	95.862	78.372
Deep_Lab-V3	Lesion	72.03	85.863	72.562	67.852
	background	96.91	94.889	94.872	71.746
U_Net++	Lesion	67.91	81.552	65.017	53.738
	background	95.65	92.994	92.811	66.201

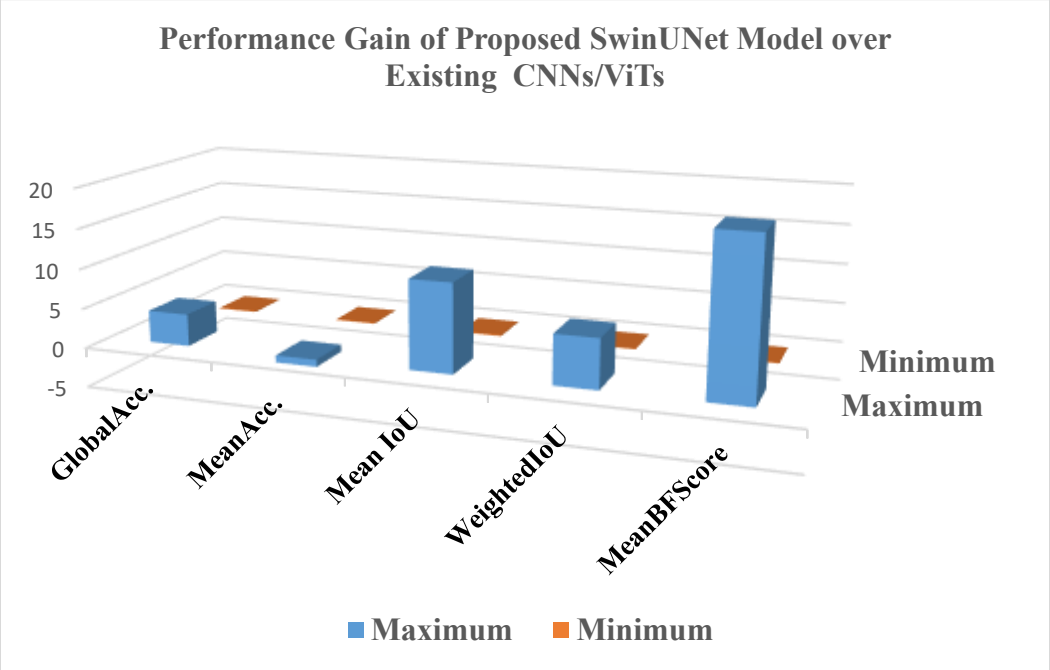

Figure 5: Performance Gain of the Proposed Model over Existing CNN/ViT's Architectures

5.2 Ablation and Comparative Analysis

The reason for the excellent performance of SwinUNet is the complementarity of the global and local feature representation mechanisms. The Swin Transformer encoder is able to capture the long-range context, which is crucial for lesion extent information, and the CNN decoder can reconstruct the output with pixel-level accuracy. The Adaptive Feature Fusion modules in the network are very helpful in giving priority to the channels that are related to the lesion. This proposed SwinUNet framework led to an increase in Mean IoU of 0.7891 to 0.8993, which is a 11.02% increase, along with an approximate increase in accuracy of 4.02%, as represented in Figure 5. The obtained experiments assert that Naïve Skip Connections inherit

unnecessary background info, and AFF successfully masks non-salient features and enhances similarity between encoder and decoder feature spaces. Furthermore, the comparative analysis emphasizes the limitations of purely CNN-based architectures, which are based on the idea of local receptive fields and are prone to issues of global consistency for the segmentation process. Although some other methods aim at the synthesis of transformer and CNN architectures, the approach based on AFF is more systematic and efficient for the integration of multi-scale features.

6. Conclusion

This paper proposes SwinUNet, a hybrid encoder-decoder network for breast lesion segmentation in ultrasound imaging. By combining a Swin Transformer encoder for contextual understanding with a CNN decoder for spatial locality, SwinUNet leverages both the strong capability of Transformers in modeling distant features and the pixel-level precise localization capability of CNNs. Perhaps the greatest strength of this network is its Adaptive Feature Fusion (AFF) module, which adjusts the reconstructed features in the skip connections to focus on lesion information and remove background noise particular to ultrasound imaging. Comprehensive experiments on the publicly available BUSI dataset using five-fold cross-validation on SwinUNet indicate a Global Accuracy of 97.68 and a Mean Intersection over Union of 89.93 for SwinUNet, surpassing both standard CNN approaches and previous hybrid CNN-Transformer networks in terms of performance metrics. The proposed framework succeeded in delineating the lesion effectively and reliably, even in tough cases with low contrast and highly irregular lesion morphology. These results indicate that SwinUNet has strong potential to be integrated into computer-aided diagnostic systems, thus enabling quantitative analysis, treatment planning, and standardized clinical assessment. Future studies will focus on multi-center validation to test generalizability and extension to 3D ultrasound volumes to increase clinical applicability.

Acknowledgment:
We thanks to Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences, Swat, for providing research environments.

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/CAAC.21660.

[2] Y. Zhou *et al.*, "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Med. Image Anal.*, vol. 70, p. 101918, May 2021, doi: 10.1016/J.MEDIA.2020.101918.

[3] A. Khan, S. H. Khan, M. Saif, A. Batool, A. Sohail, and M. W. Khan, "A Survey of Deep Learning Techniques for the Analysis of COVID-19 and their usability for Detecting Omicron," *J. Exp. Theor. Artif. Intell.*, vol. 36, no. 8, pp. 1779–1821, Apr. 2022, doi: 10.1080/0952813X.2023.2165724.

[4] B. Shareef, A. Vakanski, P. E. Freer, and M. Xian, "ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation," *Healthc. 2022, Vol. 10, Page 2262*, vol. 10, no. 11, p. 2262, Nov. 2022, doi: 10.3390/HEALTHCARE10112262.

[5] S. H. Khan *et al.*, "A new deep boosted CNN and ensemble learning based IoT malware detection," *Comput. Secur.*, vol. 133, p. 103385, Oct. 2023, doi: 10.1016/J.COSE.2023.103385.

[6] M. M. Zafar *et al.*, "Detection of tumour infiltrating lymphocytes in CD3 and CD8 stained histopathological images using a two-phase deep CNN," *Photodiagnosis Photodyn. Ther.*, vol. 37, Mar. 2022, doi: 10.1016/J.PDPDT.2021.102676.

[7] W. Weng and X. Zhu, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[8] Y. Hu *et al.*, "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," *Med. Phys.*, vol. 46, no. 1, pp. 215–228, Jan. 2019, doi: 10.1002/MP.13268,.

[9] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, Oct. 2020, Accessed: Apr. 29, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929

[10] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, Oct. 2020, Accessed: Jan. 05, 2025. [Online]. Available: https://arxiv.org/abs/2010.11929v2

[11] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, Mar. 2021, doi: 10.1109/ICCV48922.2021.00986.

[12] S. Hussain Khan and R. Iqbal, "RS-FME-SwinT: A Novel Feature Map Enhancement Framework Integrating Customized SwinT with Residual and Spatial CNN for Monkeypox Diagnosis".

[13] S. H. Khan, R. Iqbal, and S. Naz, "A Recent Survey of the Advancements in Deep Learning Techniques for Monkeypox Disease Detection," Nov. 2023, Accessed: Jan. 23, 2026. [Online]. Available: https://arxiv.org/pdf/2311.10754

[14] G. Xu, X. Zhang, X. He, and X. Wu, "LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14432 LNCS, pp. 42–53, Jul. 2021, doi: 10.1007/978-981-99-8543-2_4.

[15] M. Asam *et al.*, "Detection of Exceptional Malware Variants Using Deep Boosted Feature Spaces and Machine Learning," *Appl. Sci. 2021, Vol. 11, Page 10464*, vol. 11, no. 21, p. 10464, Nov. 2021, doi: 10.3390/APP112110464.

[16] M. H. Yap *et al.*, "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018, doi: 10.1109/JBHI.2017.2731873,.

[17] S. H. Khan, A. Sohail, A. Khan, and Y. S. Lee, "COVID-19 Detection in Chest X-ray Images Using a New Channel Boosted CNN," *Diagnostics*, vol. 12, no. 2, Feb. 2020, doi: 10.3390/DIAGNOSTICS12020267.

[18] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," Feb. 2021, Accessed: Apr. 29, 2025. [Online]. Available: https://arxiv.org/pdf/2102.04306

[19] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, 2018, doi: 10.1007/978-3-030-00889-5_1,.

[20] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," *Proc. - 2019 IEEE Int. Symp. Multimedia, ISM 2019*, pp. 225–230, Nov. 2019, doi: 10.1109/ISM46123.2019.00049.

[21] S. H. Khan, A. Sohail, M. M. Zafar, and A. Khan, "Coronavirus disease analysis using chest X-ray images and a novel deep convolutional neural network," *Photodiagnosis Photodyn. Ther.*, vol. 35, p. 102473, Sep. 2021, doi: 10.1016/J.PDPDT.2021.102473.

[22] M. Mumtaz Zahoor and S. Hussain Khan, "Brain tumor MRI Classification using a Novel Deep Residual and Regional CNN".

[23] A. Khan *et al.*, "A Recent Survey of Vision Transformers for Medical Image Segmentation," Dec. 2023, Accessed: Oct. 11, 2024. [Online]. Available: https://arxiv.org/abs/2312.00634v2

[24] A. Vakanski, M. Xian, and P. E. Freer, "Attention-Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images," *Ultrasound Med. Biol.*, vol. 46, no. 10,

pp. 2819–2833, Oct. 2020, doi: 10.1016/J.ULTRASMEDBIO.2020.06.015.

[25] S. H. Khan *et al.*, "COVID-19 detection in chest X-ray images using deep boosted hybrid learning," *Comput. Biol. Med.*, vol. 137, p. 104816, Oct. 2021, doi: 10.1016/J.COMPBIOMED.2021.104816.

[26] H. Cao *et al.*, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13803 LNCS, pp. 205–218, May 2021, doi: 10.1007/978-3-031-25066-8_9.

[27] S. Hussain Khan, A. Khan, Y. Soo Lee, M. Hassan, and W. Kyo jeong, "Segmentation of Shoulder Muscle MRI Using a New Region and Edge based Deep Auto-Encoder".

[28] S. H. Khan, N. S. Shah, R. Nuzhat, A. Majid, H. Alquhayz, and A. Khan, "Malaria parasite classification framework using a novel channel squeezed and boosted CNN," *Reprod. Syst. Sex. Disord.*, vol. 71, no. 5, pp. 271–282, Oct. 2022, doi: 10.1093/JMICRO/DFAC027.

[29] M. Abdullah, F. berhe Abrha, B. Kedir, and T. Tamirat Tagesse, "A Hybrid Deep Learning CNN model for COVID-19 detection from chest X-rays," *Heliyon*, vol. 10, no. 5, p. e26938, Mar. 2024, doi: 10.1016/J.HELIYON.2024.E26938.

[30] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12901 LNCS, pp. 14–24, Feb. 2021, doi: 10.1007/978-3-030-87193-2_2.

[31] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[32] N. Eisemann *et al.*, "Nationwide real-world implementation of AI for cancer detection in population-based mammography screening," *Nat. Med. 2025 313*, vol. 31, no. 3, pp. 917–924, Jan. 2025, doi: 10.1038/s41591-024-03408-6.

[33] M. M. S. Abdullah, H. A. Al-Lohedan, and A. M. Atta, "Expression of concern: Novel magnetic iron oxide nanoparticles coated with sulfonated asphaltene as crude oil spill collectors," *RSC Adv.*, vol. 15, no. 15, pp. 11581–11581, Apr. 2025, doi: 10.1039/D5RA90044A.

[34] X. Chen *et al.*, "Small extracellular vesicles from young plasma reverse age-related functional declines by improving mitochondrial energy metabolism," *Nat. Aging*, vol. 4, no. 6, pp. 814–838, Jun. 2024, doi: 10.1038/S43587-024-00612-4;SUBJMETA.

[35] W. Zhang *et al.*, "Enhancing CRISPR prime editing by reducing misfolded pegRNA interactions.," *Elife*, vol. 12, Jun. 2024, doi: 10.7554/eLife.90948.

[36] C. Zhou *et al.*, "LIMA: Less Is More for Alignment," *Adv. Neural Inf. Process. Syst.*, vol. 36, May 2023, Accessed: Jan. 23, 2026. [Online]. Available: https://arxiv.org/pdf/2305.11206

[37] L. Cao *et al.*, "Design of protein-binding proteins from the target structure alone," *Nat. 2022 6057910*, vol. 605, no. 7910, pp. 551–560, Mar. 2022, doi: 10.1038/s41586-022-04654-9.

[38] B. Zhou, P. Perel, G. A. Mensah, and M. Ezzati, "Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension," *Nat. Rev. Cardiol.*, vol. 18, no. 11, pp. 785–802, Nov. 2021, doi: 10.1038/S41569-021-00559-8.

[39] B. Qin *et al.*, "Cell position fates and collective fountain flow in bacterial biofilms revealed by light-sheet microscopy," *Science (80-.).*, vol. 369, no. 6499, pp. 71–77, Jul. 2020, doi: 10.1126/SCIENCE.ABB8501;PAGE:STRING:ARTICLE/CHAPTER.

[40] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019, doi: 10.1148/RADIOL.2019182716.

[41] Z. Zhou, G. Bi, and J. M. Zhou, "Luciferase Complementation Assay for Protein-Protein Interactions in Plants," *Curr. Protoc. plant Biol.*, vol. 3, no. 1, pp. 42–50, Mar. 2018, doi: 10.1002/CPPB.20066.

[42] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," Apr. 2018, Accessed: Jul. 05, 2024. [Online]. Available: https://arxiv.org/abs/1804.03999v3

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[44] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," *Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision, WACV 2022*, pp. 1748–1758, Mar. 2021, doi: 10.1109/WACV51458.2022.00181.

[45] S. Naz and S. H. Khan, "Residual-SwinCA-Net: A Channel-Aware Integrated Residual CNN-Swin Transformer for Malignant Lesion Segmentation in BUSI," Dec. 2025, Accessed: Jan. 23, 2026. [Online]. Available: https://arxiv.org/pdf/2512.08243

[46] M. A. Rahman, "HyFormer-Net: A Synergistic CNN-Transformer with Interpretable Multi-Scale Fusion for Breast Lesion Segmentation and Classification in Ultrasound Images," Nov. 2025, Accessed: Jan. 23, 2026. [Online]. Available: https://arxiv.org/pdf/2511.01013

[47] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," Jul. 2016, Accessed: Nov. 07, 2024. [Online]. Available: https://arxiv.org/abs/1607.08022v3

[48] H. Qu *et al.*, "Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images," *IEEE Trans. Med. Imaging*, vol. XX, p. 1, 2020, doi: 10.1109/TMI.2020.3002244.

[49] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/J.DIB.2019.104863.