applied sciences

MDPI

Article

Transformer-Based Explainable Model for Breast Cancer Lesion Segmentation

Huina Wang [1], Lan Wei [2], Bo Liu [3], Jianqiang Li [1], Jinshu Li [1], Juan Fang [1] and Catherine Mooney [2,*]

[1] School of Computer Science, Beijing University of Technology, Beijing 100124, China; huinawang@emails.bjut.edu.cn (H.W.); lijianqiang@bjut.edu.cn (J.L.); lijs5678@163.com (J.L.); fangjuan@bjut.edu.cn (J.F.)
[2] FutureNeuro Research Ireland Centre, School of Computer Science, University College Dublin, Belfield, D04 V1W8 Dublin, Ireland; lan.wei@ucd.ie
[3] School of Mathematical and Computational Sciences, Massey University, Palmerston North 0632, New Zealand; b.liu@massey.ac.nz
* Correspondence: catherine.mooney@ucd.ie

Abstract: Breast cancer is one of the most prevalent cancers among women, with early detection playing a critical role in improving survival rates. This study introduces a novel transformer-based explainable model for breast cancer lesion segmentation (TEBLS), aimed at enhancing the accuracy and interpretability of breast cancer lesion segmentation in medical imaging. TEBLS integrates a multi-scale information fusion approach with a hierarchical vision transformer, capturing both local and global features by leveraging the self-attention mechanism. This model addresses the limitations of existing segmentation methods, such as the inability to effectively capture long-range dependencies and fine-grained semantic information. Additionally, TEBLS incorporates visualization techniques to provide insights into the segmentation process, enhancing the model's interpretability for clinical use. Experiments demonstrate that TEBLS outperforms traditional and existing deep learning-based methods in segmenting complex breast cancer lesions with variations in size, shape, and texture, achieving a mean DSC of 81.86% and a mean AUC of 97.72% on the CBIS-DDSM test set. Our model not only improves segmentation accuracy but also offers a more explainable framework, which has the potential to be used in clinical settings.

Keywords: breast cancer lesion segmentation; transformer; explainable model

Academic Editors: Alexander Gegov, Raheleh Jafari and Farzad Arabikhan

Received: 20 December 2024
Revised: 18 January 2025
Accepted: 21 January 2025
Published: 27 January 2025

Citation: Wang, H.; Wei, L.; Liu, B.; Li, J.; Li, J.; Fang, J.; Mooney, C. Transformer-Based Explainable Model for Breast Cancer Lesion Segmentation. *Appl. Sci.* **2025**, *15*, 1295. https://doi.org/10.3390/app15031295

1. Introduction

Breast cancer is one of the most common cancers affecting women, with high rates of both incidence and mortality [1–3]. The early detection of breast cancer can effectively reduce mortality rates and improve patients' quality of life. Imaging examinations are essential for breast cancer screening and diagnosis and include mammography, breast ultrasound, and breast magnetic resonance imaging (MRI). These imaging techniques enable clinicians to detect abnormalities in breast tissue, such as masses or calcifications, supporting accurate breast cancer diagnosis. With advancements in deep learning within the medical field, their use in the automated identification and detection of breast cancer lesions has become a major focus in medical image processing [4–6].

Traditional breast cancer segmentation methods, such as threshold segmentation [7], region growing [8], and superpixel segmentation [9], are widely used in image segmentation [10,11]. These techniques assist physicians in interpreting images more effectively,

identifying potential abnormal regions, and providing more precise diagnostic information. However, traditional breast cancer detection methods are semi-automatic and rely heavily on manual intervention by radiologists, resulting in segmentation outputs with high interpretability but limited generalization capability. To achieve automated segmentation of lesion areas, CNN-based image semantic segmentation models are commonly applied in this field, typically using a local receptive field to capture image features [12–15]. Although these methods can extract local features to some extent, they lack the ability to capture global semantic information.

Additionally, the fixed sizes of convolutional kernels hinder CNN-based models from effectively interpreting complex semantic structures in images, restricting their segmentation accuracy in clinical settings. The rigid network architectures and parameter configurations of these models further impede their ability to model image semantics across different scales and complexities, diminishing their capacity to capture fine details in diverse and intricate medical images. In medical imaging, breast cancer lesions frequently present considerable variations in size, shape, and texture, which create challenges for CNN-based segmentation algorithms in effectively capturing both fine details and global semantic information.

To address these issues, recent studies have explored transformer-based image segmentation algorithms for lesion segmentation [16,17]. These models use convolutional layers to extract local features while leveraging self-attention to capture long-range feature correlations across regions, achieving multi-scale feature fusion along the spatial dimension. However, these models often overlook global context priors within multi-scale semantics and the fine-grained semantic information within channel dimensions. While deep learning-based image segmentation models exhibit strong feature-handling capabilities and potential applications in the medical field, their internal mechanisms remain unclear, which may lead to risks for clinical use. Understanding and explaining these models is essential to clarify their behaviours, limitations, and broader societal impacts [18].

Previous work [19] used a taxonomy of interpretability techniques and provided a structured summary based on transformer language model approaches. However, given the end-to-end nature and high nonlinear complexity of these models, further investigation is necessary to validate the interpretability of breast cancer image segmentation models utilizing imaging data. Therefore, we introduce a novel image segmentation method, a transformer-based explainable model for breast cancer lesion segmentation (TEBLS). This method uses a multi-scale information fusion approach centred on a swin transformer, allowing for the effective integration of spatial information across various scales. Additionally, it incorporates visualization techniques to monitor and track the segmentation results.

The main contributions of this study are as follows:

- A novel image segmentation method, TEBLS, for breast cancer lesion segmentation. TEBLS combines a multi-scale information fusion approach with a swin transformer, which allows for the effective integration of spatial information across multiple scales.
- TEBLS employs a self-attention mechanism within the transformer architecture to enhance its ability to capture both long-range dependencies and local features, addressing the limitations of traditional CNN-based methods in segmenting complex medical images.
- TEBLS incorporates explainability methods to provide insight into the segmentation results, improving interpretability and offering clinicians the ability to better understand and trust the model's decisions, which is crucial for clinical settings.
- This study overcomes the shortcomings of traditional deep learning models, such as the inability to capture fine-grained semantic information and global context

within multi-scale semantics, thereby advancing the state of the art in breast cancer lesion segmentation.
- Extensive experiments confirm the effectiveness of the proposed TEBLS over other state-of-the-art methods.

2. Related Work

This section categorizes image segmentation methods into three main types: traditional image segmentation methods, CNN-based image segmentation methods, and transformer-based image segmentation methods.

2.1. Traditional Image Segmentation Methods

Traditional image segmentation methods rely on pixel-level features to extract regions of interest (ROIs) from images. These methods have been widely used in tasks requiring rapid segmentation or unsupervised processing. Xing et al. [20] introduced a multi-threshold image segmentation method based on an improved Emperor Penguin Optimization (EPO) algorithm. This approach leverages EPO to identify optimal multi-level thresholds for color images and enhances its search capabilities using Gaussian mutation, Lévy flight, and reverse learning. Similarly, Mittal et al. [21] proposed a two-dimensional (2D) histogram-based segmentation method, using a novel non-local means 2D histogram and an exponential Kbest gravitational search algorithm to determine optimal thresholds while redefining 2D Rényi entropy.

Edge-detection-based segmentation techniques, which use gradient changes or other edge information, are commonly employed to delineate target boundaries. Cigla et al. [22] developed a graph theory-based color image segmentation algorithm. Unlike traditional pixel-based methods, this algorithm represents over-segmented regions as nodes, reducing complexity. Nodes are linked based on intensity similarity, encouraging the merging of regions with numerous shared links. This method demonstrates superior execution speed and segmentation quality compared to regularization-cut techniques. Roy et al. [11] proposed an unsupervised edge detection method that calculates local standard deviation to effectively extract nuclei edges in pathological images. Additionally, Tasli et al. [23] presented an efficient superpixel (SP) and super voxel (SV) extraction method that improves computational efficiency by employing boundary adaptation techniques instead of processing entire regions.

Traditional segmentation methods utilize principles of digital image processing and mathematical techniques. While they are computationally simple and fast, their ability to achieve detailed segmentation is often limited compared to more advanced approaches.

2.2. CNN-Based Image Segmentation Methods

The CNN-based segmentation network introduces several modifications to the standard CNN architecture, including replacing the final two fully connected layers with convolutional layers, adopting deeper architectures (such as the visual geometry group (VGG) network [24]), and employing encoder–decoder structures (such as U-Net [14]). Long et al. [25] were the first to propose adapting classification networks into fully convolutional networks (FCNs). By fine-tuning these networks, they transferred learned representations to segmentation tasks, achieving exceptional results. Building on the FCN framework, Badrinarayanan et al. [26] developed the SegNet model, which features a symmetric encoder–decoder structure designed for semantic segmentation and supports end-to-end training.

Ronneberger et al. [14] introduced the U-Net network for biomedical image segmentation, characterized by its U-shaped architecture and skip connections. Its remark-

able performance has made it a widely adopted solution in medical image segmentation. Zhao et al. [27] proposed the Pyramid Scene Parsing Network (PSPNet), which integrates contextual information from different regions to derive optimal global context representations. Dollár et al. [28] presented Mask R-CNN, an object instance segmentation framework extending Faster R-CNN [29] by incorporating a branch for predicting object masks. This approach effectively detects objects while generating high-quality segmentation masks for individual instances.

Overall, CNN-based segmentation methods use the strong feature extraction and representation capabilities of CNN architectures, leading to significant progress in image segmentation. These methods have consistently demonstrated higher segmentation accuracy compared to traditional approaches.

2.3. Transformer-Based Image Segmentation Methods

The transformer, initially developed for sequence modeling and transformation tasks, is highly effective at capturing long-range dependencies in data. Its success in natural language processing (NLP) has motivated its adoption in computer vision. Liu et al. [16] proposed the swin transformer, a versatile backbone for vision tasks. This model generates hierarchical feature maps and achieves linear computational complexity with respect to image size. Building on this work, Cao et al. [30] further introduced Swin-Unet, a U-shaped architecture comprising an encoder, a bottleneck, a decoder, and skip connections. Swin-Unet uses the swin transformer for feature extraction and employs skip connections during upsampling to integrate encoder features with decoder features, enabling accurate image segmentation predictions.

Stéphane et al. [31] developed the ConViT model, which integrates the inductive bias of CNNs into the transformer to better focus on local positional and content information. Huang et al. [32] introduced ScaleFormer, a model that combines CNN-derived local features with transformer-based global representations at multiple scales. This approach emphasizes detailed spatial information (contextual cues) alongside long-range dependencies (positional cues), effectively addressing the challenges of scale variation in image segmentation.

In the domain of volumetric medical image segmentation, Zhou et al. [33] proposed nnFormer, which integrates interleaved convolutions and self-attention mechanisms to learn volumetric representations. The model employs skip attention to replace traditional concatenation or summation operations in U-Net-style architectures, ensuring efficient local and global feature extraction. Similarly, Chen et al. [6] introduced TransUNet, which incorporates a transformer-encoder to capture long-range dependencies critical for segmentation tasks.

Ji et al. [34] proposed MCTrans, a model embedding multi-scale convolutional features into token sequences. By leveraging a self-attention mechanism, MCTrans facilitates cross-scale pixel-level context modeling and employs learnable proxy embeddings to capture class dependencies. The combination of self-attention and cross-attention mechanisms allows the model to model semantic relationships and enhance feature representations effectively.

Transformer-based segmentation models use the transformer's strength in capturing global relationships between distant pixels, making them highly effective at extracting global image features. However, their intricate architectures often face challenges in efficiently capturing fine spatial details. To address this, we propose using dense skip connections to repeatedly fuse multi-scale features, thereby capturing more detailed information.

3. Methodology

In this section, we introduce the TEBLS model proposed in this paper, experimental dataset, and experimental setup.

3.1. Data Preparation

In order to compare and demonstrate the performance of the proposed segmentation model in the task of breast cancer lesion segmentation, we used the CBIS-DDSM (Curated Breast Imaging Subset of DDSM) for training the network. The CBIS-DDSM is a breast X-ray imaging dataset designed for research in computer-aided detection and diagnosis, comprising a database of 2620 mammogram case data from multiple hospitals [35]. From this dataset, we selected 574 images with full-size masks and prominent lesion features as the experimental data for this study. We used 80% of the dataset to train the model and the other 20% for testing. To enhance data diversity, we applied random flipping and rotation techniques to the images and masks in the training dataset [36] and then resized the images to a 224 × 224 size, which were used as input and ground truth labels during model training.

3.2. Model Network Structure

Figure 1 shows the model's framework, and Figure 2 illustrates its flowchart. The input image set is denoted as $x \in R^{H \times W \times 3}$, where R represents all images in the dataset, and any individual image within this set is denoted as x. Each image has a height H, width W, and 3 color channels. To convert the image information into sequential embeddings, the image set is first input into a patch partition layer. This layer divides the medical image pixels into non-overlapping patches of size $N \times N$, transforming the feature dimensions into $N \times N \times 3$ for each patch.

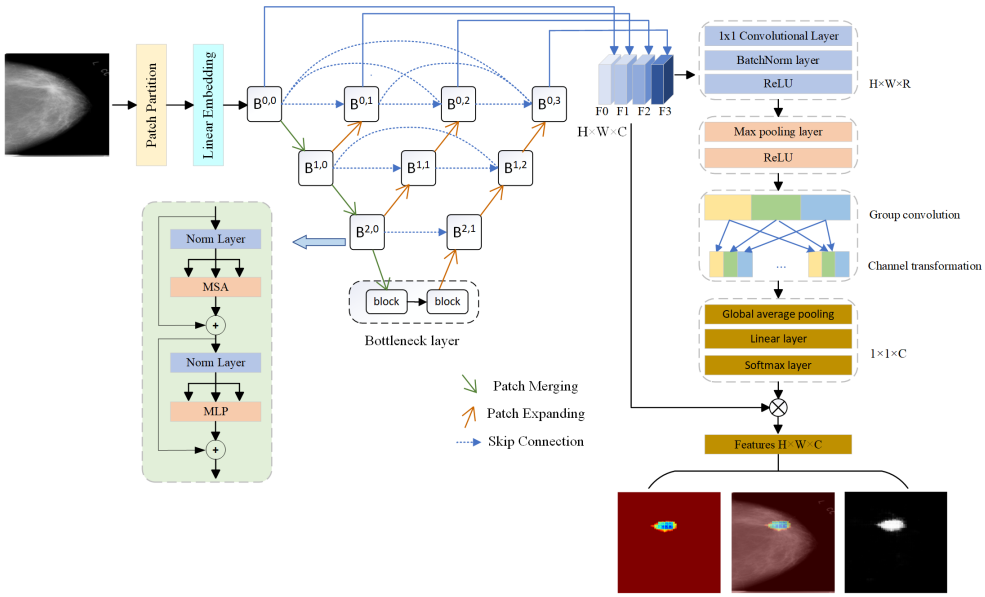

Figure 1. Overview of the framework. The model begins with patch partitioning and linear embedding to transform input images into sequential data. The encoder consists of multiple swin transformer blocks, patch merging layers, and skip connections for feature fusion. The bottleneck layer processes the encoded features with Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) modules. The decoder includes patch expansion, normalization, and linear projection layers to produce high-resolution segmentation outputs. The output is processed through global average pooling and a softmax layer for pixel-level classification, with Grad-CAM used to visualize the results.

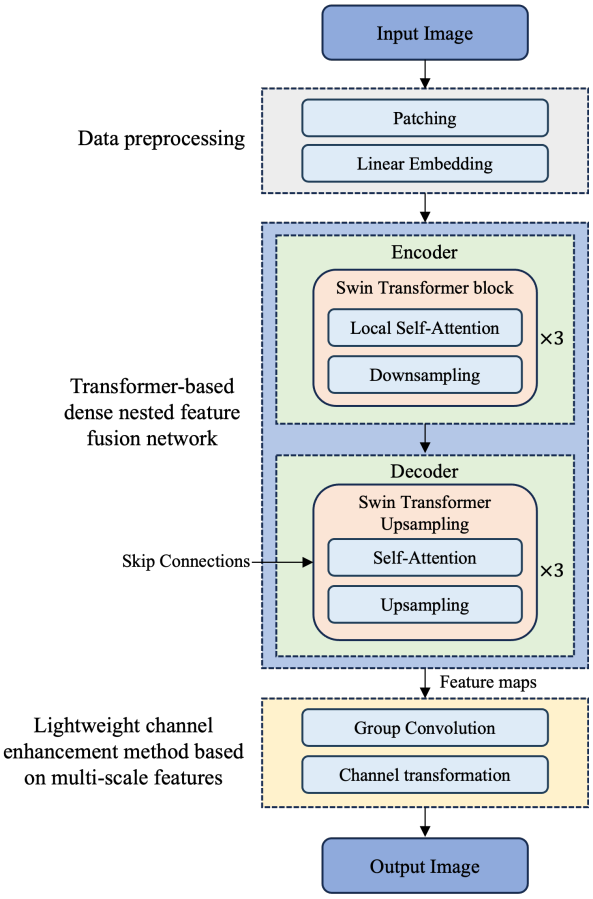

Figure 2. The flowchart illustrates the main structure and data flow of TEBLS. The rounded rectangles represent the input and output modules of the model, while the gray rectangles indicate the input data preprocessing module. The blue rectangles represent the transformer-based dense nested feature fusion network module proposed in this paper, with the green rectangular frames representing the encoder and decoder parts of this module, which consist of three swin transformer blocks and swin transformer upsampling, respectively. The yellow rectangles represent the lightweight channel enhancement method based on the multi-scale features module proposed in this paper, which includes group convolution and channel transformation.

$$x \in R^{H\times W\times 3} \rightarrow x_p \in R^{(H/N)\times(W/N)\times(N\times N\times 3)} \tag{1}$$

The linear embedding layer performs feature mapping on these patches, projecting their feature dimensions into an arbitrary dimension, C, resulting in feature maps. This operation can be expressed as

$$x_p \in R^{(H/N)\times(W/N)\times(N\times N\times 3)} \rightarrow x_p^{'} \in R^{(H/N)\times(W/N)\times C} \tag{2}$$

where $R^{(H/N)\times(W/N)\times(N\times N\times 3)}$ represents the set of all patches, where any individual patch x_p in this set has a pixel height of $\frac{H}{N}$, a pixel width of $\frac{W}{N}$, and $N \times N \times 3$ channels. $R^{(H/N)\times(W/N)\times C}$ denotes the set of images processed by the linear embedding layer, where any image in this set has a pixel height of $\frac{H}{p}$, a pixel width of $\frac{W}{p}$, and C channels.

The TEBLS model extracts image features through multiple downsampling and upsampling processes and fuses features from multiple levels. The encoder structure of the TEBLS model is as follows: the input data $R^{(H/4)\times(W/4)\times C}$ are processed through three stages, and the bottleneck layer finally outputs $R^{(H/32)\times(W/32)\times(C/8)}$. Specifically, the input data first pass through the first swin block, followed by a patch merging layer, then enter the

second swin block, followed by another patch merging layer, and finally proceed through the third swin block and another patch merging layer before being input into the bottleneck layer. Each swin block consists of a single swin transformer module.

The patch merging layer selects elements at regular intervals in the row and column directions, concatenates them into a tensor, and reshapes them. Channel normalization and a fully connected layer are then applied to adjust the channel dimensions. Through this process, feature resolution gradually decreases while feature dimensions increase, achieving downsampling.

The bottleneck layer consists of two swin transformer blocks connected sequentially. In the bottleneck layer, the first swin transformer block connects to the output of the encoder's final patch merging layer through channel normalization and a fully connected layer. The second swin transformer block connects to the input of the lowest-level patch expansion layer in the upsampling stage. This forms a connection between the second swin transformer block of the bottleneck layer and the lowest-level patch expansion layer in the upsampling stage.

Feature fusion in the TEBLS model is achieved through skip connections, which fuse the output of a swin block from one dense block with the corresponding upsampling output in the next dense block. The specific process includes first expanding the channel dimensions through a fully connected layer in the patch expansion layer and then using a rearrange operation to reshape adjacent feature map dimensions into larger feature maps and finally applying LayerNorm for channel normalization to complete the upsampling operation.

The decoder structure of the TEBLS model takes the feature maps output from the encoder and processes them through the patch expansion layer and then through a linear projection layer to generate pixel-level image predictions. Specifically, the patch expansion layer upsamples the feature maps to restore their resolution to the input resolution $W \times H$. This patch expansion layer first uses a fully connected layer to increase the channel dimensions, setting the output tensor's dimensions to 16 times that of the input. Next, the rearrange operation reshapes adjacent feature map dimensions into larger feature maps, and LayerNorm normalizes the channels to complete the upsampling. Finally, the upsampled features undergo linear projection to generate pixel-level segmentation predictions with the dimensions $W \times H \times$ Class, where Class represents the number of pixel-level classification categories and H and W represent the height and width of the image, respectively.

3.3. Transformer-Based Dense Nested Feature Fusion Network

Based on the Swin-Unet [16] and Unet++ [37] models, we propose a transformer-based densely nested feature fusion image model for medical image segmentation (Swin-Unet++). This model is built with an encoder, a bottleneck, a decoder, and skip connections based on the swin transformer module. $B^{i,j}$ represents the output of feature learning through the swin block layer, where i indexes the downsampling layers of the encoder, and the swin block layers are indexed along the skip connection path. Each swin block layer contains two swin transformer blocks. The upsampling output of $B^{0,0}$ and $B^{1,0}$ is fused to obtain $B^{0,1}$; then, $B^{1,0}$ is fused with the upsampling output of $B^{2,0}$ to obtain $B^{1,1}$. The upsampling output of $B^{1,1}$ is fused with $B^{0,0}$ and $B^{0,1}$ through skip connections to obtain $B^{0,2}$. Similarly, through layer-wise upsampling and feature fusion, $B^{2,1}$ and $B^{1,2}$ are sequentially obtained, ultimately resulting in the final output, which is the fusion of the upsampling output of $B^{0,3}$; $B^{0,3}$ is the fusion of the upsampling output of $B^{1,2}$, $B^{0,0}$, $B^{0,1}$, and $B^{0,2}$. Table 1 below shows the detailed output information corresponding to each swin block layer unit in the transformer-based densely nested feature fusion network.

Table 1. The output dimensions of the swin block unit, detailing image sizes and corresponding channel numbers across different layers.

	Image Size	Number of Channels
$B^{0,0}$	56	C
$B^{1,0}$	28	2 C
$B^{0,1}$	56	C
$B^{2,0}$	14	4 C
$B^{1,1}$	28	2 C
Bottleneck layer	7	8 C
$B^{0,2}$	56	C
$B^{2,1}$	14	4 C
$B^{1,2}$	28	2 C
$B^{0,3}$	56	C

3.4. Lightweight Channel Enhancement Method Based on Multi-Scale Features

To reduce the complexity of the model and achieve multi-channel feature fusion, we propose a lightweight channel enhancement method based on multi-scale features. The image feature maps obtained by the TEBLS model at different levels are denoted as F_0, F_1, F_2, F_3. These features are concatenated along the channel dimension to obtain a mixed feature map with a channel size of $4C$. After a linear transformation, the number of channels is reduced to C. To reduce the dimensionality of the channels for subsequent transformations, we use a 1×1 convolution layer, a batch normalization layer, and a ReLU activation function to transform the channels from C to R.

Moreover, a max. pooling layer is applied to reduce the size of the feature map while extracting the most prominent features. To further reduce the complexity and number of parameters, we use grouped convolutions to split the channels into three groups, applying convolution operations to each group separately. Furthermore, in order to capture more detailed features, we perform channel transformations on the features of each group through three stages: stage 2, stage 3, and stage 4. The number of channels increases from R to 240 in stage 2, to 480 in stage 3, and to 960 in stage 4. During this process, the image size progressively decreases: in stage 2, the image size is 4×4; in stage 3, the image size becomes 2×2; and in stage 4, the image size is reduced to 1×1.

A global average pooling operation is applied to obtain a channel feature representation of the dimension 960. Subsequently, global pooling is applied to the entire feature map. After the global pooling layer, the feature map of the size $H \times W \times R$ is reduced to $1 \times 1 \times R$. To ensure that the output channel number matches the input channel number, we use a linear layer to transform the output channels back to C. Finally, a softmax layer outputs the weights of the channel feature representations.

To effectively fuse the rich semantic information contained in the original input features and compensate for any potentially ignored semantic information during the channel transformation process, we element-wise multiply the obtained feature weights with the original input feature map. Through these operations, the resolution of the feature map is restored to the input resolution $W \times H$. These features are then linearly projected to obtain pixel-level classification prediction results.

The specific calculation formula for channel transformation is as follows: First, the number of channels C is transformed using the first 1×1 convolution layer to reduce the number of parameters. Then, a grouped convolution is applied to the obtained features, dividing them into groups. After grouping, a channel transformation operation is performed on the features to enhance the information modeling between the channels.

This transformation allows for the efficient extraction and modeling of features across different channel groups, leading to improved feature representation while minimizing computational complexity.

The computation process formula is as follows.

$$X = Fc(G(Fcf(Fgc(Fc2r(C, Wc), Wg)), Wr)) \tag{3}$$

where we have the following:

- $Fcf()$ represents the channel transformation operation.
- $Fc()$, $Fc2r()$, and $Fgc()$ correspond to the linear layer, point-wise grouped convolution, and grouped convolution operations, respectively.
- Wr, Wc, and Wg are the respective weights for these operations.
- G represents the global average pooling operation, which results in a channel feature representation of the dimension C.

To implement lightweight channel transformation, the linear layer is used to transform the number of channels. This approach reduces computational complexity while maintaining the ability to model the features effectively. For the input data $X = [x_1, x_2, \ldots, x_c] \in R_H \times W \times C$, the features of the c channel are expressed as

$$Z_c = G(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{4}$$

where $G()$ represents global average pooling and, finally, the softmax function is used to calculate the correlation weights between these channel feature representations, that is,

$$w = \tau(z) = \frac{e_i^z}{\sum_{i=1}^{c} e_i^z} \tag{5}$$

where τ indicates the softmax function and the relevant weight of the dimension C. These relevant weights and corresponding features are multiplied in the channel dimension to obtain the final output of the channel information enhancement module:

$$\widehat{x} = w_c \cdot x_c \tag{6}$$

where w_c and x_c represent the relevant weights and corresponding feature channels of the c channel.

3.5. Experimental Setup

This experiment was conducted using the PyTorch 1.7.0 framework on Windows, based on Python 3.7.16. The experimental platform was configured with an NVIDIA GeForce GTX 1650 processor (Santa Clara, CA, USA), featuring 4.295 GB of GPU memory for computations. Figure 3 shows that the loss and score function stabilized when epoch = 150; we set the number of epochs to 150 during the model training process. The learning rate LR_t in the iteration t is given by the following formula, where LR_{init} represents the initial learning rate of 0.05, t represents the current iteration number, and T_{max} represents the maximum number of iterations during the training process. The specific experimental settings are summarized in Table 2.

$$LR_t = LR_{\text{init}} \left(1 - \frac{t}{T_{\max}}\right)^{0.9} \tag{7}$$

Table 2. Configuration details for the experimental setup, including hardware specifications, software framework, and model training parameters.

Configuration	Settings	Configuration	Settings
Hardware	NVIDIA GeForce GTX1650	Image Size	244 × 244
Framework	PyTorch	Patch Size	4
GPU	4.295G	CUDA Version	11.7
Optimizer	SGD Optimizer	Momentum	0.9
Batch Size	24	Weight Decay	0.0001

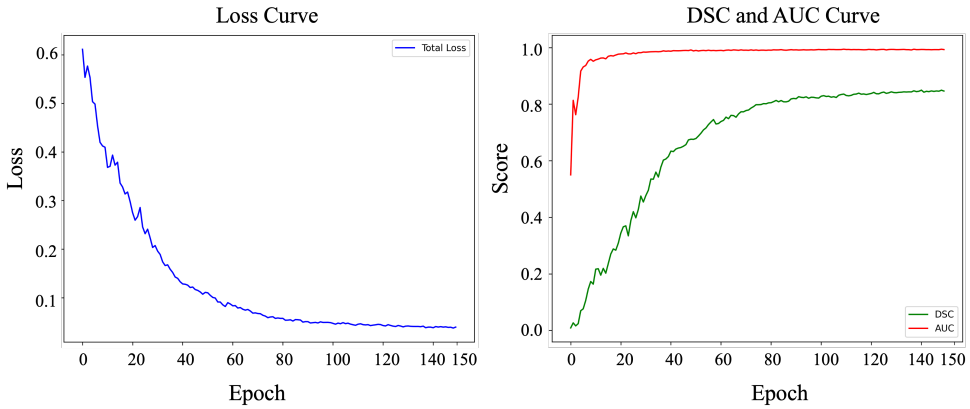

Figure 3. The loss and score curve during the model training process indicates that the loss function stabilized when epoch = 150.

3.6. Gradient-Weighted Class Activation Mapping

In this study, transformer models were used for image segmentation and classification. Due to the difficulty in understanding the internal workings of transformers, Gradient-weighted Class Activation Mapping (Grad-CAM) [38] was used to provide a way to visualize which regions of an image were most influential in the model's decision-making process. This approach involves calculating the gradient of the last convolutional layer with respect to the model's feature maps for a specific class [38]. The gradients are then averaged to generate a matrix that highlights the areas of the image most relevant to the model for identifying features of the specified class. By overlaying class activation maps on the input image, it highlights the areas the model focuses on during segmentation, offering visual explanations. This is particularly valuable in medical imaging, where understanding the model's focus is critical for clinical validation.

3.7. Evaluation Criterion

To assess the segmentation performance of the algorithm proposed in this paper, we employed the Dice Similarity Coefficient (DSC), the area under the receiver operating characteristic curve (AUC), and sensitivity.

The DSC is an indicator that measures the overlap between the model's segmentation result and the ground truth, with values ranging from 0 to 1. A higher value signifies greater similarity and, consequently, higher accuracy. The formula is as follows, where Prediction represents the total sum of pixels in the predicted region, and Ground Truth represents the total sum of pixels in the ground truth region.

$$DSC = \frac{2 \times |Prediction \cap GroundTruth|}{Prediction + GroundTruth} \tag{8}$$

The AUC is a widely used metric for evaluating the performance of binary classification models. The ROC curve plots the true positive rate (recall) on the vertical axis against the false positive rate on the horizontal axis. The AUC represents the area beneath the

receiver operating characteristic (ROC) curve, with values that range from 0 to 1. In image segmentation, AUC is typically used to assess a model's ability to distinguish between the target region and the background based on pixel classification.

Sensitivity measures the proportion of actual positive instances correctly identified by a model. A higher sensitivity indicates better performance in identifying the positive class and is calculated as follows.

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

where we have the following:

- TP is the true positives, i.e., the number of correctly predicted positive instances.
- FN is the false negatives, i.e., the number of actual positive instances that are incorrectly predicted as negative.

4. Results

4.1. Ablation Experiment

To validate the effectiveness of the proposed transformer-based densely nested feature fusion network and the lightweight multi-scale feature channel enhancement method (as shown in Table 3), we conducted ablation experiments from two perspectives:

1. For assessing the impact of the transformer-based densely nested feature fusion network, the segmentation results of the Swin-Unet++ model were compared to the Unet++ model on the CBIS-DDSM dataset. The results show that the Swin-Unet++ model achieved the best performance in terms of the mean DSC. This can be attributed to the incorporation of swin transformer blocks in the Swin-Unet++ model, which replaces traditional convolutional structures and enhances the model's ability to learn global image features, leading to improved feature representation.
2. Evaluating the effectiveness of the lightweight multi-scale feature channel enhancement method involved comparing the segmentation performance of the TEBLS model with the Swin-Unet++ model. The results show that incorporating multi-channel feature fusion significantly improved the mean AUC of the TEBLS model compared to Swin-Unet++. The mean AUC, which evaluated classification performance across various thresholds and was highly sensitive to boundary prediction errors, demonstrated notable gains. By integrating multi-channel feature fusion, the TEBLS model effectively combines features at different scales, capturing detailed image and semantic information while preserving target region details. This enhancement enhances the model's understanding of images, leading to more accurate boundary predictions.

Table 3. Ablation study results comparing model performance on the test dataset in terms of the mean DSC and mean AUC.

Model	Mean DSC	Mean AUC	Epoch
Unet++	89.52	83.00	150
Swin-Unet++	99.30	68.09	150
TEBLS	81.86	97.72	150

4.2. Quantitative Comparison with State-of-The-Art Models

We compared the segmentation performance of the proposed TEBLS model with previous segmentation models on the CBIS-DDSM dataset, including Density-ASP [39], U-Net [14], Res-Basic-U-Net [15], and AU-Net [40]. The results in terms of the mean DSC are presented in Table 4. With the same number of epochs, the segmentation performance of the TEBLS model improved by 35.3%, 21.8%, and 28.5%, respectively, compared to

that of the U-Net, AU-Net, and Res-Basic-U-Net models, which are based on simple skip connections, adaptive attention mechanisms, and residual connections.

Comparing the segmentation results of the U-Net model to those of Res-Basic-U-Net and AU-Net showed that AU-Net achieved the best performance. This was due to AU-Net's incorporation of a self-attention mechanism, which enhanced the learning of small targets and complex background features, in contrast to the residual network used in Res-Basic-U-Net. Additionally, when comparing AU-Net with Density-ASP, which uses AU-Net as its baseline, AU-Net outperformed Density-ASP. This was because Density-ASP dynamically adjusted the training priority based on sample density, making its segmentation outcomes highly dependent on dataset characteristics. In contrast, the proposed TEBLS model integrates a transformer-based structure with the Unet++ model and further improved feature learning through multi-channel feature fusion, resulting in superior segmentation performance.

Table 4. A quantitative comparison of segmentation results on the test dataset was conducted, including U-Net, Res-Basic-U-Net, AU-Net, Density-ASP, and our proposed method, TEBLS. All methods were trained on the same dataset under identical conditions and evaluated on the same test data.

Model	Mean DSC	Epoch
UNet	73.60	150
Res-Basic-UNet	77.50	150
AU-Net	81.80	150
Density-ASP	50.64	150
TEBLS	81.86	150

Table 5 presents the accuracy of the proposed TEBLS model compared to that of Swin-Unet, Vision Transformer (ViT) [17], and the Inceptionv3 model on the CBIS-DDSM dataset. The mean AUC represents the average AUC value obtained over multiple iterations, serving as a measure of the model's performance. As shown in Table 5, the proposed TEBLS model demonstrated a significant improvement in the mean AUC compared to the transformer-based Swin-Unet and ViT models. This improvement is attributed to the dense nested feature fusion in the TEBLS model, which enabled it to capture features at different levels and enhance feature representation through fusion.

Moreover, the TEBLS model outperformed the Inceptionv3 model in terms of the mean AUC. This was due to the inherent limitations of Inceptionv3, a CNN-based classification model, which relies on convolutional kernels and local receptive fields. These constraints reduce its effectiveness in capturing global features, particularly for tasks requiring long-range dependency modeling and a comprehensive understanding of global context. In contrast, transformer-based architectures leverage a global self-attention mechanism, enabling the more effective extraction of global features and long-range dependencies.

Table 5. Quantitative comparison of classification results, including Swin-Unet, ViT, Inceptionv3, and our proposed method, TEBLS.

Model	Mean AUC	Epoch
Swin-Unet	60.57	150
ViT	61.74	150
Inceptionv3	66.02	150
TEBLS	97.72	150

4.3. Model Evaluation and Complexity Analysis

Figure 4A shows the parameter complexity comparison among different models. The number of parameters is a key indicator of model complexity. A higher parameter count generally indicates stronger representational capacity, but it also requires more computational resources and data for training and inference. This necessitates greater GPU memory during training and more memory to store parameters during inference. As shown in Figure 4A, the TEBLS model has the fewest parameters, with a parameter count of 13,416,834. Compared to Swin-Unet++, the TEBLS model reduces the parameter count by approximately 0.144 M, and compared to Swin-Unet, it reduces the count by about 0.106 M. This demonstrates that the proposed TEBLS model is a lightweight image segmentation model.

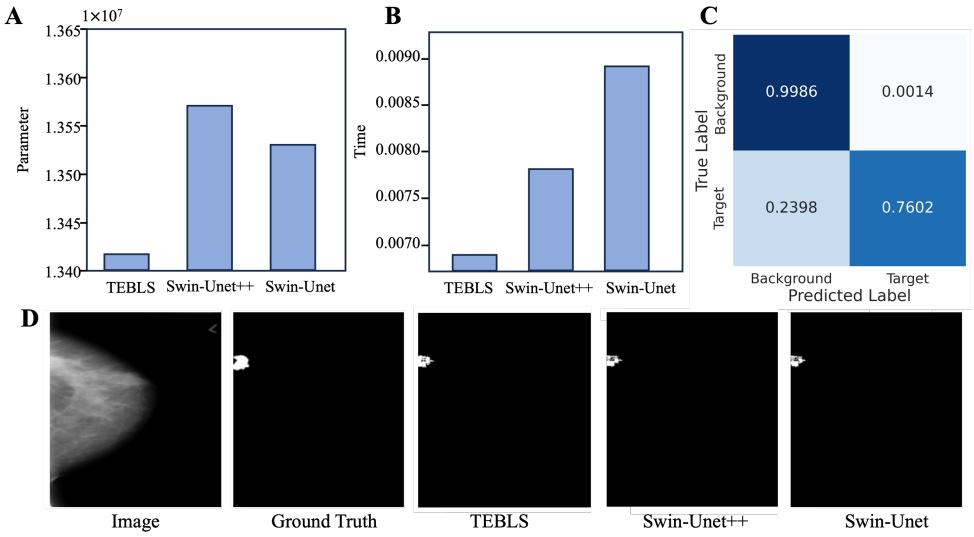

Figure 4. A performance comparison of different models in terms of parameter complexity, inference time, and segmentation accuracy. (**A**) A parameter count comparison shows that TEBLS has the fewest parameters, highlighting its lightweight nature. (**B**) An inference time comparison, demonstrating that TEBLS was the most efficient model, with faster processing compared to Swin-Unet++ and Swin-Unet. (**C**) A confusion matrix showing the results of the TEBLS model using the test set shows that the model's sensitivity was 0.7602. (**D**) Segmentation performance, where TEBLS outperformed other models by accurately capturing lesion regions with clear details at image edges and within lesion areas.

In Figure 4B, we compare the inference time of different models. Inference time measures the duration from when the model receives input data to when it generates output results. This metric evaluates the efficiency of a model in processing input data for real-world applications, with shorter times indicating higher efficiency. The inference time for TEBLS was 0.0069 s, compared to 0.0078 s for Swin-Unet++ and 0.0089 s for Swin-Unet. Although the reduction in inference time (0.0009–0.002 s) may seem marginal, experiments conducted on 574 images using an NVIDIA GeForce GTX 1650 processor demonstrated a significant reduction in the overall experimental runtime. Figure 4D illustrates the segmentation performance comparison among the models. TEBLS achieved the best segmentation performance, accurately capturing lesion regions with clear details at image edges and within lesion areas. To gain a deeper understanding of the model's performance under various conditions, Figure 4C presents the confusion matrix of the TEBLS model with the test set, which shows that the model achieved a sensitivity of 0.7602 with the test set.

4.4. Model Interpretability Analysis

Grad-CAM was employed to visualize how the TEBLS model divided lesion areas from non-lesion areas in the segmentation task, as shown in Figure 5. Figure 5 shows the Grad-CAM and superimposed images, highlighting the image regions most relevant to the features used by TEBLS to identify the lesion areas, where yellow indicates regions that TEBLS considered highly contributive to the prediction of lesion areas and purple indicates regions with a lower contribution. This enhances the clarity of the highlighted regions, assisting clinicians in evaluating whether the model's segmentation results align with their expert knowledge.

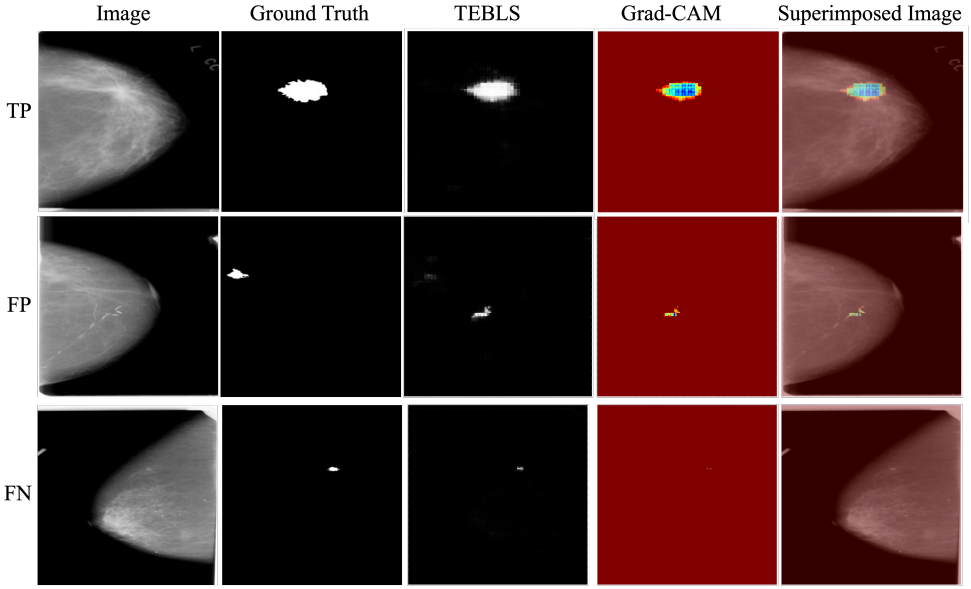

Figure 5. Visualizations of TEBLS outputs. The set includes the original input images, ground truth segmentations, TEBLS predictions, Grad-CAM visualizations highlighting model focus, and superimposed images showing the overlay of Grad-CAM heatmaps on the original images. The Grad-CAM visualizations help illustrate which areas of the image TEBLS prioritized during segmentation, providing insight into the model's decision-making process (TP: true positive; FP: false positive; FN: false negative).

5. Discussion

Segmenting lesion regions from medical imaging data is of significant importance for the early diagnosis and treatment of breast cancer. However, traditional image segmentation algorithms [11,20–23] can only extract low-level features, making them inadequate for diverse datasets and complex segmentation tasks. With advancements in computational power and data availability, CNN-based [14,24–29] and transformer-based [6,16,30–34] image segmentation models have gradually replaced traditional methods and are widely used in complex segmentation tasks. CNN-based segmentation models demonstrate strong feature extraction and representation capabilities, but due to the local receptive field of convolutional kernels, they are typically limited to processing local regions and struggle to capture long-range dependencies between pixels in an image. On the other hand, transformer-based models leverage self-attention mechanisms to effectively capture long-range dependencies but often fall short of capturing fine-grained semantic information. Moreover, the lack of interpretability in deep learning algorithms poses challenges for their application in the medical domain. To address these issues, we present an interpretable image segmentation model for the automated segmentation of breast cancer images.

To address the issue of blurred boundaries in segmentation results caused by existing models' lack of detailed spatial information capturing local features, we developed a

densely nested feature fusion model based on transformers. The model uses a locally fully connected structure to capture finer-grained features and performs multi-level feature fusion to extract multi-scale image information. Additionally, a self-attention mechanism was incorporated to capture intra-image interactions, enhancing the model's ability to learn spatial details. These mechanisms improved segmentation accuracy and detail representation and reduced boundary blurriness.

To reduce the complexity and high computational cost associated with transformer-based image segmentation models, we developed a lightweight multi-scale feature channel enhancement method. By extending and transforming feature channels, our method promoted effective information flow between channels, enabling the capture of more detailed semantic information related to breast cancer and achieving improved feature representation.

The effectiveness of the proposed model was evaluated using the mean AUC and mean DSC metrics for segmentation and classification tasks. Experimental results demonstrated that the model achieved accurate segmentation of breast cancer lesion regions. Moreover, a comparison of the TEBLS model with two other models (Swin-Unet++ and Swin-Unet) in terms of parameter complexity and inference time confirmed that the TEBLS model is a lightweight segmentation model with faster inference speed.

Additionally, Grad-CAM was used to enhance the interpretability of the transformer-based model, which can be challenging to understand due to its complex self-attention mechanisms. Figure 2 illustrates how Grad-CAM highlighted areas of focus within the model, helping to identify potential weaknesses, such as instances where the model may not have attended to the correct regions. This addresses the issues related to the data or training process.

In Figure 5, TP shows the true positive cases. TEBLS accurately captured the lesion areas, providing clear details in both the image edges and the lesion regions. In the false positive cases (Figure 5, FP), the model often misinterpreted background regions as lesions. Clinically, these regions are challenging even for radiologists due to their subtle lesion characteristics. Conversely, in false negative cases (Figure 5, FN), Grad-CAM visualizations indicated that the model may have focused on background regions instead of subtle lesion patterns. This issue might be attributed to the insufficient representation of small or low-contrast lesions in the training data. The visualization of TEBLS outputs using Grad-CAM can assist data scientists in understanding the model and improving results by analyzing the outputs. Furthermore, the insights provided by Grad-CAM can inform model refinement, allowing clinicians to evaluate whether the model's segmentation results align with their expert knowledge. The insights gained from Grad-CAM can also guide model refinement by informing adjustments that help the model focus on the most relevant features.

A limitation of the current study is that the experiments were conducted only on a breast cancer image dataset. In future work, we will extend the application of this model to other medical imaging tasks, such as 3D medical imaging and multimodal datasets. Additionally, we will integrate the lightweight channel enhancement module with techniques such as transfer learning and semi-supervised learning to further enhance the model's practicality and effectiveness in real-world clinical scenarios. Another limitation of our study is that all experiments were conducted exclusively on the CBIS-DDSM dataset. While 80% of the dataset was used as the training set and 20% as an independent validation set, there was no overlap between the training and validation data, ensuring the model's robustness and relevance. However, we plan to evaluate the effectiveness of the proposed model using multiple datasets in the future to further validate its robustness and generalizability.

6. Conclusions

In this study, we present a transformer-based explainable model for breast cancer lesion segmentation to achieve the automated and precise segmentation of breast cancer image data. This model combines a transformer-based dense nested feature fusion network with a lightweight channel enhancement module, with multi-scale features to enhance performance. The effectiveness of TEBLS was validated through extensive comparisons with existing segmentation models, demonstrating a higher accuracy in identifying breast cancer lesions. By incorporating these advanced techniques, TEBLS provides a promising solution for automated breast cancer diagnosis, offering both improved segmentation precision and greater interpretability, and it has the potential to be used in clinical settings.

Author Contributions: Conceptualization, H.W., L.W. and C.M.; Data curation, H.W., L.W. and C.M.; Formal analysis, H.W., L.W. and C.M.; Investigation, H.W., L.W., B.L., J.L. (Jianqiang Li), J.L. (Jinshu Li), J.F. and C.M.; Methodology, H.W., L.W. and C.M.; Resources, H.W., L.W., B.L., J.L. (Jianqiang Li), J.L. (Jinshu Li), J.F. and C.M.; Supervision, C.M.; Validation, H.W., L.W. and C.M.; Visualization, H.W., L.W. and C.M.; Writing—original draft, H.W., L.W. and C.M.; Writing—review and editing, H.W., L.W. and C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (Grant Number: 62076015), the 2023 International Cooperation Training Program for Innovative Talents ("Double First-class" Construction Special Program—"Artificial Intelligence + Internet of Things") of the China Scholarship Council (CSC), and in part by the CSC under the Grant No. 202406540004. This publication emanated from research supported in part by a research grant from Research Ireland under the Grant Number 21/RC/10294_P2 and co-funded under the European Regional Development Fund and by FutureNeuro industry partners.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset and accessed on 10 March 2024; reference number [35].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bray, F.; Laversanne, M.; Weiderpass, E.; Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **2021**, *127*, 3029–3030. [CrossRef] [PubMed]
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef]
4. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
5. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 36–46.
6. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
7. Houssein, E.H.; Emam, M.M.; Ali, A.A. An efficient multilevel thresholding segmentation method for thermography breast cancer imaging based on improved chimp optimization algorithm. *Expert Syst. Appl.* **2021**, *185*, 115651. [CrossRef]
8. Tang, J. A color image segmentation algorithm based on region growing. In Proceedings of the 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 16–18 April 2010; Volume 6, pp. V6-634–V6-637.

9. He, F.; Mahmud, M.P.; Kouzani, A.Z.; Anwar, A.; Jiang, F.; Ling, S.H. An improved SLIC algorithm for segmentation of microscopic cell images. *Biomed. Signal Process. Control.* **2022**, *73*, 103464. [CrossRef]
10. Jardim, S.; António, J.; Mora, C. Image thresholding approaches for medical image segmentation: Short literature review. *Procedia Comput. Sci.* **2023**, *219*, 1485–1492. [CrossRef]
11. Roy, S.; Das, D.; Lal, S.; Kini, J. Novel edge detection method for nuclei segmentation of liver cancer histopathology images. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 479–496. [CrossRef]
12. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. U-Net++: A nested U-Net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Granada, Spain, 20 September 2018; pp. 3–11.
13. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. U-Net 3+: A full-scale connected U-Net for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1–5.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
15. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
17. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6824–6835.
19. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–38. [CrossRef]
20. Xing, Z. An improved emperor penguin optimization based multilevel thresholding for color image segmentation. *Knowl. Based Syst.* **2020**, *194*, 105570. [CrossRef]
21. Mittal, H.; Saraswat, M. An optimum multi-level image thresholding segmentation using non-local means 2D histogram and exponential Kbest gravitational search algorithm. *Eng. Appl. Artif. Intell.* **2018**, *71*, 226–235. [CrossRef]
22. Cigla, C.; Alatan, A.A. Region-based image segmentation via graph cuts. In Proceedings of the 2008 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 2272–2275.
23. Tasli, H.E.; Cigla, C.; Alatan, A.A. Convexity constrained efficient superpixel and supervoxel extraction. *Signal Process. Image Commun.* **2015**, *33*, 71–85. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
26. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
27. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
30. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
31. d'Ascoli, S.; Touvron, H.; Leavitt, M.L.; Morcos, A.S.; Biroli, G.; Sagun, L. Convit: Improving vision transformers with soft convolutional inductive biases. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 2286–2296.
32. Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Tong, R. ScaleFormer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv* **2022**, arXiv:2207.14552.
33. Zhou, H.Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. nnformer: Interleaved transformer for volumetric segmentation. *arXiv* **2021**, arXiv:2109.03201.
34. Ji, Y.; Zhang, R.; Wang, H.; Li, Z.; Wu, L.; Zhang, S.; Luo, P. Multi-compound transformer for accurate biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Part I 24, Strasbourg, France, 27 September–1 October 2021; pp. 326–336.

35. Lee, R.S.; Gimenez, F.; Hoogi, A.; Miyake, K.K.; Gorovoy, M.; Rubin, D.L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **2017**, *4*, 170177. [CrossRef]
36. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
37. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
38. Vinogradova, K.; Dibrov, A.; Myers, G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13943–13944.
39. Aliniya, P.; Nicolescu, M.; Nicolescu, M.; Bebis, G. Hybrid Region and Pixel-Level Adaptive Loss for Mass Segmentation on Whole Mammography Images. In Proceedings of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 16–18 October 2023; pp. 3–17.
40. Liu, C.; Guo, X.; Jiang, J. AU-Net: A deep learning network for precise water body extraction in the middle and lower reaches of the yellow river. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2022**, *10*, 107–113. [CrossRef]