Neural Computing and Applications (2024) 36:16427–16443
https://doi.org/10.1007/s00521-024-09963-w

ORIGINAL ARTICLE

Dual-Stream CoAtNet models for accurate breast ultrasound image segmentation

Nadeem Zaidkilani[1] · Miguel Angel Garcia[2] · Domenec Puig[1]

Received: 14 September 2023 / Accepted: 3 May 2024 / Published online: 27 May 2024

Abstract

The CoAtNet deep neural model has been shown to achieve state-of-the-art performance by stacking convolutional and self-attention layers. In particular, the initial layers of CoAtNet apply efficient convolutions for extracting local features out of the input image and the initial fine-resolution feature maps. In turn, the final layers apply more cumbersome Transformers in order to extract global features from the coarse-resolution feature maps. The model's outcome directly depends on those final global features. This paper proposes an extension of the original CoAtNet model based on the introduction of a dual stream of convolution and self-attention blocks applied at the final layers of CoAtNet. In this way, those final layers automatically aggregate both local and global features extracted from the initial feature maps. Two dual-stream topologies have been proposed and evaluated. This Dual-Stream CoAtNet model exhibits a significant improvement on the segmentation accuracy of breast ultrasound images, thus contributing to the development of more robust tumor detection methods.

Keywords Breast cancer · Ultrasound image segmentation · Deep neural networks · Transformers · CoAtNet

1 Introduction

Breast cancer affects women globally. According to the National Cancer Institute, the relative survival rate for breast cancer patients in the United States is 99% provided the disease is identified and treated at an early stage. Otherwise, the survival rate drops significantly to 27% [1]. Ultrasound imaging has become a popular diagnostic tool due to its numerous benefits: fast imaging, high sensitivity, and low cost [2]. However, the sheer volume of images generated daily and the limited availability of radiologists to interpret these images can lead to misdiagnosis. The latter can have severe consequences. Researchers are thus exploring the utilization of artificial intelligence (AI) and machine learning algorithms to help radiologists interpret ultrasound images. These technologies can analyze large datasets quickly and accurately, thus reducing the likelihood of errors and improving diagnostic accuracy. Accurately spotting small tumors is essential for the early detection of breast cancer. In the initial phases, tumors are often characterized by their small size and their presence within a relatively limited area in breast ultrasound (BUS) images. This explains the difficulty in distinguishing them from normal breast tissue. AI can help radiologists detect small tumors in BUS images, leading to earlier treatment, improved patient outcomes, and better survival rates.

Computer-aided diagnosis (CAD) systems have been created to aid doctors in making accurate diagnostic decisions during early screening and diagnosis of breast cancer using BUS images [3]. Image segmentation is a crucial task in CAD systems. In our scope, it requires the identification of tumor boundaries in BUS images. This information is necessary for subsequent screening and diagnosis. However, BUS images present inherent challenges, including speckle noise and low contrast, leading to false positives during image segmentation [4]. These issues hinder the widespread use of CAD systems, making it imperative to develop robust image segmentation methods that reduce

✉ Nadeem Zaidkilani
nadeem.zaidkilani@estudiants.urv.cat;
nadimkilani@gmail.com

[1] Department of Computer Engineering and Mathematics, University Rovira i Virgili, Tarragona 43007, Spain

[2] Department of Electronic and Communications Technology, Autonomous University of Madrid, Madrid, Spain

false positives and misdetections, and enable efficient segmentation of BUS images.

We propose a new approach for image segmentation of BUS images using deep neural networks specifically trained for tumor detection. In particular, the proposed method explores two variations of the successful CoAtNet deep network [5]. The original CoAtNet exhibits state-of-the-art performance by stacking convolutional and self-attention layers. Furthermore, CoAtNet does not require any extra intensity normalization stages, such as [6, 7], which are applied for improving the performance of automated image segmentation and analysis methods.

The initial layers of CoAtNet apply efficient convolutions for extracting local features out of the input image and the initial fine-resolution feature maps. In turn, the final layers apply more complex and time-consuming Transformers in order to extract global features that express contextual visual information from the coarse-resolution feature maps. The model's outcome directly depends on those final global features. The proposed method replaces the attention layers of CoAtNet by a dual stream of both attention and convolutional layers that are automatically integrated. The aim is that those final layers also consider local information present in the coarse-resolution feature maps, similarly to what pure convolutional neural networks do, hence complementing the global information extracted by the self-attention mechanism. Experimental results with two well-known BUS image datasets show that this Dual-Stream CoAtNet model significantly improves the segmentation accuracy of breast ultrasound images, thus contributing to the development of more robust tumor detection methods.

This paper is organized as follows. Section 2 summarizes recent related work for segmenting BUS images with deep neural networks. Section 3 describes the proposed method. In particular, Sect. 3.1 presents the two proposed variations of the original CoAtNet model by introducing a dual stream of convolution and attention layers. In Sect. 3.2, the CoAtNet model is then embedded into an encoder–decoder architecture for segmenting BUS images. That architecture is also adapted to the two proposed Dual-Stream CoAtNet models. Section 3.3 describes the different loss functions that have been considered in order to train the proposed models, whereas Sect. 3.4 presents the evaluation measures that have been utilized to assess the quality of the generated segmentations. In turn, Sect. 4 shows the experimental validation. The two public datasets of BUS images that have been utilized and the data augmentation policies are described in Sect. 4.1. The preliminary experiments conducted for choosing the most appropriate loss function for training the different models are summarized in Sect. 4.2. The experimental evaluation is shown in Sect. 4.3. The robustness of the model is discussed in Sect. 4.4. Finally, conclusions and future work are given in Sect. 5.

2 Related work

Convolutional neural networks (CNNs) have rapidly evolved in the field of computer vision, becoming the primary method for automated segmentation in computer-aided diagnosis systems. Additionally, they have proven successful in both classifying and segmenting medical images [8–10]. In the medical image segmentation field, deep learning techniques relying on CNNs have been the dominant approach since the introduction of the U-Net model [11]. Many improved versions of U-Net have been suggested. One notable example is U-Net++ [12], which builds upon the foundational U-Net architecture. It concatenates the four layers of U-Net and aggregates feature maps of different scales into the decoder, thus enhancing segmentation accuracy. Another significant example is ResUNet [13], which applies a U-Net encoder/decoder backbone complemented by residual connections, atrous convolutions, pyramid scene parsing pooling and multi-tasking inference.

More recently, various deep-learning techniques have emerged for segmenting BUS images. Huang et al. [14] introduced a fully convolutional network that addressed the uncertainty problem in the original images and feature maps using fuzzy logic. Nair et al. [15] developed a deep neural network (DNN) with two decoders capable of generating ultrasound images and segmentation masks derived from raw single-plane wave channel data. This approach is producing promising results. On the other hand, Zhuang et al. [16] proposed the RDAU-Net model, which is a modification of the U-Net architecture that replaces the original skip connections and basic blocks with attention gates and dilated residual blocks, respectively. This improved the performance of tumor segmentation in BUS images.

In addition to these methods, other architectures have been proposed to enhance the performance of tumor segmentation. Among them, Zaidkilani et al. [17] presented a CAD system for breast ultrasound images utilizing a two-stage process with an encoder–decoder network for tumor segmentation and fine-tuned MobileNetv2 for classifying benign and malignant tumors. Shareef et al. [18] proposed a deep learning framework called Small Tumor-Aware Network, which combines contextual details and high-resolution image attributes. This network has been shown to enhance the segmentation performance of tumors of different sizes. Similarly, Vakanski et al. [19] introduced a salient attention-based approach that incorporates attention blocks into a U-Net architecture, allowing the network the

capability to learn feature depictions that assign priority to spatial regions with high saliency levels.

Deng et al. [20] proposed a method that preserves the transferability of the original GAN and incorporates a new optimizer to generate a continuous distribution sample space. Byra et al. [21], in a similar way, utilized selective kernel (SK) mechanisms to adjust the network's receptive fields through attention mechanisms. They also fused feature maps extracted using dilated and conventional convolutions. On the other hand, Zhou et al. [22] developed a lightweight attention encoder–decoder network, where the encoder component is a streamlined version of EfficientNet. The decoder component incorporates a Lightweight Residual Squeeze-and-Excitation (LRSE) block.

Alternatively, Xu et al. [23] proposed RMTL-Net, a multi-task learning framework that enables simultaneous breast ultrasound image segmentation and classification tasks. The framework utilizes ResNet-101 as the backbone architecture and incorporates a regional attention module to capture category-sensitive information. Similarly, Zhang et al. [24] developed a new model for routine BUS screening, comprising a classification branch to differentiate between healthy and tumorous tissue, and a segmentation branch to outline tumors. By employing a multi-task loss function that combines Binary Cross-Entropy and Dice loss, the model achieves high accuracy with low false positives for normal images while preserving sensitivity for tumor detection.

Furthermore, Tang et al. [25] presented the MGCC framework, focusing on semi-supervised learning for medical image segmentation. The framework addresses challenges associated with collecting unlabeled medical data by utilizing synthetic medical images generated by a latent diffusion model (LDM) as unlabeled data. It incorporates global context noise perturbation and ensures output consistency between decoders, resulting in improved representation ability.

Researchers have also developed innovative approaches based on deep learning networks to improve segmentation tasks. Ahmed et al. [26] presented COMANet (COMplementary Attention guided bipolar refinement-based Network). That network applies a complementary attention scheme involving positive and negative refinement modules on two encoder structures. This configuration aims to generate refined feature references for the decoder in a supervised manner. Additionally, they proposed a novel index called Foreground-to-Background Ratio (FBR) to highlight signal power differences between the target region and background due to the refinement process. Furthermore, Ta et al. [27] proposed LET-Net, a Locally Enhanced Transformer Network that adeptly tackles the challenges associated with accurately segmenting small targets. They combine the strengths of transformers and convolutions.

In addition, Heidari et al. [28] put forward HiFormer, a groundbreaking approach for medical image segmentation that seamlessly merges a CNN and a transformer. Utilizing the Swin Transformer module and a CNN-based encoder, HiFormer crafts two multi-scale feature representations and integrates global and local features through the Double-Level Fusion (DLF) module within the encoder–decoder skip connection. Moreover, Yuan et al. [29] proposed CTC-Net, an innovative CNN and Transformer Complementary Network, enhancing medical image segmentation by integrating complementary features from Swin Transformers and Residual CNNs. This approach leverages a Cross-domain Fusion Block and a Feature Complementary Module, incorporating a Swin Transformer decoder with skip connections for improved representation of long-range dependencies and multi-scale invariance. Dar et al. [30] introduced EfficientU-Net, a novel deep-learning method aimed at enhancing breast tumor segmentation in ultrasound images. This approach integrates a modified EfficientNet and an atrous convolution block in the UNet model's encoder, minimizing training parameters through depth-wise separable convolutions. Furthermore, Yang et al. [31] presented Rema-Net, an attention-grabbing multi-attention convolutional neural network designed for swift skin lesion segmentation. The network's down-sampling module incorporates spatial attention, while strategic skip-connections and a reverse attention operation further contribute to enhancing segmentation performance.

More recently, Ahmad et al. [32] proposed DoubleU-NetPlus, with enhancements to the DoubleU-Net, incorporating EfficientNetB7 as the feature encoder, and introducing new modules for better feature mapping in medical images. These improvements address issues like gradient vanishing with high-resolution features. Notably, novel triple attention gates and hybrid triple attention modules were introduced to selectively model relevant features. Standard convolution operations were replaced by attention-guided residual convolution operations to ensure effective feature map generation despite the network's increased depth. In addition, Hekal et al. [33] introduced an innovative deep learning method for breast cancer segmentation in ultrasound images, utilizing the Dual Decoder Attention ResUNet (DDA-AttResUNet). This model features a unique structure that simultaneously emphasizes tumor segmentation and captures supplementary contextual information. The incorporation of dual decoding attention and an attention mechanism within the ResUNet significantly enhances segmentation accuracy, resulting in improved breast cancer detection outcomes from ultrasound images. Furthermore, Zhang et al. [34] presented HAU-Net, a new model designed for the segmentation of

breast tumors in ultrasound images. This model seamlessly incorporates features from transformers and convolutional neural networks (CNNs). Notably, it replaces conventional skip connections with an L-G transformer block, facilitating effective long-range dependency modeling while maintaining network integrity. The incorporation of the Cross-Attention Block (CAB) improves the interaction of information among multi-size feature layers, achieving superior feature representation and enhancing segmentation accuracy. Following a different approach, Hüseyin [35] introduced ConvMixer-based Encoder-Classification-Based Decoder (CE-CD), a pioneering network architecture designed for breast lesion segmentation in ultrasound images. This novel model uniquely divides the segmentation task into image-level classification and pixel-level detection. It effectively combines ConvMixer and DenseNet121 in the encoder, capturing spatial, semantic, and long-range contextual details. The decoder integrates both a classification network and a detection network, providing lesion detection scores at the pixel level and lesion classification scores at the image level. The final lesion class is determined through a result generation algorithm. CE-CD presents a comprehensive and effective approach that harnesses the capabilities of ConvMixer and DenseNet121, leading to enhanced accuracy in breast lesion segmentation from ultrasound images.

Overall, deep learning approaches have shown promising results in the segmentation of BUS images. As this technology keeps maturing, new and improved segmentation methods will emerge, further improving the accuracy and efficiency of breast cancer diagnosis.

2.1 CoAtNet model

The proposed method for BUS image segmentation builds upon the CoAtNet model [5]. CoAtNet is a hybrid architecture that combines well-established CNNs and modern Transformers. Specifically, the first three stages of CoAtNet apply convolutional blocks for extracting local feature maps from the given image. Afterward, Transformers are applied in order to extract global feature maps through self-attention. Figure 1 shows an overview of the original CoAtNet architecture.

Specifically, the input RGB image is fed into a block of two consecutive 3×3 convolution layers (Conv2d) in the first stage (*Stage*0) of CoAtNet. Batch normalization and a GELU nonlinear activation function are applied after each convolution. The first convolution layer downsamples the given image by applying stride 2 and raises the number of channels from 3 to 64 in the minimum configuration of CoAtNet (coatnet-0 [5]). The second stage (*Stage*1) applies a sequence of efficient convolution blocks proposed in the MobileNet.v2 architecture (MBConv blocks). Each MBConv block applies a first 1×1 2D convolution (Conv2d) that quadruplicates the number of input channels, followed by a depthwise 3×3 2D convolution, and a final 1×1 2D convolution that projects the expanded channels into the desired output channels. This scheme of expanding channels, applying depthwise convolution and contracting channels is referred to in the literature as *inverted bottleneck*. Batch normalization and GELU are applied after each convolution. Each MBConv block computes the residual branch that is added to the corresponding identity branch. Batch normalization is also applied at the beginning of the residual branch (i.e., pre-activation). The number $S1$ of MBConv blocks sequentially applied within the second stage is $S1 = 2$ for the minimum CoAtNet configuration (coatnet-0). The first MBConv block in this sequence downsamples the given feature map by applying stride 2 at its first 1×1 2D convolution and also raises the number of channels from 64 to 96 (coatnet-0) at its last 1×1 2D convolution. Its corresponding identity branch also downsamples the given feature map by applying 2D max pooling and expands the number of channels to 96 by applying a 1×1 convolution. The third stage (*Stage*2) has the same structure as the previous one. The number $S2$ of MBConv blocks sequentially applied is $S2 = 3$ for coatnet-0, and the number of channels is risen from 96 to 192.

The following two stages of CoAtNet apply Transformer blocks (see Fig. 1). Each Transformer block consists of a relative self-attention layer followed by a feed-forward layer. The relative self-attention layer is applied to the residual branch and consists of 8 parallel attention heads, with each head processing 32 dimensions/channels. That attention residual branch is added with the identity branch. In turn, the feed-forward layer applies another

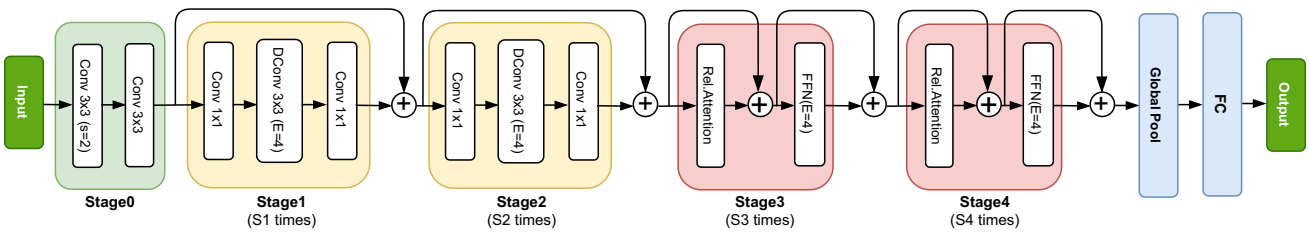

Fig. 1 Overview of the original CoAtNet model [5]

inverted bottleneck to the residual branch. It consists of a first fully connected layer (Linear) that quadruplicates the number of input channels, followed by GELU and dropout. The expanded channels are then contracted to the given number of channels with a second fully connected layer also followed by dropout. That feed-forward residual branch is added with the identity branch. Pre-activation in terms of layer normalization is applied at the beginning of both the self-attention and feed-forward layers.

The fourth stage of CoAtNet (*Stage*3) sequentially applies $S3 = 5$ Transformer blocks for coatnet-0. The first Transformer block in that stage downsamples the given feature map by applying 2D max pooling prior to applying the self-attention layer. In addition, that layer also raises the number of channels from 192 to 384. In turn, the corresponding identity branch for the first Transformer block also downsamples the input feature through 2D max pooling and duplicates the number of channels by applying a 1×1 2D convolution. Finally, the fifth stage of CoAtNet (*Stage*4) sequentially applies $S4 = 2$ Transformer blocks for coatnet-0 by following the same scheme described above. A total of 768 feature maps are thus generated for coatnet-0. These five stages of CoAtNet constitute its backbone.

Given a 256×256 input image, the size of each feature map generated by the backbone is 8×8. Those feature maps are then passed to a 2D average pooling layer (AvgPool2d) that averages the 64 elements of each map, thus generating a 768-D tensor for coatnet-0. The latter is finally fed into a fully connected classification layer that generates a 1D tensor with as many elements (logits) as desired classes (e.g., 1,024 for ILSVRC/ImageNet).

The only difference between the different configurations of CoAtNet is the number of sequential blocks in Stage2 and Stage3, as well as the number of generated channels. For example, in the largest configuration (coatnet-4), $S2 = 12$ and $S3 = 28$ ($S2 = 3$ and $S3 = 5$ in coatnet-0), and the number of generated channels is 192 for both Stage0 and Stage1, 384 for Stage2, 768 for Stage3, and 1,536 for Stage4 (64, 96, 192, 384 and 768 for coatnet-0).

3 Proposed method

3.1 Dual-Stream CoAtNet

As described in Sect. 2.1, the first three stages of the original CoAtNet model apply convolutional blocks for extracting local feature maps from the given image, whereas the following two stages apply Transformer blocks in order to extract global feature maps through self-attention. The final output directly depends on those global features. This is the main difference between the CoAtNet model and conventional CNNs, in which all stages extract local feature maps through convolutional blocks.

However, discarding local features in the last stages of the neural model in favor of global features may not necessarily be the best approach in general. With this idea in mind, we propose to add a second stream of convolutional blocks running in parallel with the last two stages of Transformer blocks. In that way, the network computes both local and global feature maps in parallel, which are finally integrated through a combination stage that is trained end-to-end as part of the whole model. The same efficient MBConv blocks utilized in *Stage*1 and *Stage*2 are applied in the new *Stage*3′ and *Stage*4′. The number of MBConv blocks applied in *Stage*3′ is $S3 = 5$ for coatnet-0, and the number of channels is risen from 192 to 384, similarly to *Stage*3. In turn, $S4 = 2$ MBConv blocks are applied in *Stage*4′ for coatnet-0, expanding the channels from 384 to 768 as in *Stage*4.

Two variations of this Dual-Stream CoAtNet architecture have been proposed and tested depending on how the two parallel streams are merged. The first variation (DSCoAtNet.v1) is summarized in Fig. 2. In this case, the new *Stage*3′ and *Stage*4′ are run in parallel with *Stage*3 and *Stage*4. The 768×2 channels generated for coatnet-0 are then reduced to 768 channels by first concatenating them along the channel dimension and then applying a 1×1 2D convolution layer followed by batch normalization and GELU. In turn, the second variation (DSCoAtNet.v2) is summarized in Fig. 3. First, the new *Stage*3′ is run in parallel only with *Stage*3. The 384×2 channels generated for coatnet-0 are reduced to 384 channels by applying the same combination block described above (channel concatenation plus 1×1 2D convolution). Finally, the new *Stage*4′ runs in parallel with *Stage*4. The 768×2 channels generated for coatnet-0 are reduced to 768 channels by applying the same combination block.

From a mathematical standpoint, the original CoAtNet pipeline depicted in Fig. 1 can be described as:

$$\mathbf{y} = \mathrm{FC}(\mathrm{Attn}(\mathrm{Attn}(\mathrm{MBConv}(\mathrm{MBConv}(\mathrm{Stage0}(\mathbf{x})))))) \tag{1}$$

In turn, our first variation (DSCoAtNet.v1) summarized in Fig. 2 can be formulated as:

$$\mathbf{y} = \mathrm{FC}(\mathrm{Dual1}(\mathrm{MBConv}(\mathrm{MBConv}(\mathrm{Stage0}(\mathbf{x}))))) \tag{2}$$

$$\mathrm{Dual1}(\mathbf{x}) = \mathrm{Conv}(\mathrm{Concat}(\mathrm{Attn}(\mathrm{Attn}(\mathbf{x})), \mathrm{MBConv}(\mathrm{MBConv}(\mathbf{x})))) \tag{3}$$

Similarly, our second variation (DSCoAtNet.v2) shown in Fig. 3 can be formulated as:

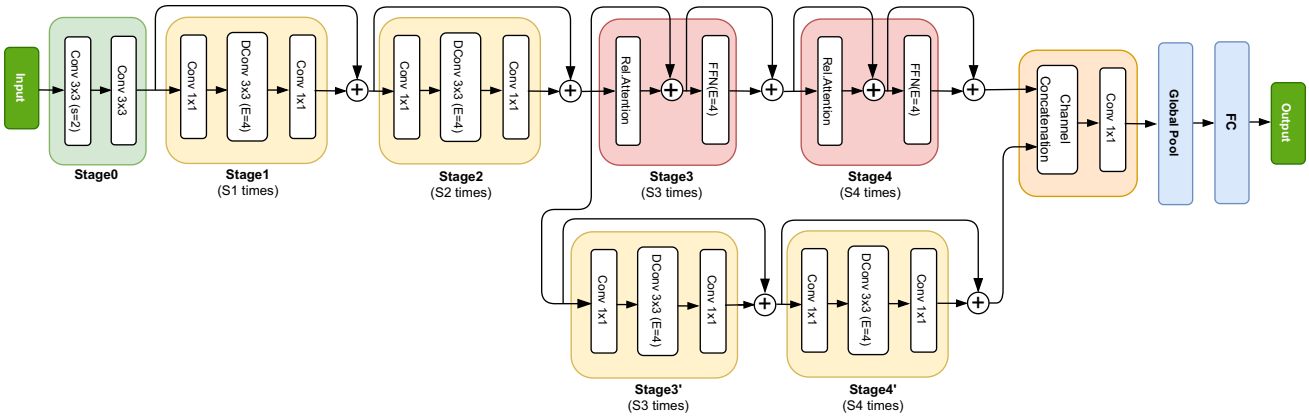

Fig. 2 First variation of Dual-Stream CoAtNet (DSCoAtNet.v1)

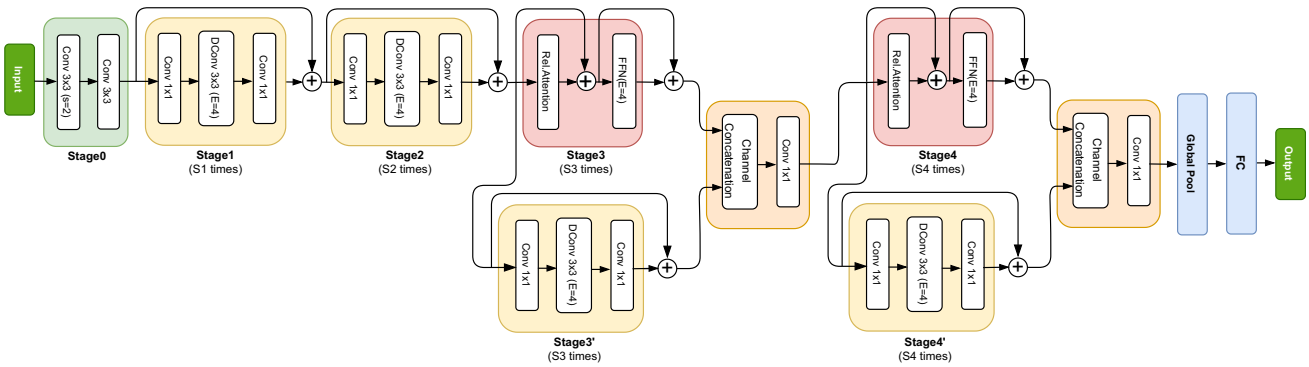

Fig. 3 Second variation of Dual-Stream CoAtNet (DSCoAtNet.v2)

$$\mathbf{y} = \mathrm{FC}(\mathrm{Dual2}(\mathrm{Dual2}(\mathrm{MBConv}(\mathrm{MBConv}(\mathrm{Stage0}(\mathbf{x})))))) \tag{4}$$

$$\mathrm{Dual2}(\mathbf{x}) = \mathrm{Conv}(\mathrm{Concat}(\mathrm{Attn}(\mathbf{x}), \mathrm{MBConv}(\mathbf{x}))) \tag{5}$$

3.2 BUS image segmentation with Dual-Stream CoAtNet

The final goal of this deep network is to generate a predicted image that allows the segmentation of the input BUS image into both tumorous and healthy regions. However, CoAtNet is a classification model. As described in Sect. 2.1, its first five stages constitute its backbone. It generates a collection of feature maps (768 for coatnet-0). Those maps are then converted into a 1D tensor with as many elements (logits) as desired classes. The given image is classified into the class with the largest associated logit.

Therefore, CoAtNet alone is not able to generate a predicted image out of a given RGB image. In order to do that, an encoder–decoder (i.e., autoencoder) scheme is applied in this work by following the U-Net architecture [11], although using the CoAtNet backbone as encoder. This is summarized in Fig. 4.

The 768 8×8 feature maps generated by the CoAtNet's backbone are fed into a convolution block that behaves as the U-Net's bottom layer (bottleneck). It is constituted by two consecutive 3×3 2D convolution layers (Conv2d). Batch normalization and ReLU are applied after each layer. The result is fed into the decoder branch, which consists of five consecutive decoder blocks.

Every decoder block has the following structure. First, a 2×2 2D transposed convolution layer (ConvTranspose2d) with stride 2 is applied. This layer halves the number of input channels (feature maps) and duplicates the height and width of those maps. The feature maps generated by that transposed convolution are then concatenated along the channel dimension with the corresponding maps of identical size generated from the encoder. This implements the skip connection in the U-Net architecture. The result is passed to a convolution block constituted by two consecutive 3×3 2D convolution layers (Conv2d), with batch normalization and ReLU applied after each layer. The first layer receives the concatenated channels and halves the number of channels, hence combining them. The output of the second layer is fed into the next decoder block.

The output of the fifth decoder block is an RGB image (3 channels), which is converted into the sought single

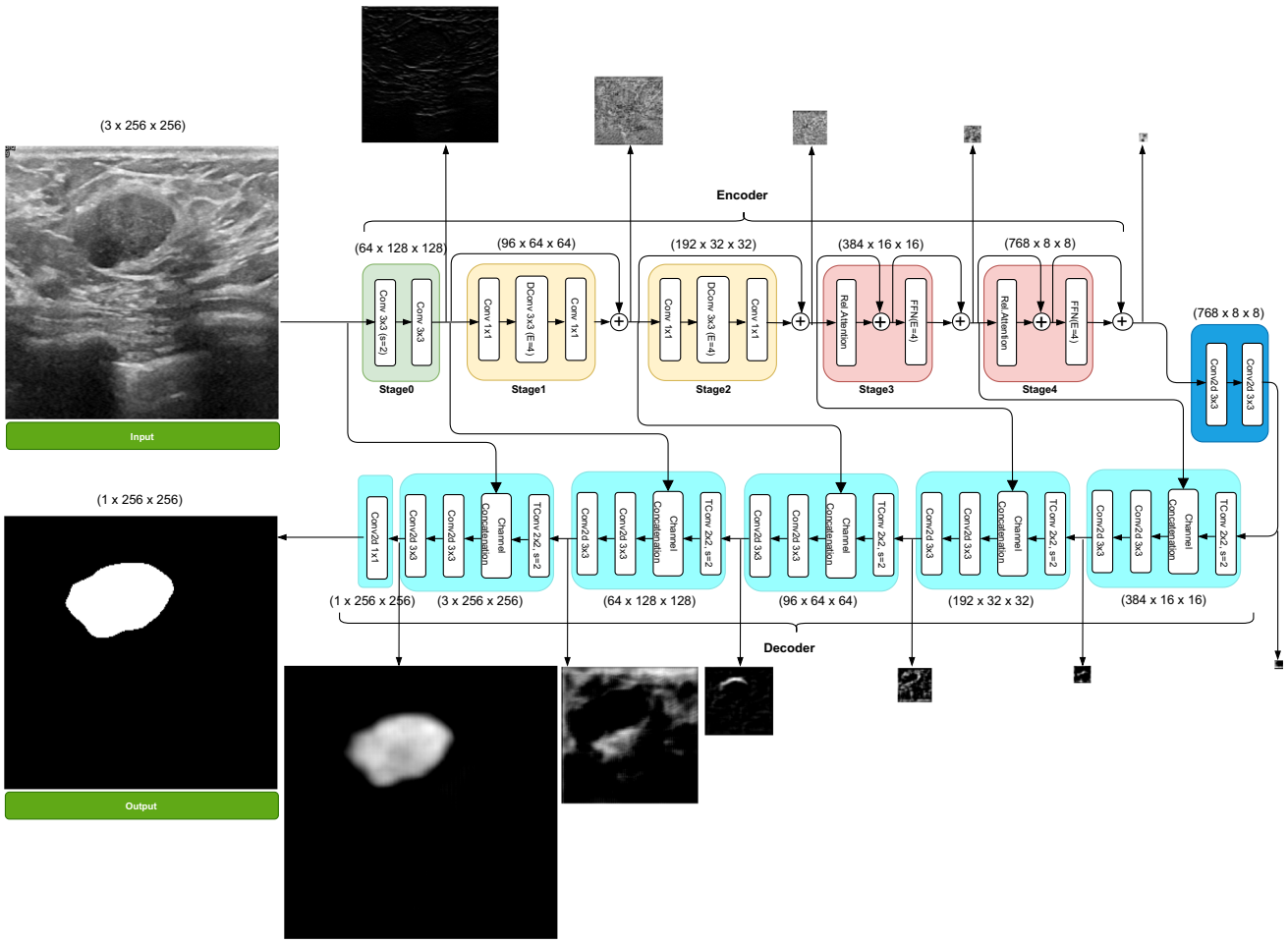

Fig. 4 BUS image segmentation with autoencoder based on CoAtNet

channel image by applying a last 1×1 2D convolution layer (Conv2d).

This basic encoder–decoder network based on the original CoAtNet has been adapted to the two variations of the proposed Dual-Stream CoAtNet model in a straightforward manner, as shown in Figs. 5 and 6, respectively.

3.3 Loss functions

Selecting an appropriate loss function is essential for training deep neural networks, as each function has its benefits and drawbacks. Furthermore, loss functions play a pivotal role in determining whether all pixels in an image contribute equally during training or if their importance varies depending on the class label. In our application scope, the minority class (tumorous regions) should receive adequate attention compared to the dominant class (healthy tissue).

Let us first define the following classification measures: *true positives* (*TP*), *true negatives* (*TN*), *false positives* (*FP*), and *false negatives* (*FN*), defined as follows. Let *TP* be the number of true positives, that is, the predicted tumorous regions that are really tumorous in the ground-truth (GT). Let also *TN* be the number of true negatives, that is, the predicted healthy regions that are really healthy. In addition, let *FP* be the number of false positives, that is, the predicted tumorous regions that are really healthy. Finally, let *FN* be the number of false negatives, that is, the predicted healthy regions that are really tumorous.

The last stage of the decoder branch (1×1 Conv2d) does not directly generate a binary image, but a 2D array of logits. We apply a sigmoid function to that array in order to generate the predicted image $\mathbf{P}$. The latter is a collection of pixels, $\mathbf{P} = \{p_i\}$, such that $p_i \in [0, 1]$ represents the probability of the corresponding pixel being tumorous. The predicted image is finally thresholded to obtain the predicted binary image $\mathbf{Y}$. The latter is a collection of pixels, $\mathbf{Y} = \{y_i\}$, such that $y_i = 1$ if the pixel is predicted to be tumorous, and $y_i = 0$ otherwise. Finally, the GT image $\mathbf{G}$ is

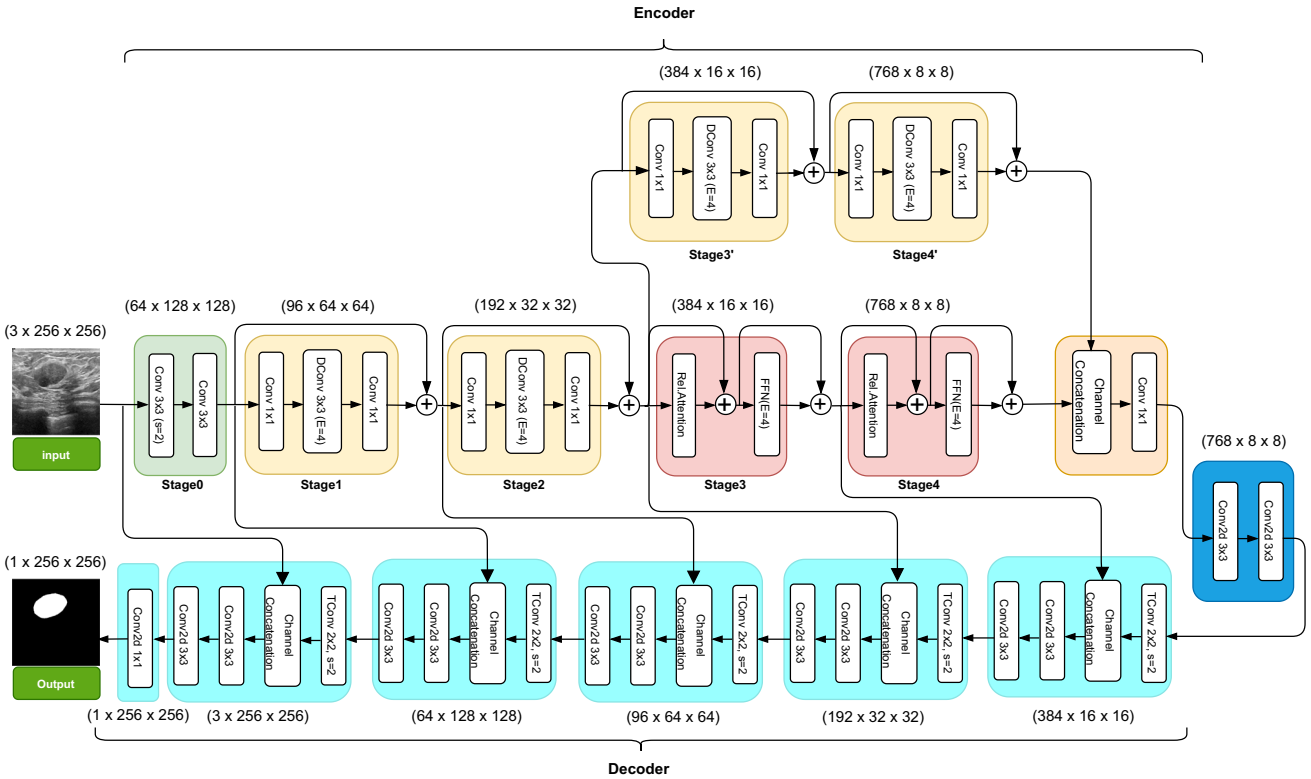

Fig. 5 BUS image segmentation with autoencoder based on DSCoAtNet.v1

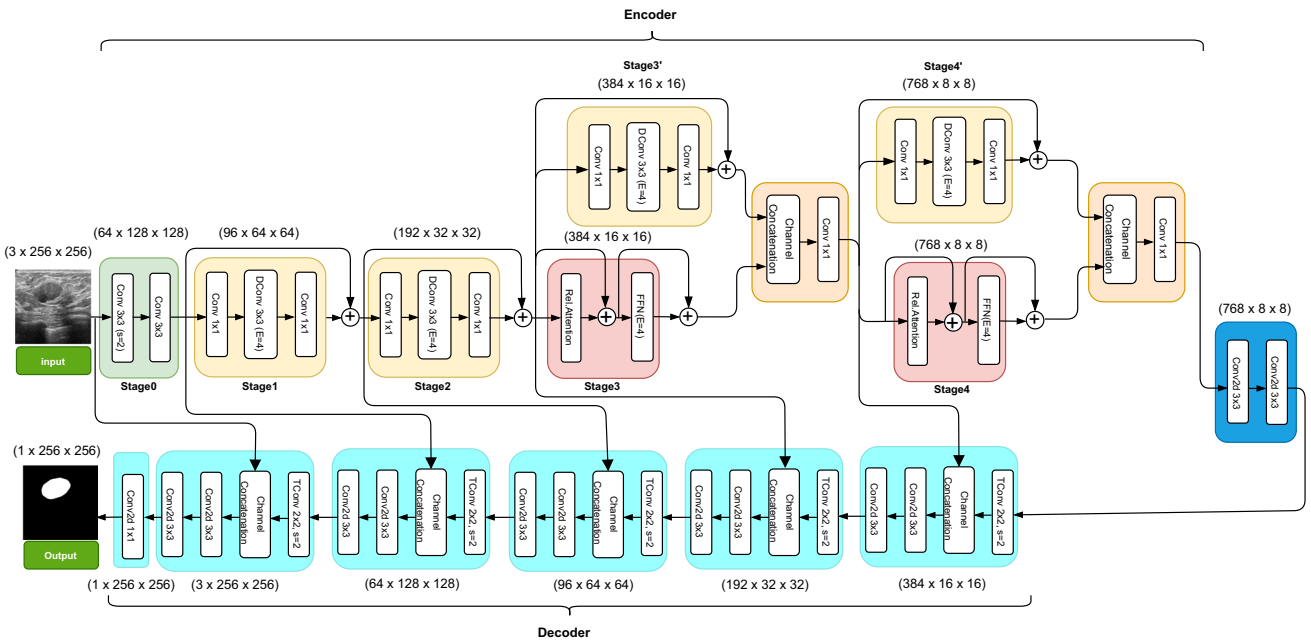

Fig. 6 BUS image segmentation with autoencoder based on DSCoAtNet.v2

a collection of pixels, $\mathbf{G} = \{g_i\}$, such that $g_i = 1$ if the corresponding pixel is known to be tumorous, and $g_i = 0$ if it is healthy.

The values of *TP*, *TN*, *FP* and *FN* used in the following loss functions were calculated in a soft-classification context by using the predicted image $\mathbf{P}$ and the GT image $\mathbf{G}$:

$$
\begin{aligned}
TP &= TP(\mathbf{P},\mathbf{G}) = \sum_{i=1}^{N} p_i g_i \\
TN &= TN(\mathbf{P},\mathbf{G}) = \sum_{i=1}^{N} (1-p_i)(1-g_i) \\
FP &= FP(\mathbf{P},\mathbf{G}) = \sum_{i=1}^{N} p_i (1-g_i) \\
FN &= FN(\mathbf{P},\mathbf{G}) = \sum_{i=1}^{N} (1-p_i) g_i
\end{aligned}
\tag{6}
$$

In this work, we have evaluated the performance of five loss functions: *Binary Cross-Entropy Loss*, *Dice Loss*, *Matthews Correlation Coefficient*, *Intersection Over Union Loss*, and *Lovász-Softmax Loss*. They are summarized below.

1. *Binary Cross-Entropy* (BCE). It is a measure that quantifies classification error. It yields significantly large values when a GT pixel is tumorous ($g_i = 1$), but the corresponding predicted pixel is healthy (p_i is close to 0), as well as when the GT pixel is healthy ($g_i = 0$) but the predicted pixel is tumorous (p_i is close to 1). The *BCE Loss* (LBCE) is obtained by averaging the BCE values calculated for all pixels:

$$
LBCE(\mathbf{P},\mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N} g_i \log p_i + (1-g_i)\log(1-p_i) \tag{7}
$$

   The BCE loss penalizes large errors between the predicted class and the true class, encouraging the model to minimize the difference and improve its binary classification performance.
2. *Dice Similarity Coefficient* (DSC), also known as F1-score. It estimates the ratio of twice the area of intersection ($2 \times TP$) to the sum of the predicted tumorous region area ($TP + FP$) and the ground truth tumorous region area ($TP + FP$):

$$
DSC(\mathbf{P},\mathbf{G}) = \frac{2TP}{2TP + FP + FN} \tag{8}
$$

   Based on it, the *Dice Loss* (LD) is defined as:

$$
LD(\mathbf{P},\mathbf{G}) = 1 - DSC(\mathbf{P},\mathbf{G}) \tag{9}
$$

3. *Matthews Correlation Coefficient* (MCC) [36]. It is commonly used to measure the quality of binary classification based on the confusion matrix. Low MCC values indicate a good agreement between the predicted tumorous regions and the GT:

$$
MCC(\mathbf{P},\mathbf{G}) = \frac{TP\ TN - FP\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}
$$

   Based on it, the *MCC Loss* (LMCC) is defined as:

$$
LMCC(\mathbf{P},\mathbf{G}) = 1 - MCC(\mathbf{P},\mathbf{G}) \tag{11}
$$

4. *Intersection over Union* (IoU), also known as Jaccard index. It is a measure of dissimilarity between predicted and GT regions in object detection and segmentation tasks:

$$
IoU(\mathbf{P},\mathbf{G}) = \frac{TP}{TP + FP + FN} \tag{12}
$$

   Based on it, the *IoU Loss* (LIoU) is defined as:

$$
LIoU(\mathbf{P},\mathbf{G}) = 1 - IoU(\mathbf{P},\mathbf{G}) \tag{13}
$$

   The IoU loss encourages better segmentation accuracy by minimizing the dissimilarity between predicted and GT regions.
5. *Lovász-Softmax Loss* [37]. It is a differentiable loss function used in semantic segmentation tasks. It combines the Lovász extension with the softmax function to directly minimize the IoU Loss in neural networks.

3.4 Evaluation measures

The proposed deep networks have been evaluated by comparing their predicted segmentation masks, $\mathbf{Y}$, against the corresponding GT, $\mathbf{G}$. The *TP*, *TN*, *FP* and *FN* measures defined in (6) were computed in a hard-classification context by using binary values, $\mathbf{Y}$, instead of probabilities, $\mathbf{P}$.

The following evaluation measures have been considered in this work:

1. *Jaccard index* or *Intersection over Union* (IoU), already defined in (12). It behaves similarly to the *DSC* (8):

$$
\text{Jaccard}(\mathbf{Y},\mathbf{G}) = IoU(\mathbf{Y},\mathbf{G}) \tag{14}
$$

2. *Dice index*. It is the Dice Similarity Coefficient (*DSC*) already defined in (8), also referred to as F1-score. It behaves similarly to the Jaccard index (14):

$$
Dice(\mathbf{Y},\mathbf{G}) = DSC(\mathbf{Y},\mathbf{G}) \tag{15}
$$

3. *Recall*. It is the percentage of truly tumorous regions detected by the network over the total number of tumorous regions in the GT.

$$\text{Recall}(\mathbf{Y}, \mathbf{G}) = \frac{TP}{TP + FN} \tag{16}$$

4. *Precision*. It is the percentage of truly tumorous regions predicted by the network over the total number of regions predicted as tumorous by the network.

$$\text{Precision}(\mathbf{Y}, \mathbf{G}) = \frac{TP}{TP + FP} \tag{17}$$

4 Experimental results

4.1 Datasets and data augmentation

Two datasets of BUS images have been utilized to evaluate the proposed network models. The first dataset (UDIAT) was provided by UDIAT Diagnostic Centre in Sabadell, Spain. It consists of 163 BUS images depicting various breast tumors. Each image is accompanied by a ground-truth segmentation of the tumor. These images were captured using a Siemens ACUSON Sequoia C512 ultrasound system equipped with an 8.5 MHz linear array ultrasound transducer. The second dataset (BUSI) was introduced in [38]. It comprises BUS images of 600 female patients. They were captured using both LOGIQ E9 and LOGIQ E9 Agile ultrasound systems. Among the 780 images available, we selected those that featured either benign or malignant tumors, resulting in a subset of 697 images. Ground-truth segmentations are also available for the tumors within this dataset. For both datasets, a split of 70% of the images was used for training, while the remaining 30% was allocated for testing. All images were resized to 256×256 to feed the evaluated models.

Augmentation algorithms must be applied to increase the reliability and robustness of deep networks. Although different augmentation techniques have been applied to overcome different issues caused by deep networks [39–41], efficient augmentation methods must be properly chosen according to the characteristics of the images. In this work, several augmentation methods have thus been chosen and applied carefully, including rotation, width and height shifts, shear, horizontal and vertical flips, shift scale rotation, median blur, hue saturation value, and random brightness contrast. Following data augmentation, the training set of the UDIAT dataset expanded to 1610 images, while the training set of the BUSI dataset increased to 6342 images.

4.2 Preliminary experiments with different loss functions

Our first research goal was to identify the most suitable loss function for this application scope. In particular, we evaluated the five loss functions summarized in Sect. 3.3 and measured their accuracy by considering the UDIAT dataset and the baseline BUS image segmentation autoencoder model depicted in Fig. 4, which applies the original CoAtNet model as encoder.

Among the tested loss functions, BCE demonstrated the highest accuracy, achieving a Dice score of 73.83%. In comparison, the alternative loss functions yielded slightly lower Dice scores: 73.56% for the Dice Loss, 72.57% for the Lovász-Softmax Loss, 71.87% for the Jaccard index (IoU Loss), and 71.43% for the MCC Loss. Based on these results, we applied BCE Loss for training the proposed network models and their evaluated alternatives with the two proposed datasets.

4.3 Experimental validation

In this study, we evaluated the performance of the two proposed Dual-Stream CoAtNet models applied to the encoder branch of the proposed autoencoder: DSCoAtNet.v1 (Fig. 5) and DSCoAtNet.v2 (Fig. 6). We compared those two models against three alternative autoencoders obtained by only changing the encoder branch using: the original CoAtNet (Fig. 4), WideResNet [42], and ResNext [43], respectively. Moreover, we also evaluated six advanced semantic segmentation networks: UNet [11], UNet++ [12], Attention-UNET [44], DoubleU-Net [32], DeepLab [45], and SegNet [46].

Table 1 Performance measures of the proposed Dual-Stream CoAtNet models and alternative modern deep networks on the UDIAT dataset

Dataset 1		
Network	Jaccard	Dice
CoAtNet (Fig. 4)	64.65	73.83
DSCoAtNet.v1 (Fig. 5)	**69.25**	**78.84**
DSCoAtNet.v2 (Fig. 6)	**68.05**	**76.61**
WideResNet	64.91	73.67
ResNext	63.26	71.51
DeepLab	52.91	63.68
Attention-UNET	63.10	72.69
SegNet	66.39	76.18
UNet	66.73	74.72
UNet++	66.22	74.86
DoubleU-Net	63.39	70.69

Tables 1 and 2 show the performance of the evaluated models for the UDIAT and BUSI datasets, respectively. Notice that the proposed Dual-Stream CoAtNet models perform significantly better than the other alternatives, including the original (single-stream) CoAtNet model, which exhibits a performance slightly superior to Wide-ResNet for the UDIAT dataset, but slightly inferior to ResNext for the BUSI dataset. Interestingly, the first version (DSCoAtNet.v1) is superior to the second (DSCoAtNet.v2) for UDIAT, and vice versa for BUSI. Therefore, the two proposed dual-stream versions are equally relevant.

In terms of qualitative results, Figs. 7 and 8 show examples of specific tumor segmentations generated by the proposed Dual-Stream CoAtNet models and the other state-of-the-art deep networks for the UDIAT and BUSI datasets, respectively. Notice that the proposed models tend to reduce the false negative regions (FN) in yellow, that is, tumorous regions wrongly detected as healthy, as well as the false positive regions (FP) in blue, that is, healthy regions wrongly detected as tumorous.

In order to assess the distribution of precision and recall for both datasets, Fig. 9 shows the boxplots of both measures corresponding to the different evaluated networks, ordered from left to right as: CoAtNet, DSCoAtNet.v1, DSCoAtNet.v2, Attention-UNET, DeepLab, SegNet, ResNext, WideResNet, UNet, UNet++, and DoubleU-Net. The behavior of recall and precision is crucial in assessing the performance of a model for detecting tumorous regions as described in Sect. 3.4. The proposed Dual-Stream CoAtNet models exhibit a consistent performance with few outliers and a small variance, indicating a limited number of false negatives and false positives in their predictions. Overall, both models exhibit strong recall and precision values, effectively detecting tumorous regions while minimizing errors.

Table 2 Performance measures of the proposed Dual-Stream CoAtNet models and alternative modern deep networks on the BUSI dataset

Dataset 2		
Network	Jaccard	Dice
CoAtNet (Fig. 4)	66.29	74.93
DSCoAtNet.v1 (Fig. 5)	**67.73**	**76.13**
DSCoAtNet.v2 (Fig. 6)	**69.22**	**77.91**
WideResNet	64.22	72.83
ResNext	67.39	75.68
DeepLab	63.59	73.22
Attention-UNET	54.62	64.46
SegNet	63.50	72.52
UNet	63.37	71.74
UNet++	63.10	71.60
DoubleU-Net	66.46	75.05

The BUSI dataset has also been used by other recent state-of-the-art tumor segmentation methods: EnGAN [20], SK-U-Net [21], LAEDNet [22], and MGCC [25]. Table 3 shows the segmentation performance of those previous models along with the best proposed dual-stream model for the BUSI dataset (DSCoAtNet.v2) in terms of Dice index. The proposed model yields the best performance: Dice = 77.91%. The latter is 4.11% larger than the performance of the closest method LAEDNet [22], whose Dice index is 73.8%.

Finally, we have applied the open-source package Ptflops [47] for estimating the computational complexity and number of trainable parameters of the original CoAtNet model and the two proposed dual-stream variations (DSCoAtNet.v1 and DSCoAtNet.2). The computational complexity is estimated in terms of GMacs (Giga Multiply-Accumulate operations). One GMac is roughly equivalent to two GFlops. Table 4 shows the results when processing 256×256 images. For comparison purposes, we have also included the two state-of-the-art models (ResNext 101_32x8d and WideResNet 101_2) considered in the experimental comparison (Tables 1 and 2), which are also used in the encoder branch, equivalently to CoAtNet and the proposed variations.

The two variations of CoAtNet have a computational complexity between 41.7% and 43.2% higher than CoAtNet, respectively, whereas the number of trainable parameters almost doubles: between 90.3% and 92% higher. However, these increases are one order of magnitude lower than the ones of the competing state-of-the-art deep models: ResNext has 374.7% more computational complexity than CoAtNet, whereas WideResNet is 555.6% higher. As for the trainable parameters, ResNext requires 398.8% more parameters than CoAtNet, whereas WideResNet has 612.8% more parameters. Notably, the two proposed dual-stream variations of CoAtNet yield significantly better results than CoAtNet and the other tested models.

4.4 Model robustness

Images are often corrupted by noise during acquisition, compression, and transmission processes. In the realm of image processing, noise refers to random variations in the signal, impacting the brightness and color of image observations and information extraction. The presence of such noise significantly hampers tasks related to image processing, including video processing, image analysis, and segmentation, potentially resulting in inaccurate diagnoses [48].

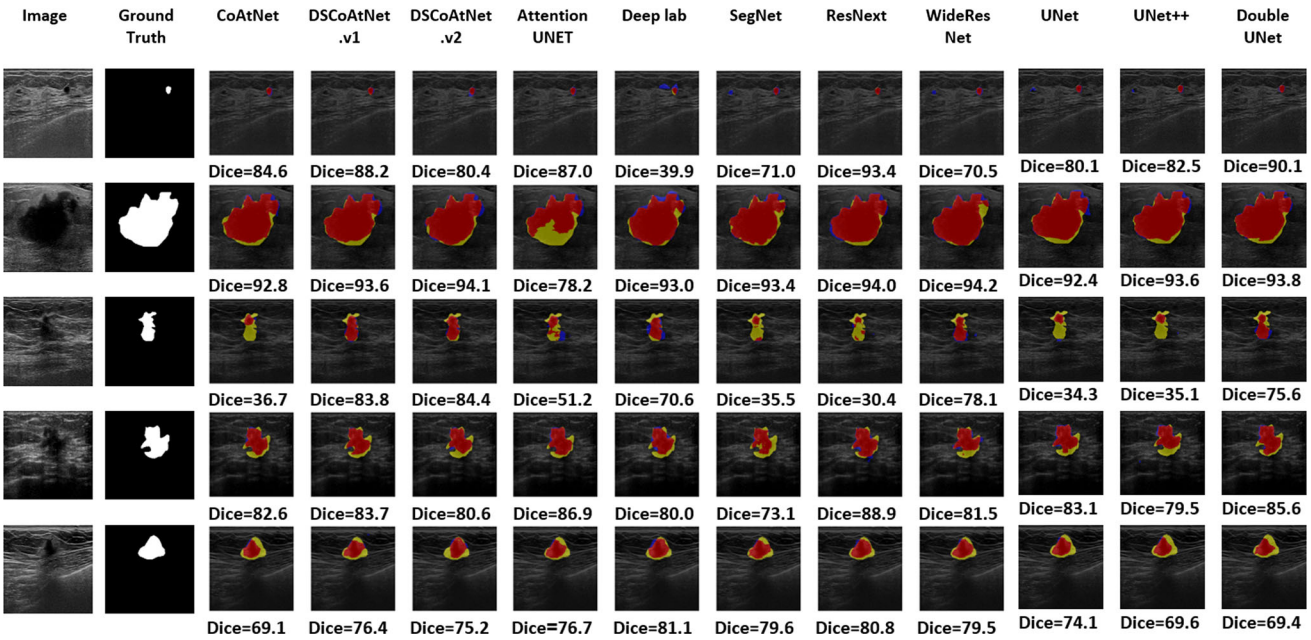

Fig. 7 Examples of tumor segmentation with the proposed Dual-Stream CoAtNet models and alternative state-of-the-art deep networks for the UDIAT dataset. Color encoding: Red (TP), Blue (FP), Yellow (FN)

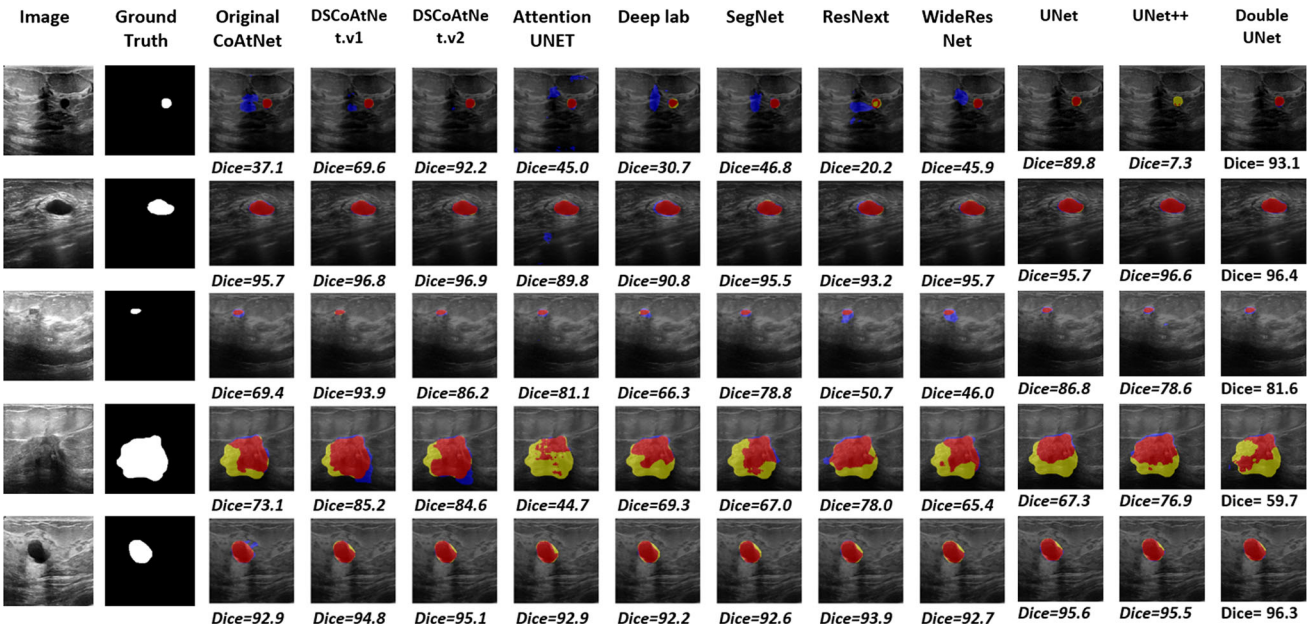

Fig. 8 Examples of tumor segmentation with the proposed Dual-Stream CoAtNet models and alternative state-of-the-art deep networks for the BUSI dataset. Color encoding: Red (TP), Blue (FP), Yellow (FN)

A significant challenge in image denoising lies in the differentiation among noise, edges, and textures, given their shared high-frequency components. The following types of noise have been extensively discussed in the literature: *Blur* [49], which is caused by factors like atmospheric noise and camera setup errors, significantly degrading image quality. *Gaussian noise* [50], which is characterized by random intensity variations following a normal distribution, adversely affecting image quality by introducing a grainy appearance. *Impulse noise* [50], also known as salt-and-pepper noise, which disrupts images with random and abrupt intensity changes, appearing as isolated bright and dark pixels and adversely affecting the overall image quality. *Poisson noise* [51], which stems

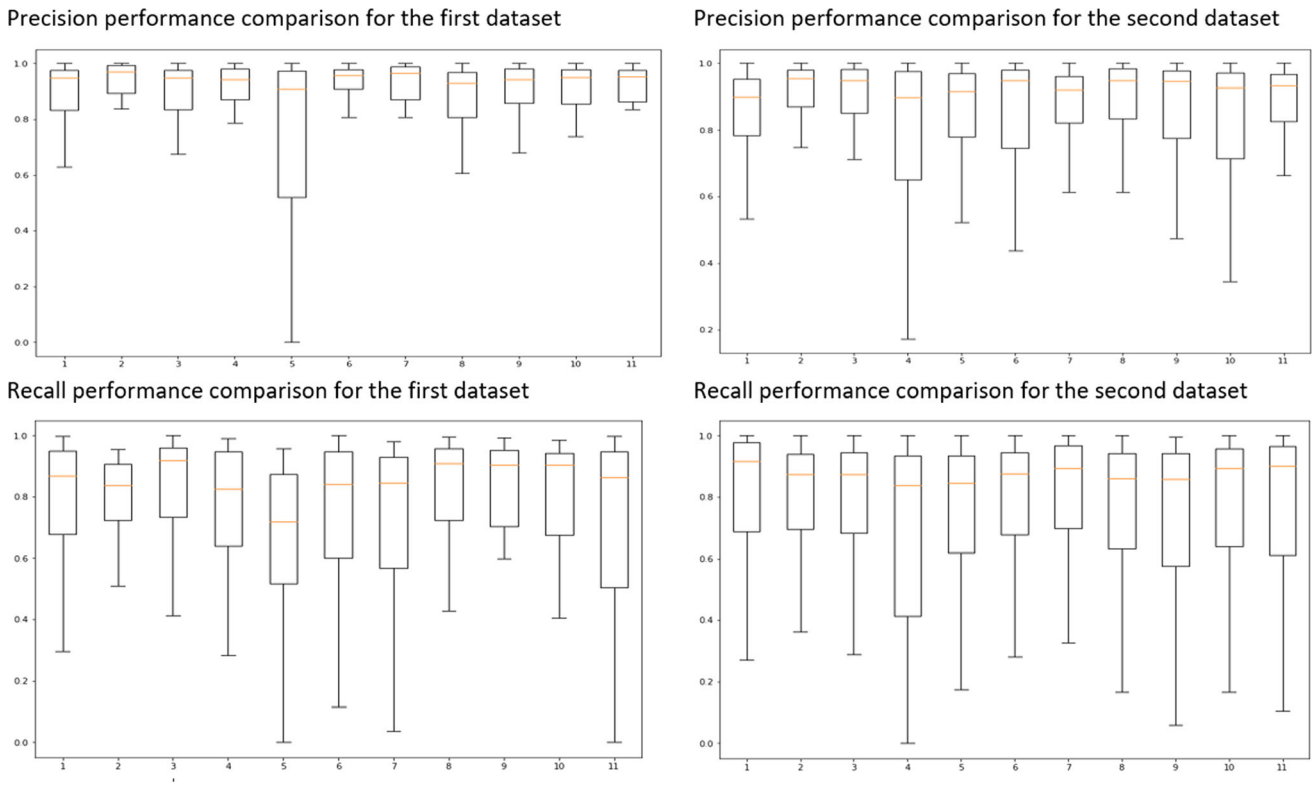

Fig. 9 Boxplot of recall and precision for the evaluated models on the two tested datasets. Ordered from left to right: CoAtNet, DSCoAtNet.v1, DSCoAtNet.v2, Attention-UNET, DeepLab, SegNet, ResNext, WideResNet, UNet, UNet++, and DoubleU-Net

Table 3 Comparison with state-of-the-art tumor segmentation models on the BUSI dataset. The proposed model applies DSCoAtNet.v2 as encoder

References	Date	Dice (%)
EnGAN [20]	Jan 2022	68.4
SK-U-Net [21]	Aug 2020	70.9
LAEDNet [22]	Apr 2022	73.8
MGCC [25]	May 2023	65.0
DSCoAtNet.v2		**77.91**

Table 4 Computational complexity and number of trainable parameters of proposed models and SoA alternatives

Model	GMacs (1 GMac ≈ 2 GFlops)	#Parameters (Millions)
CoAtNet_0	4.55	17.8
DSCoAtNet.v1	6.45 (+41.7%)	33.89 (+90.3%)
textbfDSCoAtNet.v2	6.52 (+43.2%)	34.18 (+92.0%)
ResNext 101_32x8d	21.6 (+374.7%)	88.79 (+398.8%)
WideResNet 101_2	29.83 (+555.6%)	126.89 (+612.8%)

from the probabilistic nature of photon arrival and that introduces pixel intensity variability and a grainy appearance in medical images, thus having an impact on diagnostic accuracy. *Speckle noise* [52], which is characterized by a multiplicative nature and linked to echoes, leading to the deterioration of image contrast, the distortion of brightness, and the disruption of image luminance [53, 54].

The evaluation of breast tumor segmentation models is far more exhaustive and better reflects real-world scenarios by injecting a diverse array of noise types. This ensures the model's resilience against challenges commonly encountered in medical imaging applications. To assess the robustness of the two proposed Dual-Stream CoAtNet models under demanding noise conditions, the aforementioned noise types were intentionally applied to a subset of ultrasound images from the two tested datasets. As shown in Figs. 10 and 11, the two dual-stream models consistently achieved accuracy levels that closely matched their performance on the original images, even when confronted with these challenging noise conditions. These findings strongly confirm an excellent resilience to noise and, hence, practical utility of the proposed models in real-world scenarios, as ultrasound images often exhibit noise and artifacts.

Finally, Fig. 12 zooms into one of the samples in order to show a clear picture of the different types of noise that have been tested.

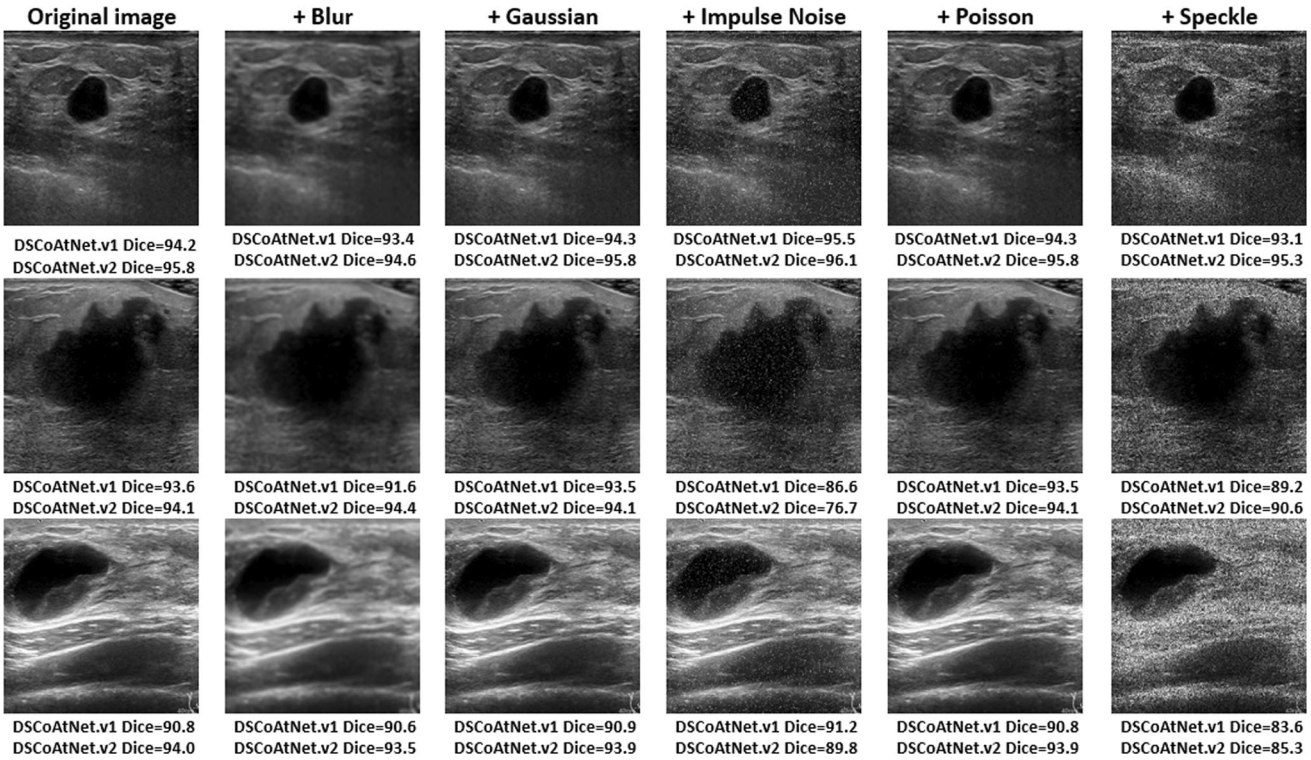

Fig. 10 Sample from the UDIAT dataset to check the robustness to noise of the two proposed Dual-Stream CoAtNet models

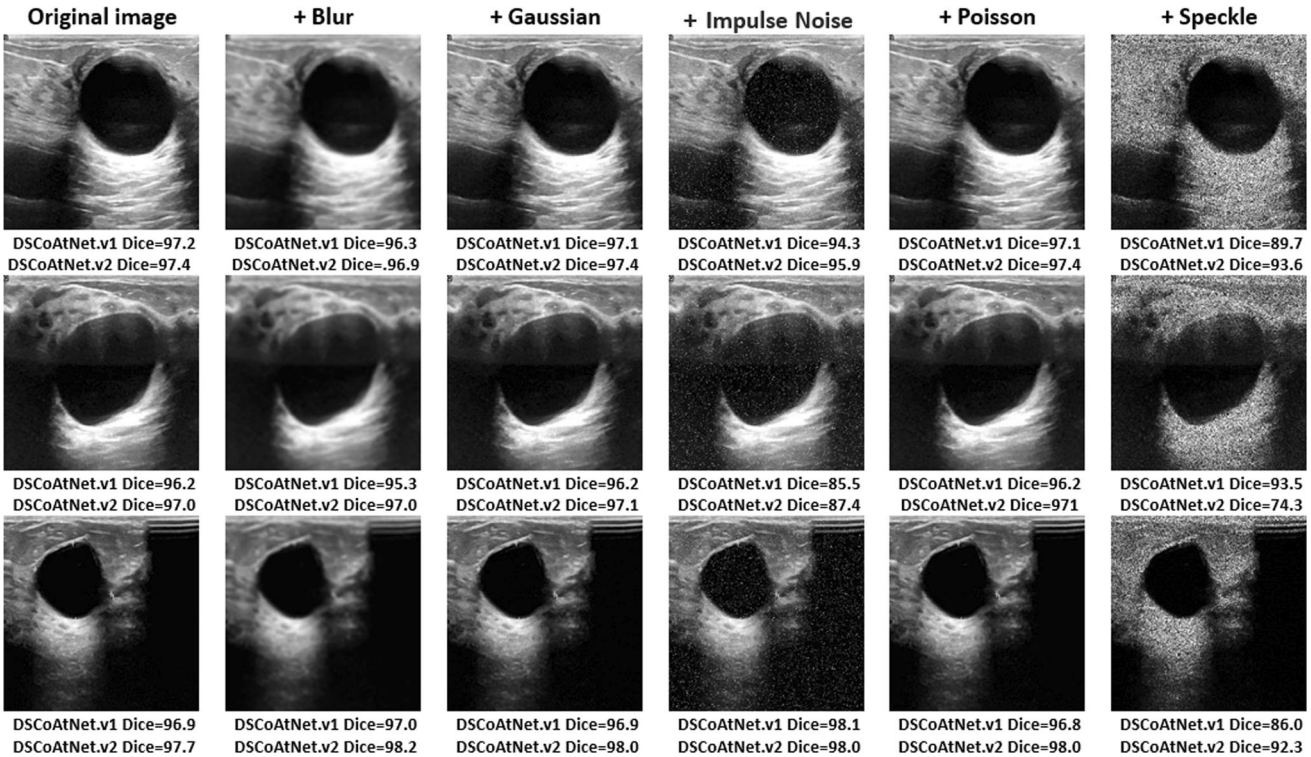

Fig. 11 Sample from the BUSI dataset to check the robustness to noise of the two proposed Dual-Stream CoAtNet models

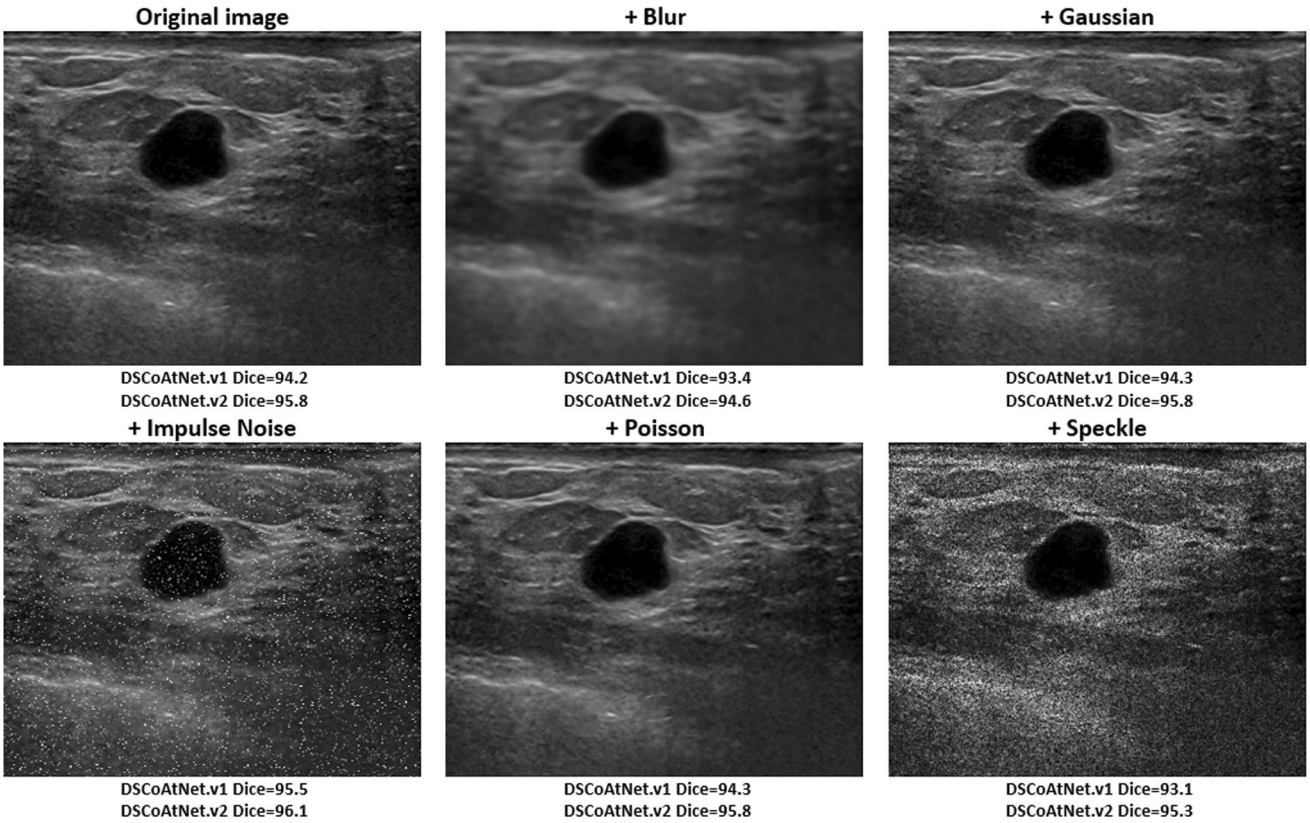

Fig. 12 Enlargement showing the different types of noise that have been tested

5 Conclusions and future work

A new approach for image segmentation of BUS images using deep neural networks has been proposed. In particular, two variations of the successful CoAtNet deep network [5] have been considered. The proposed method replaces the attention layers of CoAtNet by a dual stream of both attention and convolutional layers that are automatically integrated. The aim is that those final layers also consider local information present in the coarse-resolution feature maps, similarly to what pure convolutional neural networks do, hence complementing the global information extracted by the self-attention mechanism. Experimental results with two well-known BUS image datasets (UDIAT and BUSI) show that this Dual-Stream CoAtNet model significantly improves the segmentation accuracy of breast ultrasound images, thus contributing to the development of more robust tumor detection methods.

The two deep models proposed in this work are still far from perfect, as their best Dice performance is 78.84% for UDIAT and 77.91% for BUSI, respectively. This is their major limitation in our opinion. Therefore, it is necessary to keep exploring a variety of promising research lines as future work. Firstly, we aim to extend the capabilities of the proposed model to multi-class semantic segmentation in breast ultrasound images. We also intend to distinguish between malignant, benign, and healthy regions. To address the limited annotated datasets, we will apply pixel-wise self-supervised contrastive learning, leveraging unlabeled images for training. Additionally, we recognize the need to enhance the quality of breast ultrasound image data by using various machine and deep learning methods, such as preprocessing, noise reduction, image enhancement, and data augmentation. This will improve the accuracy and robustness of the model's segmentation performance by enhancing the clarity and resolution of the input images. Furthermore, in recent works on image segmentation and classification, hybrid loss functions applied to different network models have demonstrated their efficiency [55, 56]. Therefore, we aim to apply various hybrid loss functions to the proposed models for assessing their performance. Additionally, the performance of the proposed Dual-Stream CoAtNet model can be compared with the performance of recent capsule network-based models, which can keep spatial relationships of learned features [57–59]. Finally, two additional research lines deserve further efforts: feeding the deep models with pseudocolor images and considering level set approaches in the deep learning models.

Acknowledgements The Spanish Government partly supported this research through Project TED2021-130081B-C21 and Project PDC2022-133383-I00.

Data availability The availability of the public dataset used in this study is openly accessible through the provided source reference. For access to the private dataset utilized in this research, interested researchers can directly contact the UDIAT Diagnostic Centre as indicated in the citation.

Declarations

Conflict of interest The authors declare that they have no conflict of interest

References

1. Feng J, Polychronidis G, Heger U, Frongia G, Mehrabi A, Hoffmann K (2019) Incidence trends and survival prediction of hepatoblastoma in children: a population-based study. Cancer Commun 39:1–9
2. Huang Q, Huang Y, Luo Y, Yuan F, Li X (2020) Segmentation of breast ultrasound image with semantic classification of superpixels. Med Image Anal 61:101657
3. Guo Z, Xie J, Wan Y, Zhang M, Qiao L, Yu J, Chen S, Li B, Yao Y (2022) A review of the current state of the computer-aided diagnosis (cad) systems for breast cancer diagnosis. Open Life Sci 17:1600–1611
4. Xian M, Zhang Y, Cheng H-D, Xu F, Huang K, Zhang B, Ding J, Ning C, Wang Y (2018) A benchmark for breast ultrasound image segmentation (BUSIS), Infinite Study
5. Dai Z, Liu H, Le QV, Tan M (2021) Coatnet: Marrying convolution and attention for all data sizes, CoRR abs/2106.04803. arXiv:2106.04803
6. Göçeri E (2017) Intensity normalization in brain mr images using spatially varying distribution matching, in: International conference on computer graphics, visualization, computer vision and image processing, pp. 300–304
7. Göçeri E (2018) Fully automated and adaptive intensity normalization using statistical features for brain mr images, Celal Bayar University Journal of. Science 14:125–134
8. Hardaha S, Edla DR, Parne SR (2023) A survey on convolutional neural networks for mri analysis. Wireless Pers Commun 128:1065–1085
9. Göçeri E (2020) Convolutional neural network based desktop applications to classify dermatological diseases, in, (2020) IEEE 4th international conference on image processing, applications and systems (IPAS). IEEE 138–143
10. Idlahcen F, Idri A, Göçeri E (2024) Exploring data mining and machine learning in gynecologic oncology. Artif Intell Rev 57:20
11. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, pp. 234–241
12. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, pp. 3–11
13. Diakogiannis FI, Waldner F, Caccetta P, Wu C (2019) Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data, CoRR abs/1904.00592. arXiv:1904.00592
14. Huang K, Zhang Y, Cheng H-D, Xing P, Zhang B (2019) Fuzzy semantic segmentation of breast ultrasound image with breast anatomy constraints, arXiv preprint arXiv:1909.06645
15. Nair AA, Washington KN, Tran TD, Reiter A, Bell MAL (2020) Deep learning to obtain simultaneous image and segmentation outputs from a single input of raw ultrasound channel data. IEEE Trans Ultrason Ferroelectr Freq Control 67:2493–2509
16. Zhuang Z, Li N, Joseph Raj AN, Mahesh VG, Qiu S (2019) An rdau-net model for lesion segmentation in breast ultrasound images. PLoS ONE 14:e0221535
17. Zaidkilani N, Abdel-Nasser M, Garcia MA, Puig D (2022) Breast ultrasound cad system based on efficient tumour segmentation network and transfer-learned features, in: 2022 5th International conference on multimedia, signal processing and communication technologies (IMPACT), IEEE, pp. 1–5
18. Shareef B, Xian M, Vakanski A (2020) Stan: small tumor-aware network for breast ultrasound image segmentation, in, (2020) IEEE 17th International symposium on biomedical imaging (ISBI). IEEE 1–5
19. Vakanski A, Xian M, Freer PE (2020) Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. Ultrasound Med Biol 46:2819–2833
20. Deng E, Qin Z, Chen D, Qin Z, Ding Y, Geng J, Zhang N (2022) Engan: Enhancement generative adversarial network in medical image segmentation
21. Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M (2020) Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. Biomed Signal Process Control 61:102027
22. Zhou Q, Wang Q, Bao Y, Kong L, Jin X, Ou W (2022) Laednet: a lightweight attention encoder-decoder network for ultrasound medical image segmentation. Comput Electr Eng 99:107777
23. Xu M, Huang K, Qi X (2023) A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification. IEEE Access 11:5377–5392
24. Zhang S, Liao M, Wang J, Zhu Y, Zhang Y, Zhang J, Zheng R, Lv L, Zhu D, Chen H et al (2023) Fully automatic tumor segmentation of breast ultrasound images with deep learning. J Appl Clin Med Phys 24:e13863
25. Tang F, Ding J, Wang L, Xian M, Ning C (2023) Multi-level global context cross consistency model for semi-supervised ultrasound image segmentation with diffusion model, arXiv preprint arXiv:2305.09447
26. Ahmed S, Hasan MK (2023) Coma-net: towards generalized medical image segmentation using complementary attention guided bipolar refinement modules. Biomed Signal Process Control 86:105198
27. Ta N, Chen H, Liu X, Jin N (2023) Let-net: locally enhanced transformer network for medical image segmentation. Multimedia Syst 29:3847–3861
28. Heidari M, Kazerouni A, Soltany M, Azad R, Aghdam EK, Cohen-Adad J, Merhof D (2023) Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 6202–6212
29. Yuan F, Zhang Z, Fang Z (2023) An effective cnn and transformer complementary network for medical image segmentation. Pattern Recogn 136:109228
30. Dar MF, Ganivada A (2023) Efficientu-net: a novel deep learning method for breast tumor segmentation and classification in ultrasound images. Neural Process Lett 55:10439–10462

31. Yang L, Fan C, Lin H, Qiu Y (2023) Rema-net: an efficient multi-attention convolutional neural network for rapid skin lesion segmentation. Comput Biol Med 159:106952
32. Ahmed MR, Ashrafi AF, Ahmed RU, Shatabda S, Islam AM, Islam S (2023) Doubleu-netplus: a novel attention and context-guided dual u-net with multi-scale residual feature fusion network for semantic segmentation of medical images. Neural Comput Appl 35:14379–14401
33. Hekal AA, Elnakib A, Moustafa HE-D, Amer HM (2024) Breast cancer segmentation from ultrasound images using deep dual-decoder technology with attention network, IEEE Access
34. Zhang H, Lian J, Yi Z, Wu R, Lu X, Ma P, Ma Y (2024) Hau-net: hybrid cnn-transformer for breast ultrasound image segmentation. Biomed Signal Process Control 87:105427
35. Üzen H (2024) Convmixer-based encoder and classification-based decoder architecture for breast lesion segmentation in ultrasound images. Biomed Signal Process Control 89:105707
36. Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 21:1–13
37. Berman M, Triki AR, Blaschko MB (2018) The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4413–4421
38. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020) Dataset of breast ultrasound images. Data Brief 28:104
39. Göçeri E (2023) Medical image data augmentation: techniques, comparisons and interpretations. Artif Intell Rev 56:12561–12605
40. Göçeri E (2023) Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets. Int J Imaging Syst Technol 33:1727–1744
41. Göçeri E (2020) Image augmentation for deep learning based lesion classification from skin images, in, (2020) IEEE 4th International conference on image processing, applications and systems (IPAS). IEEE 144–148
42. Zagoruyko S, Komodakis N (2016) Wide residual networks, arXiv preprint arXiv:1605.07146
43. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500
44. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al (2018) Attention u-net: learning where to look for the pancreas, arXiv preprint arXiv:1804.03999
45. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587
46. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39:2481–2495
47. Sovrasov V (2019) Flops counter for convolutional networks in pytorch framework. https://github.com/sovrasov/flops-counter.pytorch/
48. Göçeri E (2023) Evaluation of denoising techniques to remove speckle and gaussian noise from dermoscopy images. Comput Biol Med 152:106474
49. Muthana R, Alshareefi AN (2020) Techniques in de-blurring image, in: Journal of physics: conference series, volume 1530, IOP Publishing, p. 012115
50. Awad A (2019) Denoising images corrupted with impulse, gaussian, or a mixture of impulse and gaussian noise. Eng Sci Technol Int J 22:746–753
51. Rajagopal A, Hamilton RB, Scalzo F (2016) Noise reduction in intracranial pressure signal using causal shape manifolds. Biomed Signal Process Control 28:19–26
52. Ilesanmi AE, Idowu OP, Chaumrattanakul U, Makhanov SS (2021) Multiscale hybrid algorithm for pre-processing of ultrasound images. Biomed Signal Process Control 66:102396
53. Hooi FM, Kripfgans O, Carson PL (2016) Acoustic attenuation imaging of tissue bulk properties with a priori information. J Acoust Soc Am 140:2113–2122
54. Biswas B, Sen BK, Dey KN (2018) Ultrasound medical image deblurring and denoising method using variational model on cuda. Adv Comput Syst Secur 5:95–108
55. Göçeri E (2024) Polyp segmentation using a hybrid vision transformer and a hybrid loss function, J Imaging Inform Med 1–13
56. Göçeri E (2021) An application for automated diagnosis of facial dermatological diseases. İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi 6:91–99
57. Göçeri E (2023) Classification of skin cancer using adjustable and fully convolutional capsule layers. Biomed Signal Process Control 85:104949
58. Göçeri E (2021) Analysis of capsule networks for image classification, in: International conference on computer graphics, visualization, computer vision and image processing
59. Göçeri E (2021) Capsule neural networks in classification of skin lesions, in: International conference on computer graphics, visualization, computer vision and image processing, pp. 29–36