Local and Context-Attention Adaptive LCA-Net for Thyroid Nodule Segmentation in Ultrasound Images

Zhen Tao, Hua Dang, Yueting Shi, Weijiang Wang, Xiaohua Wang and Shiwei Ren

Abstract: The thyroid nodule segmentation of ultrasound images is a critical step for the early diagnosis of thyroid cancers in clinics. Due to the weak edge of ultrasound images and the complexity of thyroid tissue structure, it is still challenging to accurately segment the delicate contour of thyroid nodules. A local and context-attention adaptive network (LCA-Net) for thyroid nodule segmentation is proposed to address these shortcomings, which leverages both local feature information from convolution neural networks and global context information from transformers. Firstly, since most existing thyroid nodule segmentation models are skilled at local detail features and lose some context information, we propose a transformers-based context-attention module to capture more global associative information for the network and perceive the edge information of the nodule contour. Secondly, a backbone module with $7 \times 1, 1 \times 7$ convolutions and the activation function Mish is designed, which enlarges the receptive field and extracts more feature details. Furthermore, a nodule adaptive convolution (NAC) module is introduced to adaptively deal with thyroid nodules of different sizes and positions, thereby improving the generalization performance of the model. Simultaneously, an optimized loss function is proposed to solve the pixels class imbalance problem in segmentation. The proposed LCA-Net, validated on the public TN-SCUI2020 and TN3K datasets, achieves Dice scores of 90.26% and 82.08% and PA scores of 98.87% and 96.97%, respectively, which outperforms other state-of-the-art thyroid nodule segmentation models. This paper demonstrates the superiority of the proposed LCA-Net for thyroid nodule segmentation, which possesses strong generalization performance and promising segmentation accuracy. Consequently, the proposed model has wide application prospects for thyroid nodule diagnosis in clinics.

Keywords: thyroid nodule segmentation; transformers; local details; ultrasound images; computer-aided diagnosis

1. Introduction

The thyroid gland is honored as the ‘life gland’ of human beings, which is an essential endocrine organ in the body [1]. The thyroid regulates the body’s metabolism, growth and development by secreting thyroid hormones. Thyroid nodules are irregular lumps due to lesions in the thyroid gland, which are considered to be the main clinical manifestations of thyroid abnormalities [1–3]. Consequently, the diagnosis of thyroid lesions through thyroid nodules has become the most common diagnostic method. For the evaluation and diagnosis of thyroid nodules [3], ultrasound technology has become the optimal choice to provide nodules’ information owing to its non-invasive, practical and painless characteristics [4,5]. The thyroid nodule segmentation of ultrasound images is an indispensable step for the analysis of nodules’ characteristics and a prerequisite for accurate diagnosis [6]. Generally, radiologists need to carefully read ultrasound images and manually draw the edge contour.
of the nodules. However, the manual segmentation of a nodule contour is time-consuming and tedious. Meanwhile, on account of the intrinsic characteristics of ultrasound images, such as weak edges, heterogeneity [5] and low contrast [2], it is challenging to segment the delicate contours of thyroid nodules from ultrasound images. In order to mitigate these issues, several computer-aided diagnosis (CAD) systems are employed to precisely and automatically segment nodules to assist radiologists.

Before the rise of deep learning, the thyroid nodule segmentation of ultrasound images relied on conventional image processing technologies, which could be roughly divided into three categories [2]: contour and shape-based methods, region-based methods, and traditional machine learning methods [5,6]. Du et al. [7] proposed a novel DRLSE algorithm based on the contour and shape to segment thyroid nodules in ultrasound images. This method overcame a certain amount of ultrasonic image noise and nodule echogenicity and prevented the leakage of some boundaries of nodules. However, the DRLSE algorithm performed poorly in the segmentation of weak edges and was sensitive to the locations of manually initialized contours [7]. Zhao et al. [8] proposed a normalized cut method based on regions which could segment the thyroid nodules and the trachea areas in ultrasound images. The algorithm reduced the number of parameters, but the fineness of segmentation still needed to be improved. Keramidas et al. [9] utilized a fuzzy grey-level histogram for feature extraction and an SVM algorithm for pixels classification in order to segment the thyroid nodules. Nevertheless, this algorithm was sensitive to the selected features, and the segmentation accuracy remained to be improved. In summary, these conventional thyroid nodule segmentation methods cannot overcome the characteristics of ultrasound and thyroid tissue well and the segmentation accuracy still needs improvement.

With the remarkable development of deep learning, it has been applied to medical image diagnosis and provides auxiliary diagnosis advice in clinics [10]. In the field of thyroid nodules segmentation, models based on deep learning have significantly outperformed traditional algorithms [1,2,5]. Based on supervised learning, deep learning surmounts the inherent characteristics of ultrasound images and could distinguish the contour boundary between nodules and other thyroid tissues. In 2017, Ma et al. [11] first applied deep convolutional networks to the thyroid nodule segmentation of ultrasound images. This overcame the shortcomings of conventional methods such as manually extracting nodule features and improved the efficiency of automatic segmentation. However, this method lost plenty of feature details, leading to less-than-ideal segmentation results. In 2020, Yang et al. [5] adopted a modified UNet for the initial segmentation and then a level set algorithm based on the image feature direction to refine the segmentation performance. However, due to the inherent properties of convolution, such as gathering information only from neighborhood pixels [12], this network could not explicitly capture global dependency [13]. In 2021, Gong et al. [6] proposed a thyroid region prior guided feature enhancement model for thyroid nodule segmentation. This model could segment the thyroid gland and its internal nodules simultaneously, making it easier for radiologists to compare the tissues around the nodules for subsequent diagnosis. However, this network pays too much attention to global information but ignores the local feature details. Recently, Zhang et al. [3] reported segmenting thyroid nodules with a cascade UNet framework. The network utilized the first UNet to roughly locate the nodule position and employed the second UNet to perform fine segmentation to obtain the final result. Nevertheless, it is not robust to thyroid nodules of diverse sizes and locations. Given the original convolution networks capturing multi-scale features [12] in the limited scaling range by employing the $3 \times 3$ convolution and pooling operations [14,15] and lacking the ability to explicitly capture global dependency [13], it is still challenging to achieve high accuracy in the segmentation of thyroid nodules. It can be seen from the above that due to the various and complex shapes of thyroid nodules in the actual clinics, the existing thyroid nodule segmentation networks could not satisfactorily segment the contours of the nodules.

In this paper, a local and context-attention adaptive network (LCA-Net) is proposed to further overcome the shortcomings of the current thyroid nodule segmentation methods.
This model could more accurately and automatically depict the contour of nodules and even better assist radiologists in the subsequent diagnosis of clinics. Considering the characteristic of convolutional neural networks focusing on local details and the merits of transformers in capturing global context information, this paper mainly employs CNNs and transformers to design the whole novel end-to-end network structure. Specifically, the backbone modules based on CNNs and the context-attention modules based on transformers designed in this paper are utilized to extract local feature details and global associative information. Simultaneously, a nodules adaptive convolutions (NAC) module is embedded into the encoder of the network, which assists the network to have promising adaptability for various sizes and positions of nodules. Furthermore, an optimized loss function is proposed to solve the pixels class imbalance problem in segmentation. Experiments are performed on the public TN-SCUI2020 dataset and TN3K dataset to evaluate our method, and the results are superior to the current thyroid nodule segmentation models.

The main contributions of our algorithm are five-fold:

- A novel local and context-attention adaptive network (LCA-Net) is proposed for thyroid nodule segmentation in ultrasound images. Compared with the existing state-of-the-art methods of thyroid nodule segmentation, the LCA-Net shows promising segmentation accuracy and strong generalization performance.
- A context-attention module based on transformers is proposed to gain global associative information to the model.
- A backbone module with $7 \times 1, 1 \times 7$ convolutions and activation function Mish is designed to more finely extract the local feature details of nodules.
- The nodules adaptive convolutions (NAC) module is embedded into the encoder of the LCA-Net to enhance the robustness of the network to various sizes and positions of nodules.
- A novel loss function is proposed to solve the pixels’ class imbalance problem for thyroid nodule segmentation.

The rest of the paper is organized as follows. In Section 2, we provide the detailed description of our proposed method, introducing the overall architecture, each sub-module design and a novel loss function. In Section 3, the performance and robustness of the proposed method are evaluated over the two public thyroid nodule datasets using the standard performance metrics and the results are compared with the previous state-of-the-art works. Finally, Section 4 concludes the paper and suggests topics for future work.

2. Materials and Methods

This paper aimed to establish a high-accuracy and strong generalization thyroid nodule segmentation model to assist radiologists in clinical diagnosis. In this section, we propose a novel thyroid nodule segmentation scheme. Firstly, we introduce the proposed LCA-Net architecture which includes backbone modules, context-attention modules and a nodules adaptive convolution module. Afterward, we present the details of the context-attention module, the backbone module, the nodules adaptive convolution module and a novel loss function, respectively.

2.1. LCA-Net Architecture

Currently, existing thyroid nodule segmentation models in the literature are often not comprehensive enough in design, which ignores the global information capture ability or the generalization performance of the practical clinical application. Considering the weak edge and low contrast of thyroid nodules in ultrasound images, it is easy to misdiagnose using convolution neural networks to extract the local features of thyroid nodules. Hence, we designed a novel thyroid nodule segmentation model called LCA-Net from the perspective of taking global context, local details and generalization performance into account. Figure 1 shows the specific structure of the LCA-Net. This network based on the encoder–decoder framework is mainly composed of backbone modules and context-attention modules. Since the convolution neural networks are experts in extracting local
detail features, our backbone module primarily consists of two consecutive $7 \times 1$ and $1 \times 7$ large convolution kernels, which are responsible for extracting more detailed edge features of nodules. However, the convolution kernel has the inherent characteristics of excessively paying attention to the detailed features of nodules and the middle part of feature maps. It is easy to lose some significant features only by using convolution neural networks to segment the contour of nodules. In actual clinical ultrasound images, the contour of the thyroid nodule is generally located at the edge of the region of interest [15], and convolution neural networks could not be fully considered. In view of the ability of Transformers to capture high-resolution context information in image segmentation, we designed a context-attention module based on transformers. Consequently, this module was integrated into the network, making it feasible for the LCA-Net to precisely segment the edge contours. Instead of naively integrating the Context-attention modules on top of the feature maps from the CNN backbone, we embedded the context-attention modules into each level of the encoder and decoder to collect long-range dependency from multiple scales. Moreover, the skip connection enables the context-attention module to capture more global features through the fusion of low-level features and high-level features. This framework enables each network encoding and decoding to give the extraction of local features and the context association information of ultrasound images into consideration to more finely perceive the contour information of thyroid nodules. It is worth noting that the LCA-Net employs the convolution kernels of $3 \times 3$ for the original image resolution because the input and output ports of the network pay more attention to the detailed textures. Therefore, simply embedding the backbone module or the context-attention module will affect the performance of the whole model and increase additional computation. Finally, we embedded the nodule adaptive convolution (NAC) module in the last part of the network encoder, which enables the network to adaptively deal with the thyroid nodules of various sizes and positions without additional computation [16,17]. The integration of this module also improves the generalization performance of the whole model in practical clinical application.

![Figure 1](image_url)

**Figure 1.** Illustration of the proposed local and context-attention adaptive network (LCA-Net) for the thyroid nodule segmentation of ultrasound images. The LCA-Net has the structure of an encoder and decoder. It mainly contains the backbone modules, context-attention modules and a nodule adaptive convolutions module, which could extract local details and global context information. See the text for more details.

### 2.2. Context-Attention Module

Figure 1 highlights the locations of the context-attention modules. It is worth noting that we apply the context-attention modules to each level of the encoder and decoder to capture global association information from multiscale feature maps. In clinical practice, the edge contour of the thyroid nodule is often distributed at the boundary of ultrasound
images. In this case, barely utilizing convolution neural networks cannot finely segment the nodule contour. In view of the outstanding performance of transformers in capturing the global pixel of images, we established a context-attention module based on transformers. The specific structure of the context-attention module is shown in Figure 2. However, we cannot directly capture the global information of the nodule feature map with the classical vision transformers [14,15]. Due to a large number of pixels in the ultrasound image, employing the classical vision transformer easily leads to problems such as excessive network calculation, over-fitting and challenging training. In order to alleviate the above problems and make the model aware of the global context, we preprocessed the nodule feature map before inputting the transformer. As seen in Figure 3a,b, firstly, three $1 \times 1$ convolution kernels are exploited to project input to query, key and value embeddings. Hence, the dimensions of query, key and value input into the module are: $Q, K, V \in \mathbb{R}^{C_i \times H_1 \times W_1}$, where $C_i$ denotes the dimension of embedding in self-attention mechanism, $H_1$, $W_1$ are the spatial height and width. Afterward, we divide the query, key and value embeddings into a fixed number of small patches instead of simple pixels. Specifically, the window size of the query patch is set to $8 \times 8$, while the size of the key-value patch is set to $10 \times 10$ and crossed. This leads to the feature size of the query of $Q \in \mathbb{R}^{N_1 \times C_i \times 8 \times 8}$, where the total number patches of the query is $N_1 = \frac{H_1}{8} \times \frac{W_1}{8}$. Similarly, the feature size of the key-value is $K, V \in \mathbb{R}^{N_2 \times C_i \times 10 \times 10}$, and the number patches of the key-value is $N_2 = (\frac{H_1-10}{8}+1) \times (\frac{W_1-10}{8}+1)$, where $s$ is the step size of moving the key-value patches. Notably, the query is different from the key-value patch sizes, and the key-value patch is overlapped, which is larger than the query patch. Extensive experiments show that keeping larger and overlapping key-value patches leads to significant performance gains due to the small information exchange between two neighborhood windows. This design aims to save the amount of computation and maximize the transformer’s ability to capture contextual information as much as possible. Additionally, the function permute is used to transform the dimensions of $Q, K$ and $V$ to facilitate subsequent matrix operations. At this point, the dimensions of $\hat{Q}, \hat{K}$ and $\hat{V}$ after transformation are: $\hat{Q} \in \mathbb{R}^{C_i \times M_1}$, $\hat{K}, \hat{V} \in \mathbb{R}^{C_i \times M_2}$, where $M_1 = N_1 \times 8 \times 8$ and $M_2 = N_2 \times 10 \times 10$. Simultaneously, the input of the transformer as the decoder is different from that of the encoder. In decoders, the query input comes from the skip connections of the encoder, while the key-value utilizes the output of the previous module as the input. Through the fusion of low-level features and high-level features, the context-attention module retains more global context information. Furthermore, considering the fact that ultrasound images are highly structured, we also introduce relative position index [15,18]. By recording the relative position between patches, the performance of the module is further promoted. Therefore, the proposed self-attention is defined as:

$$Attention(\hat{Q}, \hat{K}, \hat{V}) = \text{softmax}(\frac{\hat{Q} \otimes \hat{K}^T + \varphi}{\sqrt{C_i}}) \otimes \hat{V}, \quad (1)$$

where $\varphi \in \mathbb{R}^{M_1 \times M_2}$ denotes relative position bias and $\otimes$ denotes matrix multiplication.

The proposed self-attention takes into account the information and position of all pixels in the feature map, so it can extract more global context features. We define the input feature map of the context-attention module as $X_1 \in \mathbb{R}^{C_i \times H_1 \times W_1}$, where $C_i$, $H_1$, $W_1$ represent the channel number, the height and the width of the feature map. First, the channel number of $X_1$ is changed by $1 \times 1$ convolution operation, and $X_1$ is, respectively, split into a sequence of patches and transformed the dimensions: $X_1 = [x_1, ..., x_{N_1}] \in \mathbb{R}^{C_i \times (N_1 \times 8 \times 8)}$ for $\hat{Q}$, $X_1^* = [x_1^*, ..., x_{N_1}^*] \in \mathbb{R}^{C_i \times (N_2 \times 10 \times 10)}$ for $\hat{K}$, and $X_1^* = [x_1^*, ..., x_{N_2}^*] \in \mathbb{R}^{C_i \times (N_2 \times 10 \times 10)}$ for $\hat{V}$, where $N_1, N_2$ are the number of patches. The specific calculation process is shown in the following formula:

$$X_1, X_1^*, X_1^* = \Phi_{qkv}[f_{ps}(g(LN(X_1)))]$$

(2)
Figure 2. The specific structure of the context-attention module, where LN is the layer normalization operator, and $C_1 \times H_1 \times W_1$ represents the map shape (channel, height and width).

Figure 3. The proposed novel self-attention, where $N_1$, $N_2$ are the patch number, $\otimes$ is matrix multiplication and $\oplus$ denotes concatenation operation. (a,b) denote the encoder and decoder of self-attention, respectively. They share similar concepts. Significantly, the decoder of the self-attention takes high-resolution features from the encoder for query, and low-resolution features from the decoder for key and value.

We denote $(\cdot)$ as the dot-product operation, where $LN(\cdot)$ denotes the layer normalization operator, $g(\cdot)$ is the $1 \times 1$ convolution layer, $f_{ps}(\cdot)$ is the patches split operation, and $\Phi_{glo}(\cdot)$ is the permute function. Subsequently, we input $X_1, X'_1, X''_1$ into the self-attention to capture more global context information, and the specific process is summarized as follows:

$$Y'_1 = g[Attention(X_1, X'_1, X''_1)] \oplus X_1 \in \mathbb{R}^{C_1 \times H_1 \times W_1}. \quad (3)$$

Furthermore, the feature map passes through the rest of the module, and the calculation process can be defined as:

$$Y_1 = g[LN(Y'_1)] \oplus Y'_1 \in \mathbb{R}^{C_1 \times H_1 \times W_1}, \quad (4)$$
where $\oplus$ denotes the concatenation operation and $g(\cdot)$ represents the $1 \times 1$ convolution layer.

### 2.3. Backbone Module

As shown in Figure 1, similarly to the locations of context-attention modules, we embed the backbone module to each level of the encoder and decoder to extract local feature details from multiple scales. The critical technology of thyroid nodule segmentation depends on extracting the detailed features of nodule contour. Due to the calculation principle of the sliding window, the convolution neural networks pay more attention to the information in the sliding window, that is, the local details. Therefore, we establish the backbone module based on CNNs. The specific structure of the backbone module is shown in Figure 4.

**Figure 4.** The specific structure of the backbone module, where BN denotes the batch normalization operator.

In general, because the smaller convolution kernel has merits such as better nonlinear expression and fewer parameters, researchers use it to build deep neural networks. Nevertheless, the smaller convolution kernel will also lead to the incomplete extraction of local information, resulting in the loss of thyroid nodule contour information. Simultaneously, the application of pooling layers further leads to the reduction in the image resolution and the loss of spatial feature information [17,19]. To alleviate these problems, we adopt a large convolution kernel with a size set of 7. This design not only enlarges the receptive field of the network and extracts more local detailed information but also offsets the loss of feature information as much as possible. Considering the shortcomings of the large convolution kernel, such as too large parameters and easy over-fitting, we employ two consecutive $7 \times 1$ and $1 \times 7$ convolution layers. Extensive experiments show that the effect of extracting nodule features from a $7 \times 7$ large convolution kernel is essentially the same as that of two consecutive convolutions with the size of 7. This design makes the network end up with more disentangled parameters and therefore with faster training [20]. At this point, the computational parameters of the module are reduced by approximately 72%. Consequently, the convolution layers of the module have better local feature extraction performance and fewer parameters.

As is known to all, the activation function is nonlinear, which improves the expression ability of the neural network to the model. Due to its fast convergence speed, no gradient saturation and gradient disappearance, the ReLU function has become the mainstream neural network activation function. Meanwhile, this function also has shortcomings such as neuron necrosis and not being smooth enough. Generally speaking, the smoother activation function can better transfer information and improve the accuracy and generalization ability of the model. Considering the fact that the loss of information should be avoided as much as
possible when extracting the detailed features of nodules, an advanced activation function Mish [21] is introduced in this paper:

\[ y = x \cdot \tanh(\ln(1 + e^x)) \]  

(5)

As seen in Figure 5, compared with the widely adopted activation function ReLU, the Mish function is smoother, so it shows more significant superiority in transmitting information and anti-interference.

![Figure 5. The activation function diagram of Mish and ReLU.](image)

The remaining two 1 × 1 convolution kernels in the backbone module are mainly used to transform feature map channels. We denote the input and output of the module as \( X_2, Y_2 \in \mathbb{R}^{C_2 \times H_2 \times W_2} \), where \( C_2, H_2 \) and \( W_2 \) represents the channel number for feature map, the length and width of the feature map. In the backbone module, the local detail features of the feature map are first extracted by two large convolution kernels as shown in Formula (5):

\[ Y'_2 = f_2[f_1(X_2)] \in \mathbb{R}^{C_2 \times H_2 \times W_2}, \]  

(6)

where \( f_1(\cdot), f_2(\cdot) \), respectively, represent two consecutive 7 × 1 and 1 × 7 convolution layers. Afterward, we utilize two 1 × 1 convolution kernels to change the number of feature map channels. The specific calculation process is shown in the following formula:

\[ Y_2 = g[\sigma(g(BN(Y'_2)))) \oplus X_2, \]  

(7)

where \( g(\cdot) \) represents the 1 × 1 convolution layer, \( BN(\cdot) \) is the batch normalization operator, \( \sigma(\cdot) \) denotes the activation function Mish and \( \oplus \) is concatenation operation.

2.4. Nodule Adaptive Convolution Module

In clinical practice, the size and location of thyroid nodules are various. Hence, it is difficult to merely match the variable nodules with the convolution kernel of fixed size [16]. The large convolution kernel could often extract the features of large or medium position objects, while the small convolution kernel is better for small or remote position targets. Therefore, it is of great significance to flexibly adjust the appropriate convolution according to the size of the nodule. In order to better adapt to the practical clinical application, we designed a nodule adaptive convolution (NAC) module to match various thyroid nodules. The specific structure of the module is shown in Figure 6. The main body of the module is the input preprocessing and aggregation of multiple convolution kernels of different sizes. The preprocessing part dynamically adjusts the weights of each convolution kernel according to the input, thereby generating adaptive dynamic convolutions. Simultaneously,
Dilated convolution is widely applied for neural network optimization because it enlarges the receptive field without adding additional parameters. As illustrated in Figure 6, we introduced four different ratios of dilated convolutions [17,19], from 1 to 2, 3 and 5, then the receptive field of each branch is 1, 5, 7 and 11. After four multi-scale feature maps are achieved, they will be spliced. This design could deal with nodules of different sizes and locations in actual clinical practice and improve the generalization ability of the model.

**Figure 6.** The specific structure of nodule adaptive convolution module. FC denotes the fully connected layer, and $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\alpha_4$ represent the adaptive ratios of each convolution when dealing with different nodules.

Let us denote the input and output of the module as $X_3$, $Y_3 \in \mathbb{R}^{C_3 \times H_3 \times W_3}$, where $C_3$ is the number of feature channels, $H_3$ and $W_3$ are the length and width of the feature map. We define the calculation process of the NAC module by aggregating multiple dilated convolutions with different weights as follows:

$$Y'_3 = \sum_{k=1}^{K} \alpha_k(X_3) \cdot \sigma[BN(f_k(X_3))],$$

$$Y_3 = \sigma[BN(g(Y'_3))],$$

$$0 \leq \alpha_k(X_3) \leq 1, \sum_{k=1}^{K} \alpha_k(X_3) = 1,$$

where $f_k(\cdot)$, $\alpha_k$ represent each dilated convolution and its assigned weight, respectively, $\sigma(\cdot)$ is the Mish activation function, $BN(\cdot)$ is the Batch normalization operator and $g(\cdot)$ denotes a $1 \times 1$ convolution layer which is used to adjust the number of channels $C_3$.

### 2.5. Loss Function for Pixels Class Imbalance

In actual ultrasound images, the thyroid nodule occupies only a small part of the image, while the rest are mostly background. This can easily cause pixel class imbalance, which affects the segmentation performance of the model. Therefore, if the conventional dice loss or cross-entropy loss is directly used to optimize the network parameters, the prediction segmentation maps in the test stage are not ideal. Motivated by focal loss structure [22], an optimized loss function is devised to overcome this problem. First, we introduce a
weight factor \((1 - p_{x,y,c})^2\) into the cross-entropy loss, where \(p_{x,y,c}\) represents the prediction probability of the real value at pixel \((x, y)\), and \(c\) is the pixel class. Hence, the prediction probability of misclassified pixels is \((1 - p_{x,y,c})\). At this point, the value of the weighed factor is large, which is equivalent to increasing the weight of misclassified pixels in the loss function. Nevertheless, with the progress of the training process, if the model forever pays attention to the misclassified pixels, it will affect the accuracy of pixel classification. In order to further develop the performance of the loss function, we add \(p_{x,y,c}^2\) as the fine-tuning of the function. As a consequence, we redefine the loss function \(L_{LCA}\) as follows:

\[
L_{LCA} = -\frac{1}{HW} \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \sum_{c=0}^{1} R_{x,y,c} \log \left( \frac{(1 - p_{x,y,c})^2}{2} + p_{x,y,c}^2 \right),
\]

where \(H, W\) are the height and width of map, \(c\) denotes the pixel class and \(R_{x,y,c}\) represents the real value of pixels.

3. Experiments and Results

3.1. Datasets

In this paper, in order to realize the function of the LCA-Net, we needed to utilize the suitable thyroid nodule datasets of ultrasound images to train and test the model. The TN-SCUI2020 [2,23] and TN3k [5] datasets come from different institutions with different data sources and dataset sizes. Therefore, using these two datasets to verify the model performance can also reflect the generalization ability of the model.

3.1.1. TN-SCUI2020

The TN-SCUI2020 dataset comes from the challenge of Thyroid Nodule Segmentation and Classification in Ultrasound Images, which provides a public 2D dataset of a thyroid nodules with approximately 7288 ultrasound images. These images were collected from different ages, genders and in different sites using various ultrasound machines such as Mindray DC-8, Philips-cx50 and TOSHIBA Aploio300 [23]. Each image is provided with a detailed delineation of the nodule. For preprocessing, we crop the thyroid nodule area in ultrasound images and resize the areas to 256 \(\times\) 256 in the dimension. The dataset was randomly split into a training set (5830 images), a verification set (729 images) and a test set (729 images) in a ratio of 8:1:1.

3.1.2. TN3k

In order to facilitate the development of thyroid nodule segmentation, Gong et al. [5] contributed an open access dataset of thyroid nodule images called TN3k. This dataset includes 3493 ultrasound images taken from 2421 patients, and each patient retains only one ultrasound image of a case. Each image is converted into grayscale and corresponds to a high-resolution mask labeling. Compared with the TN-SCUI2020 dataset, this dataset has fewer data and lower ultrasound image quality. All images are randomly cropped to 256 \(\times\) 256 before entering the models. To verify the performance of the network, we divided the TN3k dataset into the training set (2879 images) and test set (614 images).

3.2. Evaluation Metrics

Different metrics are employed to test the results of the proposed model. The performance of the thyroid nodule segmentation will be evaluated by dice, Jaccard, recall and precision, which are defined as:

\[
\text{Dice}(\%) = \frac{2TP}{2TP + FP + FN} \times 100
\]

\[
\text{PA}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100
\]
Jaccard(%) = \frac{TP}{TP + FP + FN} \times 100 \quad (14)

Recall = \frac{TP}{TP + FN} \quad (15)

Precision = \frac{TP}{TP + FP} \quad (16)

FPR = \frac{FP}{FP + TN} \quad (17)

FNR = \frac{FN}{TP + FN} \quad (18)

where TP, FP, FN and TN, respectively, represent the true positive, false positive, false negative and true negative, which are commonly used in image processing. Dice is also called the dice similarity coefficient, which can be used to compare the similarity between ground truth and prediction [24]. The PA is called pixel accuracy, which represents the proportion of pixels with the correct prediction class out of the total number of pixels. Jaccard, also known as intersection over union, is the primary measure of overlap between ground truth and segmentation results [25]. Furthermore, since our designed thyroid nodule segmentation network needs to assist radiologists in clinically diagnosing the disease, recall and precision are also included in our comprehensive experiments. Moreover, false positive rates (FPR) and false negative rates (FNR) are also the most commonly used evaluation metrics of thyroid nodule segmentation model, which are of great significance for practical clinical practice.

3.3. Implementation Details

The hyperparameters of each experiment in this work are generally consistent. During training, these experiments appropriately use data enhancement and apply data enhancement with a probability of 10/11. The used transformations are in rotation with an angle in degrees \( \theta \sim U(-10, 10) \), horizontal flip with a possibility of 0.5. For our self-attention mechanism, we set the values of s to 1. The AdaBound [26] optimizer is employed, along with the Cosine annealing learning rate scheduler with a batch size of 16 and an initial learning rate of \( 1 \times 10^{-3} \) until convergence. We train all models from scratch for 200 epochs on four GeForce RTX 2080Ti.

3.4. Comparison to State-of-the-Art Models

In this section, we compare our proposed LCA-Net with other state-of-the-art thyroid nodule segmentation networks, such as SGU-Net [4], TRFE-Net [6], RUL-Net [5] and WU-Net [27]. The SGU-Net based on UNet extracts single-channel pixel semantic maps from the high-dimensional features of each decoding step, which serves as a high-level semantic guidance to low-level features for capturing more details. The TRFE-Net simultaneously predicts the segmentation of thyroid regions and nodules, forcing the same backbone network to infer the localization of nodules. The RUL-Net integrates a modified UNet with a level set evolution as the segmentation network. The WU-Net uses the DenseNet-121 network as the backbone network and integrates the Atrous spatial pyramid pooling (ASPP) module for thyroid nodule segmentation in ultrasound images. Experiments are conducted under the same loss function, optimizer and GPU conditions to ensure fairness.

Firstly, we calculate the nodule segmentation evaluation metrics defined above based on the TN-SCUI2020 dataset, and the comparison results for all methods are listed in Table 1. It can be seen from Table 1 that the proposed LCA-Net outperforms almost all state-of-the-art methods. Compared with several other models, the WU-Net has dramatically achieved the segmentation effect of thyroid nodules in ultrasound images (89.07% for dice, 98.24% for PA, 81.13% for Jaccard, 0.8957 for precision, 0.9082 for recall, 0.0076 for FPR and 0.1207 for FNR). However, thanks to the performance of LCA-Net for extracting both local details and global context information, the proposed LCA-Net achieves a certain improvement...
and reaches 2.48% for dice, 0.27% for PA, 3.06% for Jaccard, 0.0197 for precision, 0.027 for recall, 0.0013 for FPR and 0.0252 for FNR, respectively. The SGU-Net achieves poor results with 84.21%, 97.02%, 74.47%, 0.8346 and 0.8568 for dice, PA, Jaccard, precision and recall, respectively. The metrics of our method achieve a remarkable improvement and reach 90.26%, 98.87%, 82.65%, 0.9068 and 0.9184, respectively. Several examples of segmentation results with different methods are shown in Figure 7, from which it can be seen that our proposed method achieves the best thyroid nodule segmentation results. We select four different cases of thyroid nodules: a small nodule at the edge of the image, a large nodule at the edge of the image, a large nodule at the center of the image and a medium nodule at the center of the image. Other segmentation models have better results in the first row of the small nodule, except SGU-Net. For the nodule in the second row, the segmentation results of other models are far from the ground truth, while LCA-Net can still accurately segment the contour at the edge of the image and maintain good regional continuity owing to its ability to capture global associative information. Additionally, LCA-Net can still achieve better segmentation results than other models for the large nodule in the middle of the image. This is mainly due to the expansion of the receptive field of the backbone module for extracting detailed features and the flexible adaptive convolutions of the NAC module. For the medium-sized nodules in the last row, other models mistakenly segment two small nodules, which can easily cause major medical accidents in clinics. Fortunately, our model segments the more accurate nodule contour without misdiagnosis. In summary, the proposed LCA-Net has high segmentation accuracy and strong robustness to various nodules. Moreover, a boxplot diagram of the dice scores based on 5-fold cross verification has been presented in Figure 8a. Our model has the superiorities of median, interquartile and maximum dice values and has the least abnormal values, proving strong robustness for the segmentation of various nodules in the test set.

Table 1. Comparison of the results of state-of-the-art thyroid nodule segmentation models on the TN-SCU12020 dataset and TN3k dataset.

| Dataset  | Model     | Dice     | PA       | Jaccard | Precision | Recall    | FPR      | FNR     |
|----------|-----------|----------|----------|---------|-----------|-----------|----------|---------|
| TN-SCU12020 |          |          |          |         |           |           |          |         |
|          | SGUNet [3]| 84.21%   | 97.02%   | 74.47%  | 0.8346    | 0.8568    | 0.0111   | 0.1328  |
|          | TRFE-Net [5]| 85.03%  | 98.04%   | 75.48%  | 0.8639    | 0.8652    | 0.0093   | 0.2252  |
|          | RUL-Net [4]| 86.59%   | 97.97%   | 78.07%  | 0.8760    | 0.8812    | 0.0082   | 0.1543  |
|          | WU-Net [27]| 89.07%   | 98.24%   | 81.13%  | 0.8957    | 0.9082    | 0.0076   | 0.1207  |
|          | Proposed Method | 90.26% | 98.87%   | 82.65%  | 0.9068    | 0.9184    | 0.0063   | 0.0955  |
| TN3k     |          |          |          |         |           |           |          |         |
|          | SGUNet [3]| 75.03%   | 95.35%   | 64.25%  | 0.7719    | 0.7936    | 0.0158   | 0.2149  |
|          | TRFE-Net [5]| 77.56% | 96.23%   | 67.68%  | 0.7838    | 0.8316    | 0.0143   | 0.1915  |
|          | RUL-Net [4]| 78.08%   | 96.72%   | 68.05%  | 0.8024    | 0.8113    | 0.0175   | 0.1606  |
|          | WU-Net [27]| 81.27%   | 96.82%   | 70.54%  | 0.7822    | 0.8395    | 0.0195   | 0.1843  |
|          | Proposed Method | 82.08% | 96.97%   | 71.18%  | 0.8055    | 0.8534    | 0.0138   | 0.1473  |

In addition, to further prove the generalization performance and effectiveness of the models, we conducted a new set of segmentation experiments on the TN3k dataset. The calculated evaluation metrics and experimental conditions are precisely the same as above. The specific experiment results are illustrated in Table 1. Although the TN3k dataset has a small amount of data and low image quality, the proposed LCA-Net achieves the best nodule segmentation results, including 82.08% for dice, 96.97% for PA, 71.18% for Jaccard, 0.8055 for precision, 0.8534 for recall, 0.0138 for FPR and 0.1473 for FNR. Other state-of-the-art models cannot perform very stably on this dataset compared with our outstanding results, proving that LCA-Net has good generalization performance for different datasets. Figure 9 also reflects the visualization results of each model nodule segmentation. We
selected three examples: a medium-sized nodule in the middle of the image, a large nodule whose outline is all at the edge of the image, and a small nodule at the edge. It can be seen from the figure that our model can still more finely segment the thyroid contour. This also attributes the success to the capability of the context-attention modules to extract global features and the NAC module to adapt to nodules of various sizes and locations. Similarly, the boxplot diagram of the dice scores based on five-fold cross verification in Figure 8b also reflects the segmentation accuracy and generalization performance of our model. It can be seen from the figure that LCA-Net has minor abnormal points in the test results. Hence, it has good robustness for various types of nodule images in the test set. Simultaneously, these experiments also further demonstrate that our model has great application value in practical clinical thyroid segmentation.

![Figure 7](image1.jpg)

**Figure 7.** Qualitative comparison of different models on the TN-SCUI2020 dataset. Our predictions are closest to the ground truth and keep finer information.

![Figure 8](image2.jpg)

**Figure 8.** Boxplot of dice scores for each segmentation model. (a) Boxplot of dice scores on the TN-SCUI2020 dataset; and (b) boxplot of dice scores on the TN3k dataset. Median, interquartiles, minimum and maximum dice are provided.

Finally, we mixed the two datasets to further test the generalization ability and clinical application value of the model. We utilized the TN-SCUI2020 dataset as the training set and the TN3k dataset as the test set for experiments and vice versa. The specific experimental results are shown in Table 2. We can find that the proposed LCA-Net still achieves the highest test segmentation accuracy when the training set is completely different from the test set. This also proves that the LCA-Net has a strong application value and development prospects in clinical medicine.
3.5. Ablation Studies

In this section, in order to thoroughly evaluate the proposed LCA-Net framework and verify the performance of the selected loss function and optimization method, a variety of ablation studies were carried out on the TN-SCUI2020 dataset. Considering the fact that the TN-SCUI2020 dataset has wider data sources and more clinical representation than the TN3k dataset, we only use TN-SCUI2020 as the dataset for ablation experiments.

3.5.1. Effect of Modules Selection

Since our proposed LCA-Net integrates the merits of context-attention modules, backbone modules and a nodule adaptive convolutions module concurrently to design the model, it is essential to investigate the impact of each kind of module on the performance of LCA-Net. We conduct this ablation experiment and calculate the metrics with different modules selections. The results are summarized in Table 3. Compared to just using the backbone modules, combining backbone and context-attention modules achieves an improvement in the main evaluation metrics (2.83% for dice, 0.39% for PA, 3.06% for Jaccard, 0.0154 for precision, 0.0229 for recall, 0.0057 for FPR and 0.0219 for FNR), which benefit from the fact that the context-attention modules can capture more global context information to improve the segmentation performance for nodule contours at image boundaries. In addition, the adoption of the nodule adaptive convolutions module is the best means of enhancing the generalization of the model, which further gains breakthroughs in the evaluation metrics (1.04% for dice, 0.23% for PA, 1.76% for Jaccard, 0.0164 for precision, 0.0146 for recall, 0.0011 for FPR and 0.0173 for FNR). However, due to the characteristics of ultrasound images and the complexity of thyroid nodule structure, the lack of backbone modules in the network to extract local details easily leads to a declining segmentation...
performance. Moreover, it can also be seen from the above that these experiments indicate the effectiveness of the module’s design and the combination of the models.

Table 3. Comparison of thyroid nodule segmentation performance using various modules combinations.

| Dataset          | Module Construction | Results                      |
|------------------|---------------------|------------------------------|
| TN-SCU2020       | Backbone            | 86.39% 98.25%                |
|                  | Context-Attention    | 84.53% 97.65%                |
|                  | Nodule Adaptive      | 85.81% 97.94%                |
|                  | Dice                | 77.83% 75.51%                |
|                  | PA                  | 0.8750 0.8542                |
|                  | Jaccard             | 0.8809 0.8320                |
|                  | Precision           | 0.8897 0.8629                |
|                  | Recall              | 0.0138 0.0138                |
|                  | FPR                 | 0.1347 0.1435                |
|                  | FNR                 | 0.0109 0.1253                |

Several examples of segmentation results with modules variations are shown in Figure 10. Two different cases of thyroid nodules are selected: a large nodule at the center of the image and a small nodule at the edge of the image. If context-attention modules or context-attention modules and NAC module are selected, it is easy to cause the segmentation result of nodules imprecisely. Simultaneously, only employing the backbone modules or backbone modules and the NAC module will make the contour segmentation at the edge of the image less effective because the network cannot capture global context features. However, combining the context-attention modules and the backbone modules makes it possible to segment both edge contours and local details better. Furthermore, incorporating the NAC module further enhances the robustness of the network for the various nodules. It can be seen from these images that the proposed complete network structure can segment the nodule contour more finely and present more details, which effectively improves the accuracy of thyroid nodule segmentation and reduces false positives. In conclusion, our network structure is highly reasonable and has solid clinical value.

Figure 10. Qualitative comparison of various modules selections by visualization. A large nodule at the center of the image and a small nodule at the edge of the image are selected. The predictions of the complete network structure of LCA-Net are closest to the ground truth and keep more delicate information.
3.5.2. Effect of Loss Function

Furthermore, we demonstrate the effect of different loss functions on the segmentation performance of the model. In most ultrasound images, the background size is more significant than that of the nodule, which can easily lead to the problem of class imbalance during segmentation and make it difficult to continue training. The classical dice loss [28] is often employed for medical image segmentation, but its performance in small target segmentation is not ideal. Once the small target has some pixel prediction errors, it will significantly change the loss function, resulting in a drastic change in the gradient and the instability of training. Meanwhile, due to the same weight assigned by the class, the binary cross-entropy (BCE) loss function cannot adequately deal with the imbalance of pixel class.

This paper compares the segmentation results of binary cross-entropy loss, dice loss and our proposed loss function under the same network structure and AdaBound optimizer. It can be seen from Table 4 that the evaluation metrics (such as dice, PA and Jaccard) of our proposed loss function are higher than the segmentation results of the dice loss function and binary cross-entropy loss function. It is worth noting that there is a significant gap in the precision and recall of the test results of the models obtained by the dice and BCE loss function. The main reason is that neither solves the problem of pixel class imbalance. Moreover, the segmentation of thyroid nodules in clinics has high requirements for both precision and recall. The variation trend of each loss function is shown in Figure 11a,b. By comparing the two figures, we can infer that the BCE loss and dice loss functions make the model over-fitting in the training stage. Although the proposed loss function curve has a few fluctuations in the middle process, it is generally stable. Meanwhile, the loss function curve converges faster, and the final function value remains in a reasonable range. In summary, the overall performance of our proposed loss function is the best. Therefore, our novel loss function is suitable for optimizing the thyroid nodule segmentation model.

Table 4. Comparison of thyroid nodule segmentation performance using various loss functions.

| Dataset      | Loss Function | Dice   | PA     | Jaccard | Precision | Recall  | FPR    | FNR    |
|--------------|---------------|--------|--------|---------|-----------|---------|--------|--------|
| TN-SCUI2020  | BCE Loss      | 83.55% | 98.31% | 74.20%  | 0.6035    | 0.9672  | 0.0083 | 0.1496 |
|              | Dice Loss [28]| 86.17% | 98.49% | 77.41%  | 0.7419    | 0.9539  | 0.0065 | 0.1321 |
|              | Proposed Loss | 90.26% | 98.87% | 82.65%  | 0.9068    | 0.9184  | 0.0063 | 0.0955 |

Figure 11. The various loss function curves with the number of iterations. (a) Loss function in training stage; and (b) Loss function in validation stage.

3.5.3. Effect of Optimizer Selection

Finally, we also conducted a set of ablation experiments for optimizer selection. The optimizer method is employed to assist in training to obtain the best network parameters
and find the minimum loss of the loss function. Extensive experiments have shown that the SGD [29] optimizer has a slow convergence speed during training, which causes the result to easily fall into a local minimum. Simultaneously, the generalization of the model trained by the Adam [30] optimizer is not ideal. The AdaBound optimizer combines the merits of the SGD optimizer and Adam optimizer. At the beginning of training, it can be as fast as Adam and has as good a convergence as SGD in the later stage. This paper uses the SGD optimizer, Adam optimizer and AdaBound optimizer to train the LCA-Net under our proposed loss function. We select the appropriate optimizer by comparing different test results. It can be seen from Table 5 that the AdaBound optimizer achieved the best results in the evaluation indexes, which is higher than the commonly used SGD optimizer (3.49% for dice, 0.61% for PA, 3.6% for Jaccard, 0.0312 for precision and 0.0362 for recall). In addition, the test results using the Adam optimizer are not ideal, which proves that Adam is not suitable as the optimizer of the proposed model.

### Table 5. Comparison of thyroid nodule segmentation performance using various optimizers.

| Dataset      | Optimizer     | Dice   | PA     | Jaccard | Precision | Recall | FPR    | FNR    |
|--------------|---------------|--------|--------|---------|-----------|--------|--------|--------|
| TN-SCUI2020  | Adam [30]     | 78.93% | 96.93% | 68.41%  | 0.7901    | 0.8594 | 0.0213 | 0.1434 |
|              | SGD [29]      | 86.77% | 98.26% | 79.05%  | 0.8756    | 0.8822 | 0.0075 | 0.1215 |
|              | AdaBound [26] | 90.26% | 98.87% | 82.65%  | 0.9068    | 0.9184 | 0.0063 | 0.0955 |

The optimization of our proposed loss function by the three optimizers is shown in Figure 12. Compared with the SGD optimizer, the loss function curve corresponding to the AdaBound optimizer drops rapidly in the initial stage. As the number of iterations increases, the loss function tends to converge smoothly in the later stage. Meanwhile, the fluctuation of the loss function corresponding to the Adam optimizer is considerable, proving that the Adam optimizer is not suitable for the optimization of our model. In conclusion, our choice of the AdaBound optimizer for training the model is sensible and superior.

![Figure 12. The novel loss function curve under different optimizers with the number of iterations.](image)

## 4. Conclusions

In this paper, a local and context-attention adaptive network (LCA-Net) for thyroid nodule segmentation in ultrasound images is proposed. The transformers-based context-attention module is proposed to capture more global associative information and perceive edge information of nodule contours to make up for the deficiency of convolutional neural...
networks in capturing context information. Additionally, the backbone module enables the network to extract more local features. Furthermore, a nodule adaptive convolutions (NAC) module is designed to deal with thyroid nodules of different sizes and positions adaptively. For the training of the model, an optimized loss function for pixel class imbalance is introduced to further enhance the segmentation performance. Compared with the latest algorithms, our method achieves promising segmentation accuracy and strong generalization performance. We believe that the proposed model could also serve other medical image segmentation tasks and has broad application prospects for diagnosing thyroid nodules.

Author Contributions: Conceptualization, Z.T. and Y.S.; methodology, Z.T.; investigation and validation, Z.T., Y.S. and S.R.; data curation and formal analysis, Z.T., H.D., X.W. and W.W.; writing—original draft preparation, Z.T., Y.S. and S.R.; writing—review and editing, H.D., S.R., X.W. and W.W.; funding acquisition, H.D., S.R. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Chongqing Natural Science Foundation under Grant No. cstc2021jcyj-msxmX1096.

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

References

1. Song, W.; Li, S.; Liu, J.; Qin, H.; Zhang, B.; Zhang, S.; Hao, A. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE J. Biomed. Health Inform. 2018, 23, 1215–1224. [CrossRef]

2. Chen, J.; You, H.; Li, K. A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Comput. Methods Programs Biomed. 2020, 185, 105329. [CrossRef] [PubMed]

3. Zhang, Y.; Lai, H.; Yang, W. Cascade UNet and CH-UNet for thyroid nodule segmentation and benign and malignant classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 129–134.

4. Pan, H.; Zhou, Q.; Latecki, L.J. SGU.NET: Semantic Guided UNET For Thyroid Nodule Segmentation. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 630–634.

5. Yang, B.; Yan, M.; Yan, Z.; Zhu, C.; Xu, D.; Dong, F. Segmentation and classification of thyroid follicular neoplasm using cascaded convolutional neural network. Phys. Med. Biol. 2020, 65, 245040. [CrossRef] [PubMed]

6. Gong, H.; Chen, G.; Wang, R.; Xie, X.; Mao, M.; Yu, Y.; Chen, F.; Li, G. Multi-Task Learning For Thyroid Nodule Segmentation With Thyroid Region Prior. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 257–261.

7. Du, W.; Sang, N. An effective method for ultrasound thyroid nodules segmentation. In Proceedings of the 2015 International Symposium on Bioelectronics and Bioinformatics (ISBB), Beijing, China, 14–17 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 207–210.

8. Zhao, J.; Zheng, W.; Zhang, L.; Tian, H. Segmentation of ultrasound images of thyroid nodule for assisting fine needle aspiration cytology. Health Inf. Sci. Syst. 2013, 1, 5. [CrossRef] [PubMed]

9. Keramidas, E.G.; Maroulis, D.; Iakovidis, D.K. TND: A thyroid nodule detection system for analysis of ultrasound images and videos. J. Med. Syst. 2012, 36, 1271–1281. [CrossRef] [PubMed]

10. Chen, Y.; Li, D.; Zhang, X.; Jin, J.; Shen, Y. Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning. Med. Image Anal. 2021, 67, 101819. [CrossRef] [PubMed]

11. Ma, J.; Wu, F.; Jiang, T.; Zhao, Q.; Kong, D. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 2017, 12, 1895–1910. [CrossRef] [PubMed]

12. 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.

13. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. arXiv 2015, arXiv:1511.07122.

14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.

15. Gao, Y.; Zhou, M.; Metaxas, D.N. UBTN: A hybrid transformer architecture 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. 61–71.

16. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. CE-Net: Context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging 2019, 38, 2281–2292. [CrossRef]
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 2015, 37, 1904–1916. [CrossRef] [PubMed]

18. 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.

19. Ouahabi, A.; Taleb-Ahmed, A. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. *Pattern Recognit. Lett.* 2021, 144, 27–34. [CrossRef]

20. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

21. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* 2019, arXiv:1908.08681.

22. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

23. Zhao, Z.; Yang, G. Unsupervised Contrastive Learning of Radiomics and Deep Features for Label-Efficient Tumor Classification. 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. 252–261.

24. Li, L.; Ma, H. RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation. *Sensors* 2022, 22, 2452. [CrossRef] [PubMed]

25. Yao, C.; Wang, M.; Zhu, W.; Huang, H.; Shi, F.; Chen, Z.; Wang, L.; Wang, T.; Zhou, Y.; Peng, Y.; et al. Joint segmentation of multi-class hyper-reflective foci in retinal optical coherence tomography images. *IEEE Trans. Biomed. Eng.* 2021, 69, 1349–1358. [CrossRef] [PubMed]

26. Wu, D.; Yuan, Y.; Huang, J.; Tan, Y. Optimize TSK fuzzy systems for regression problems: Minibatch gradient descent with regularization, DropRule, and AdaBound (MBGD-RDA). *IEEE Trans. Fuzzy Syst.* 2019, 28, 1003–1015. [CrossRef]

27. Wu, Y.; Shen, X.; Bu, F.; Tian, J. Ultrasound image segmentation method for thyroid nodules using ASPP fusion features. *IEEE Access* 2020, 8, 172457–172466. [CrossRef]

28. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 565–571.

29. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* 2016, arXiv:1609.04747.

30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* 2014, arXiv:1412.6980.