applied sciences

MDPI

Article

Residual-Attention UNet++: A Nested Residual-Attention U-Net for Medical Image Segmentation

Zan Li, Hong Zhang *, Zhengzhen Li and Zuyue Ren

School of Computer Science and Technology, Minzu University of China, Beijing 100081, China; 20301827@muc.edu.cn (Z.L.); 19301533@muc.edu.cn (Z.L.); 21302022@muc.edu.cn (Z.R.)
* Correspondence: zhanghong751103@muc.edu.cn

Abstract: Image segmentation is a basic technology in the field of image processing and computer vision. Medical image segmentation is an important application field of image segmentation and plays an increasingly important role in clinical diagnosis and treatment. Deep learning has made great progress in medical image segmentation. In this paper, we proposed Residual-Attention UNet++, which is an extension of the UNet++ model with a residual unit and attention mechanism. Firstly, the residual unit improves the degradation problem. Secondly, the attention mechanism can increase the weight of the target area and suppress the background area irrelevant to the segmentation task. Three medical image datasets such as skin cancer, cell nuclei, and coronary artery in angiography were used to validate the proposed model. The results showed that the Residual-Attention UNet++ achieved superior evaluation scores with an Intersection over Union (IoU) of 82.32%, and a dice coefficient of 88.59% with the skin cancer dataset, a dice coefficient of 85.91%, and an IoU of 87.74% with the cell nuclei dataset and a dice coefficient of 72.48%, and an IoU of 66.57% with the angiography dataset.

Keywords: residual unit; attention mechanism; UNet++; medical image segmentation

Citation: Li, Z.; Zhang, H.; Li, Z.; Ren, Z. Residual-Attention UNet++: A Nested Residual-Attention U-Net for Medical Image Segmentation. *Appl. Sci.* **2022**, *12*, 7149. https://doi.org/10.3390/app12147149

Academic Editor: Nikolaos Dikaios

Received: 6 June 2022
Accepted: 14 July 2022
Published: 15 July 2022

1. Introduction

Image semantic segmentation is one of the core tasks in the field of computer vision. From the microscopic point of view, semantic segmentation can be understood as classifying each pixel in the image. It is a pixel-level space-intensive prediction task. In other words, semantic segmentation attempts to understand the meaning of each pixel in the image semantically, such as identifying whether it is a car, a building, or a pedestrian. From a macro point of view, semantic segmentation can be seen as assigning consistent semantic labels to a class of things instead of each pixel. It has important application value in video surveillance, automatic driving, scene understanding, and 3D scene modeling.

Interactive image segmentation methods include Graphcut [1] and Grabcut [2]. Graphcut uses the minimum-cut maximum-flow algorithm for image segmentation, which can segment the image into the foreground and background. When using this algorithm, it is necessary to draw a few strokes at the foreground and background as input. The algorithm will establish a weighted map of the similarity between each pixel and the foreground and background, and distinguish the foreground and background by solving the minimum cut. Grabcut algorithm is an iterative algorithm of the Graphcut method, which utilizes the color information and boundary information in the RGB image, and manually selects the area of interest. The inside of the frame is regarded as the unknown area, and the outside of the frame is regarded as the background area. A GMM (Gaussian Mixture Model) is established between the unknown area and the calibrated background area, the foreground GMM and the background GMM are initialized according to the result of the manual selection, and then the initialized area (unknown area) is divided into the target and the background area.

Early research on fully automatic semantic segmentation were dominated by traditional methods such as the threshold method [3], the edge detection-based method [4],

the conditional random field (CRF)-based method [5], and the cluster-based method [6]. Among them, the threshold method uses a single threshold for direct classification, which is mainly applied to the segmentation of grayscale images, such as Otsu [7], which belongs to an adaptive threshold determination method. The image fixed threshold binarization with this threshold has the largest inter-class variance. It divides the image into background and foreground according to the gray characteristics of the image so that the segmentation with the largest inter-class variance means the minimum misclassification probability. The core idea is to find a threshold, T, which divides all pixels of the image into two categories. The pixel values of one category are less than or equal to T (background area), and the pixel values of the other category are greater than T (foreground area). When the variance between these two categories reaches the maximum, the T value is considered to be the most appropriate threshold. The edge detection-based method classifies objects by boundary information. The CRF method is a probabilistic framework for labeling and segmentation; the interaction process between pixels can be modeled, and the efficiency is usually higher than the threshold method and edge detection method. The clustering-based method divides the image into K groups by clustering pixels or regions with similar characteristics; K-means clustering is the mainstream method.

In the past few years, medical image segmentation has attracted increasing attention, the purpose of which is to separate the parts with special meanings from medical images. In the medical field, doctors mainly rely on medical film images and subjective judgments to diagnose diseases. Medical image segmentation is a key step in medical 3D reconstruction and quantitative analysis. It is also an important technical premise for quantification of lesion areas, selection of treatment methods, and radiotherapy. The accuracy of segmentation results directly affects the follow-up treatment effect. Medical images mainly come from different imaging technologies, including ultrasound, X-ray, magnetic resonance imaging (MRI), computer tomography (CT), and positron-emission tomography (PET). On the one hand, different segmentation methods should be selected according to different imaging technologies and imaging parts. On the other hand, medical images are different from natural images and have high complexity. Due to the single pixel of the image, the boundary, shape and other information of the lesion are also fuzzy. In addition, automatic preprocessing without human involvement also reduces human errors and overall time and cost [8]. Considering the slow process and the complexity of manual segmentation methods, there is a great demand for a fast and accurate computer method for fully automatic segmentation [9].

To meet the need for more accurate medical image segmentation, inspired by the attention mechanism and residual unit, we propose Residual-Attention UNet++, an extension of the UNet++ architecture using a residual unit and attention mechanism.

The contributions of this work can be summarized as follows:

- The residual unit and attention mechanism were introduced to UNet++ to increase the weight of target areas and to solve the degradation problem.
- The proposed model Residual-Attention UNet++ was introduced for medical image segmentation.
- The experiments conducted on three medical imaging datasets demonstrated better performance in segmentation tasks compared with existing methods.
- Residual-Attention UNet++ could increase the weight of the target area and suppress the background area irrelevant to the segmentation task.
- Comparison against some UNet-based methods showed superior performance.
- The pruned Residual-Attention UNet++ enabled faster inference at the cost of minimal performance degradation.

The remaining content is organized as follows. Section 2 discusses some related work, Section 3 presents the introduction of the proposed architectures, Section 4 describes the datasets, experimental details, and results, and finally, Section 5 concludes this paper.

2. Related Work

Considering the limitations of traditional segmentation methods and the need for high accuracy in medical image segmentation, rapidly developing deep learning techniques have been widely used in the field of medical images, mainly due their its advantages in segmentation speed and accuracy, which can significantly reduce the time and help doctors diagnose diseases more efficiently. Fully Convolutional Networks (FCN) proposed by Long et al. [10] have become the mainstream framework in the field of image semantic segmentation, which is the pioneering work of the most successful and advanced deep learning technology. For the first time, FCN realizes the pixel-level semantic segmentation task that can accept inputs of any image size. It replaces the fully connected layer in the CNN model with a fully convolutional layer to achieve pixel-level dense prediction, uses deconvolution to upsample the feature map, and proposes skip layer connections to fully integrate global semantic information and local location information to achieve precise segmentation. However, the disadvantages of FCN are also obvious. First, the upsampling process is rough, resulting in serious loss of semantic information of feature maps, which seriously affects the segmentation accuracy. Second, skip connections fail to make full use of the context information and spatial location information of the image, resulting in low utilization of global information and local information.

Deep convolutional neural network uses repeated convolution and pooling to increase the range of the receptive field and improve the ability of the feature to express the global information, but it will cause the problem of reduced resolution and loss of detailed information. Furthermore, the features acquired by convolution are spatially invariant, which hinders the segmentation task [11]. For the above problems, DeepLab v1, proposed by Chen et al. [12], reduces the loss of detailed information through atrous convolution while increasing the receptive field and uses CRF [13] to improve the ability to obtain boundary details. It is based on VGGNet as the backbone network to obtain image semantic information, upsample the features, and finally use the CRF to obtain a more accurate segmentation effect.

DeepLab v2 [14] is an improved method based on DeepLab v1. It uses ResNet-101 as the backbone network and add continuous atrous convolution to replace the original downsampling layer, which improves the resolution of the output layer while maintaining the same receptive field. Another contribution of DeepLab lies in employing the ASPP module to extract and fuse multi-scale information of feature layer information, effectively fuse local and global context information, and improve the segmentation accuracy of the model. In addition, pixel-level classification belongs to low-level semantic information, so it appears very vague in local details [15]. DeepLab v2 makes the position of the segmentation boundary more accurate with the help of a fully connected CRF.

DeepLab v3 [16] continues to uses ResNet-101 as the backbone network. For multi-scale target segmentation, the ASPP module has been redesigned, which is composed of atrous convolution and BN layers with different sampling rates, and the modules are arranged in a serial or parallel manner to obtain larger receptive fields and obtain multi-scale information. In addition, DeepLab v3 removes CRF and the experimental result showed that it is better than DeepLab v1 and v2. DeepLab v3+ [17], further expands DeepLab v3, and adds a simple and effective decoding module to improve the segmentation efficiency of the model, especially in the performance of target boundary segmentation, with a greater degree of optimization. Besides, through further study of the Xception model, the maximum pooling is replaced by deep separable convolution, which is used in ASPP and the decoder network structure to obtain a faster and more robust encoder–decoder network.

SegNet [18] is a symmetrical segmentation network of an encoder and decoder, which achieves end-to-end pixel-level image segmentation [15]. The encoder network is basically the same as VGG16 in structure, which is stacked by several convolution layers and pooling layers. SegNet discards the full connection layer to retain the high-resolution feature map at the encoder output position, and greatly reduces the number of parameters in the encoder. The decoder corresponds to the encoder. Its function is to restore the low-resolution coding

feature map to the complete input resolution, and the output will finally be passed to the multi category softmax classifier, which can be used in the pixel-level classification task to generate the independent probability of pixels in different categories.

Based on FCN, Ronneberger et al. [19] proposed a deep model named "U-Net" in 2015, which is dedicated to biomedical image segmentation. Convolutional encoding and decoding units are the two main parts of this structure, as shown in Figure 1. In these two parts of the network, the ReLU activation follows the basic convolution operation while 2×2 max pooling follows the activation operation in the encoding unit. On the other hand, the up-convolution operation is implemented in the decoding unit, which can up-sample the feature maps. Crop and copy are used to perform feature fusion between encoding unit and decoding unit, which are pivotal due to the loss of border pixels in every convolution. U-Net model ensures the accuracy of positioning and the acquisition of context information at the same time. In addition, this model can achieve higher segmentation results with less training samples, which is very important for the small number of medical image samples.

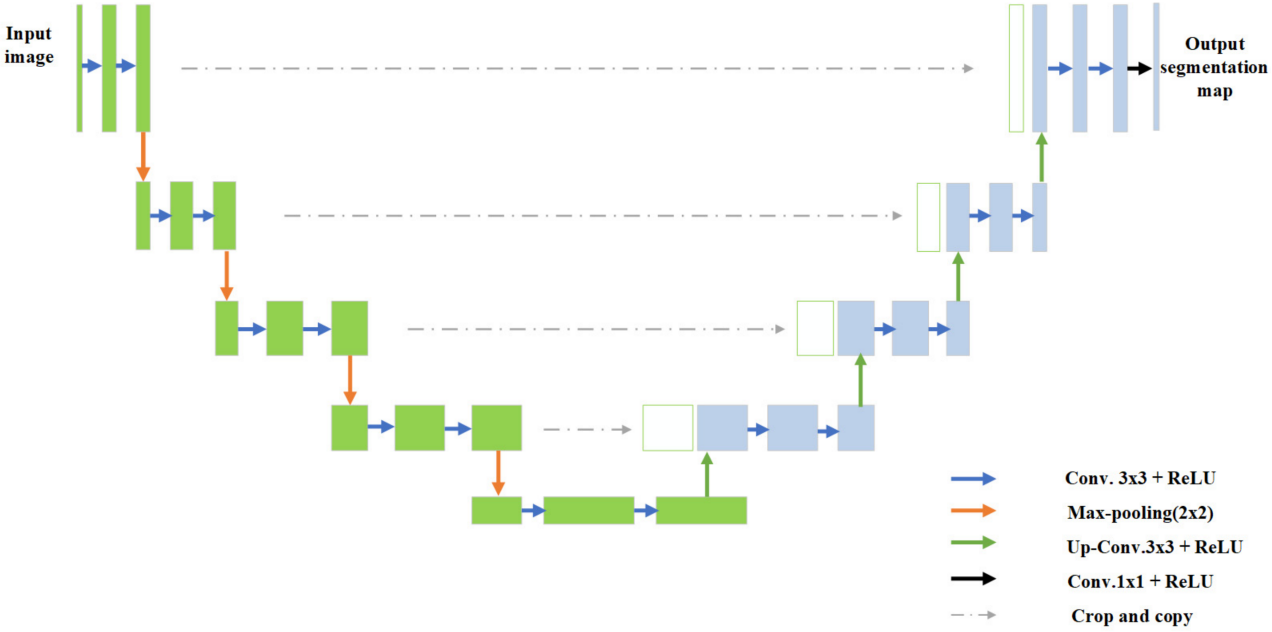


Figure 1. The structure of the U-Net.

Different variants of U-Net have been proposed in the past few years. For instance, Çiçek et al. [20] proposed a network called "3DU-Net" for volumetric segmentation that learns from sparsely annotated volumetric images, which extends the previous U-Net architecture by replacing all 2D operations with their 3D counterparts. Milletari et al. [21] proposed a volume-based, fully convolutional, end-to-end deep learning model for 3D medical images, which is called "V-Net". It consists of upsampling and downsampling. The main function of the former is to extract image features, including five convolution units, each of which is composed of convolution layer, pooling layer, and activation function. Convolution adopts $5 \times 5 \times 5$, with a filling size of 2 and a stride of 1. The pooling layer can reduce the parameters in the training process, reduce the memory occupation rate, and improve the training speed. The activation function is the PReLU [22] function. The residual learning structure is also added to the convolution block to effectively alleviate the problem of gradient disappearance. There is a cascade between downsampling and upsampling. The Cascade fuses the feature map in the downsampling process with the feature map after the deconvolution of the upsampling, and then the convolution operation is carried out, which can effectively prevent the loss of features in the downsampling process, and retain the target location information in the downsampling part and the edge features of the image. Alom et al. [8] proposed a network called "R2U-Net" for medical image

segmentation, which replaces the plain convolution block in the encoder–decoder network with the recurrent residual convolution units (RRCU). The advantage of R2U-Net model is that it performs element-by-element feature summation outside the U-Net model, and integrates high-dimensional abstract feature information and low-dimensional detailed feature information. Besides, the feature accumulation inside the RRCU structure can ensure a stronger feature representation and effectively increase the depth of the network structure. H-Dense UNet [23] was proposed for liver and tumor segmentation in 2017, which is a novel, hybrid, densely connected U-Net. H-Dense UNet transforms 3D volume blocks into 2D adjacent slices by using the transform processing function F; then these 2D slices are sent to 2D DenseUNet to extract on-chip features. The original input of 3D and the prediction result after 2D DenseUNet conversion are concatenated and sent to the 3D network to extract inter-slice features. After the fusion training, the results obtained from the training and the results obtained from the previous 2D training are added to increase the feature fusion. Finally, Finally, the final result is predicted by HFF. Attention U-Net [24] introduces the attention mechanism into U-Net, intensively extract significant features useful for specific tasks (such as relevant tissues or organs) from medical images, and suppresses the irrelevant inputs [25]. Before splicing the features on each resolution of the encoder with the corresponding features in the decoder, an attention module is used to readjust the output characteristics of the encoder. The module generates a gating signal to control the importance of features at different spatial locations. Zhou et al. [26] proposed a new medical image segmentation model called "UNet++", in which the encoder and decoder sub-networks were connected through a series of nested, dense skip pathways. The improvement of UNet++ from U-Net is that UNet++ adopts redesigned skip pathways to connect the encoder network and the decoder network. In U-Net, the feature map in the encoder network is directly transmitted to the decoder network; however, through the redesigned skip pathways in UNet++, the feature map of the encoder network is mapped to the decoder network through dense convolution blocks. In fact, the feature graph semantic level in the encoder is close to the feature graph semantic level in the decoder through dense convolution blocks.

Relying on the redesigned skip pathways, UNet++ improves the semantic gap between the feature maps of encoder and decoder subnetworks; however, dense convolution blocks also affect the gradient propagation, which will make the training process more complicated. According to the residual neural network proposed by He et al. [27], it can facilitate training and address the degradation problem. The residual network consists of sequentially stacked residual units, which can be illustrated as the following form:

$$\begin{cases} y_l = x_l + \mathcal{F}(x_l, \mathcal{W}_l) \\ x_{l+1} = L(y_l) \end{cases} \tag{1}$$

where x_l and x_{l+1} are the input and output of the l^{th} residual unit respectively, $\mathcal{F}(\cdot)$ is the residual function, and $\mathcal{W}_l$ are parameters of the block. $L(y_l)$ is the activation function. Figure 2a presents the plain and residual unit. As shown in Figure 2b, the residual unit contains two combinations of batch normalization (BN), ReLU, and convolutional layer. Although the residual unit improves the gradient propagation, it also increases the network parameters, which can easily lead to overfitting. For this, we introduced another module called dropout [28] in the residual unit. Dropout is a regularization technique, which drops a unit (along with connections) at training phase with a certain probability p (a common value is $p = 0.5$). It was mostly applied on top layers that had a large number of parameters to prevent feature coadaptation and overfitting [29]. As shown in Figure 2c, the dropout layer was added behind the first convolution layer in the residual unit.

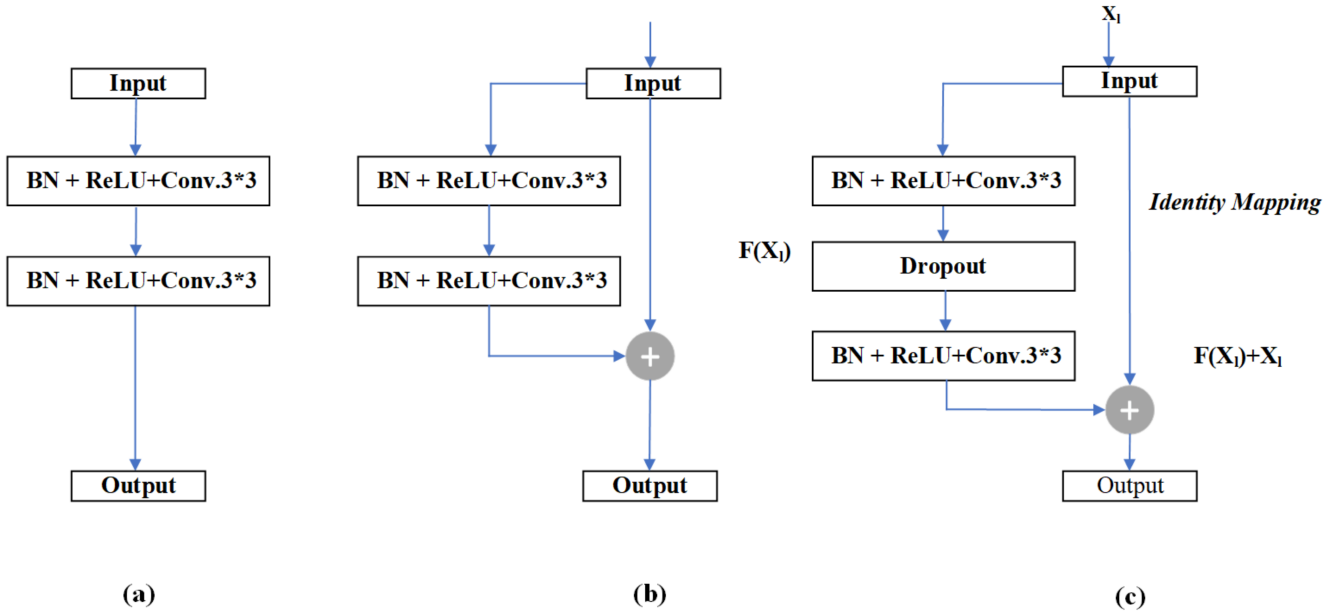


Figure 2. Different variants of convolutional and residual units. (**a**) Plain convolution units, (**b**) residual unit, (**c**) residual unit with dropout layer.

Attention mechanism is a core technology that has been widely used in natural language processing (NLP), statistical learning, image detection, speech recognition, and other fields. It was the proposal of Non-local [30] that applied the attention mechanism to computer vision for the first time. In order to focus on the target organs related to the segmentation task, we refer to the method proposed by Attention U-Net, which adds an Attention Gate to the network architecture. The structure of Attention Gate is shown in Figure 3, with the upsampling features from the expansion path and the corresponding features of the encoder as its inputs. The former is used as a gating signal to enhance the learning of target regions relevant to the segmentation task and to suppress task-irrelevant regions [31]. Firstly, both the inputs are passed through the operation of Convolution and BatchNorm and added to obtain A. A is obtained through the first activation function ReLU and Convolution and BatchNorm operation to obtain B. Next, B passes through the second activation function Sigmoid and Resample to obtain the attention coefficient α, and finally, the encoder feature is multiplied pixel-by-pixel by the coefficient α to obtain the output.

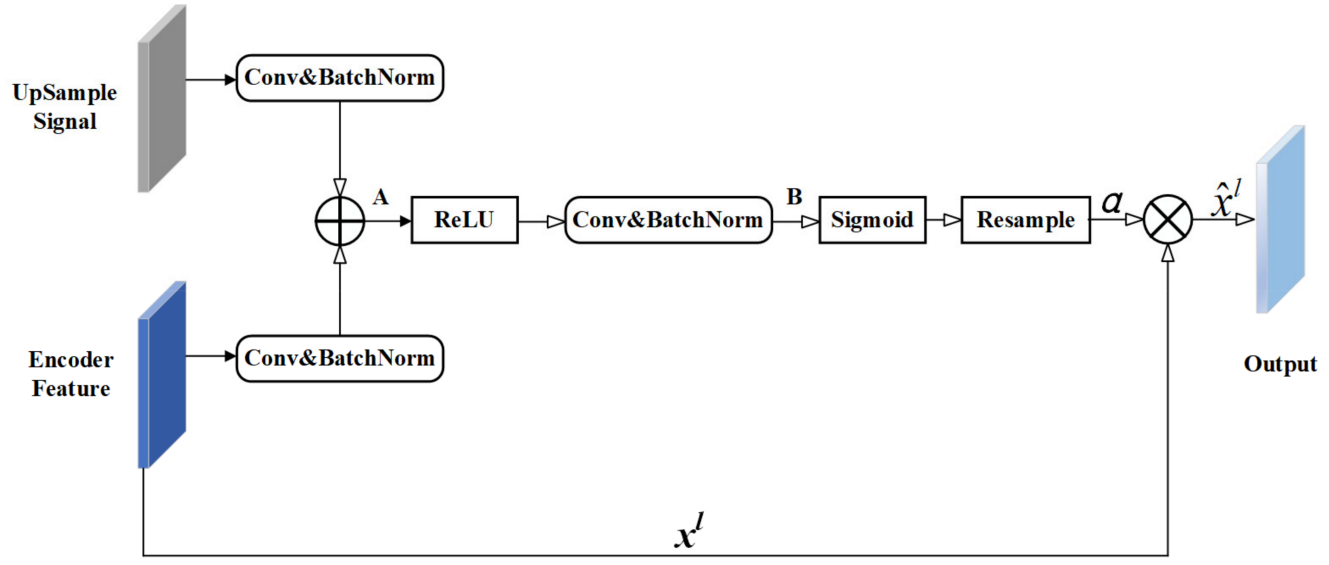


Figure 3. Architecture of Attention Gate.

3. Methodology

3.1. Residual-Attention UNet++

Here, we proposed the Residual-Attention UNet++, an integrated neural network that combines strengths of UNet++, residual unit, and attention mechanism for medical image segmentation, Figure 4 shows its overall structure. As we can see, the proposed model uses UNet++ as the basic network framework, which adopts redesigned skip pathways to connect the encoder network and the decoder network. The feature map of the encoder network was mapped to the decoder network through dense convolution blocks. In the above way, the feature graph semantic level in the encoder is close to the feature graph semantic level in the decoder.

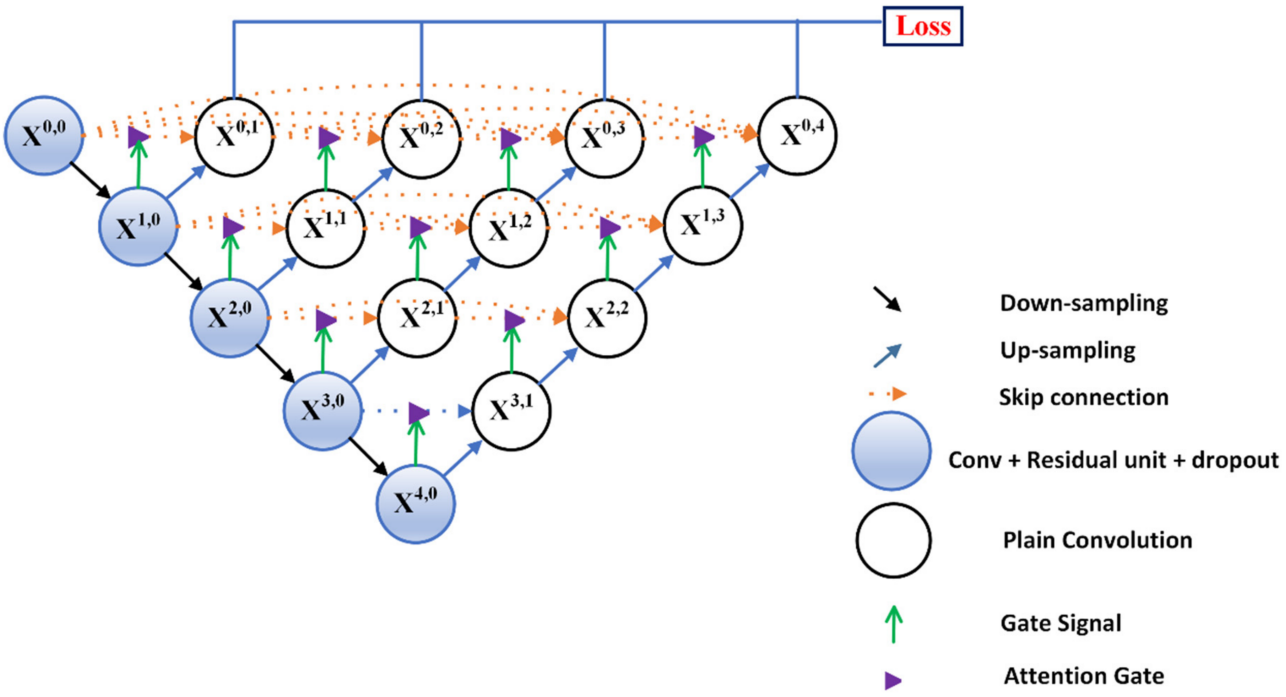


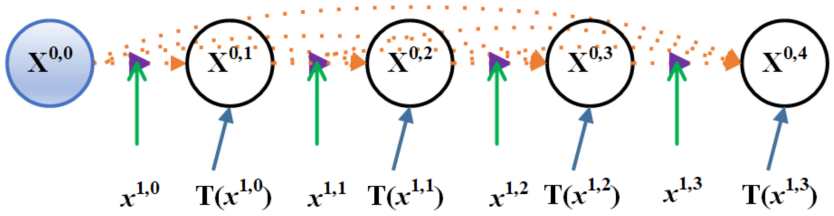
Figure 4. Architecture of Residual-Attention UNet++.

The skip pathway was formulated in the following form: $x^{i,j}$ represents the output of node $X^{i,j}$, i indexes the down-sampling layer according to the encoder sub-network, and j indexes the convolution layer of the dense block along the skip pathway. $x^{i,j}$ can be calculated by the following mathematical formula:

$$x^{i,j} = \begin{cases} \mathcal{H}\{x^{i-1,j}\}, & j = 0 \\ \mathcal{H}\left\{\left[\int_{k=0}^{j-1} AG\left(x^{i,k}\right), T\left(x^{i+1,j-1}\right)\right]\right\}, & j > 0 \end{cases} \quad (2)$$

where $\mathcal{H}\{\cdot\}$ represents a convolution operation followed by an ReLU activation, $AG(\cdot)$ and $T(\cdot)$ are used for attention gate and upsampling operations respectively, and [] denotes the concatenation layer. Figure 5 further explains the first skip pathway in Residual-Attention UNet++.

$$x^{0,1} = \mathrm{H}\{[\mathrm{AG}(x^{0,0}), \mathrm{T}(x^{1,0})]\} \qquad x^{0,3} = \mathrm{H}\{[\mathrm{AG}(x^{0,0}), \mathrm{AG}(x^{0,1}), \mathrm{AG}(x^{0,2}), \mathrm{T}(x^{1,2})]\}$$



$$x^{0,2} = \mathrm{H}\{[\mathrm{AG}(x^{0,0}), \mathrm{AG}(x^{0,1}), \mathrm{T}(x^{1,1})]\} \qquad x^{0,4} = \mathrm{H}\{[\mathrm{AG}(x^{0,0}), \mathrm{AG}(x^{0,1}), \mathrm{AG}(x^{0,2}), \mathrm{AG}(x^{0,3}), \mathrm{T}(x^{1,3})]\}$$

Figure 5. Detailed analysis of the first skip pathway of Residual-Attention UNet++.

This combination brings three benefits: (1) UNet++ improves the semantic gap between the feature maps of encoder and decoder subnetworks; (2) the residual unit eases training of the network and addresses the degradation problem; and (3) the attention mechanism can increase the weight of the target area and suppress the background area irrelevant to the segmentation task, so the accuracy of Residual-Attention UNet++ is improved.

3.2. Deep Supervision

In the proposed model, we also introduce deep supervision [32]. On the other hand, with the help of dense skip connections in nested blocks, Residual-Attention UNet++ gains full resolution feature maps at different semantic levels from $\{x^{0,j}, j \in \{1,2,3,4\}\}$, which are manageable to deep supervision. We used a combination of binary cross-entropy and dice coefficient as the loss function to add to each of the above four nodes, which is described as:

$$\mathcal{L}(h, \hbar) = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2}\cdot h_b \cdot log\hbar_b + \frac{2\cdot h_b \cdot \hbar_b}{h_b + \hbar_b}\right) \tag{3}$$

where h_b and $\hbar_b$ denote the flatten predicted probabilities and the flatten ground truths of the b^{th} image respectively, and N indicates the batch size.

3.3. Model Pruning

Figures 6–9 show how the choice of segmentation branch results in architectures of varying complexity [33]. We used Residual-Attention UNet++ L^i to denote Residual-Attention UNet++ pruned at level i. For example, Residual-Attention UNet++ L^1 means the prediction result is from node $X^{0,1}$, which is a maximally pruned model. Moreover, Residual-Attention UNet++ L^4, which is not pruned, indicates that the prediction result comes from node $X^{0,4}$.

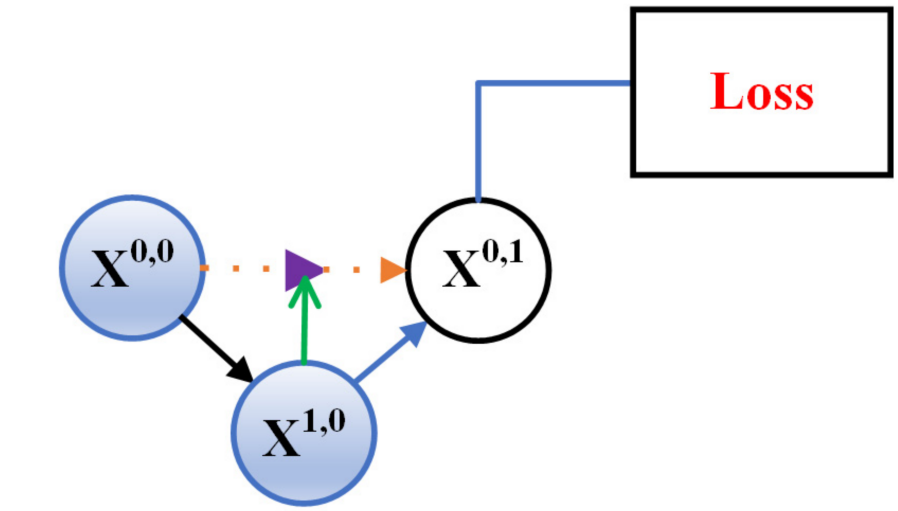


Figure 6. Residual-Attention UNet++ can be pruned to Residual-Attention UNet++ L^1 if trained with deep supervision.

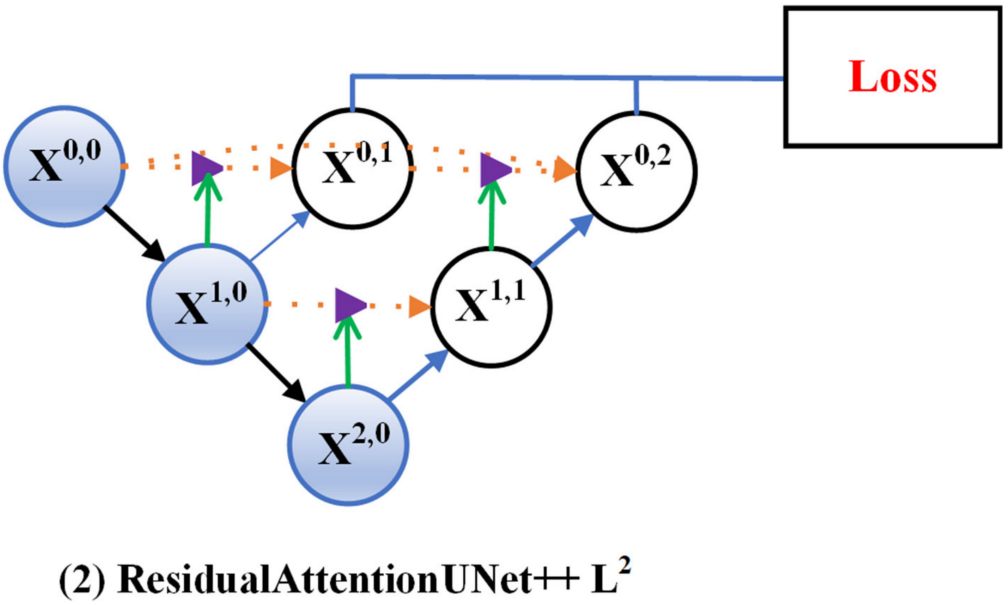


Figure 7. Residual-Attention UNet++ can be pruned to Residual-Attention UNet++ L^2 if trained with deep supervision.

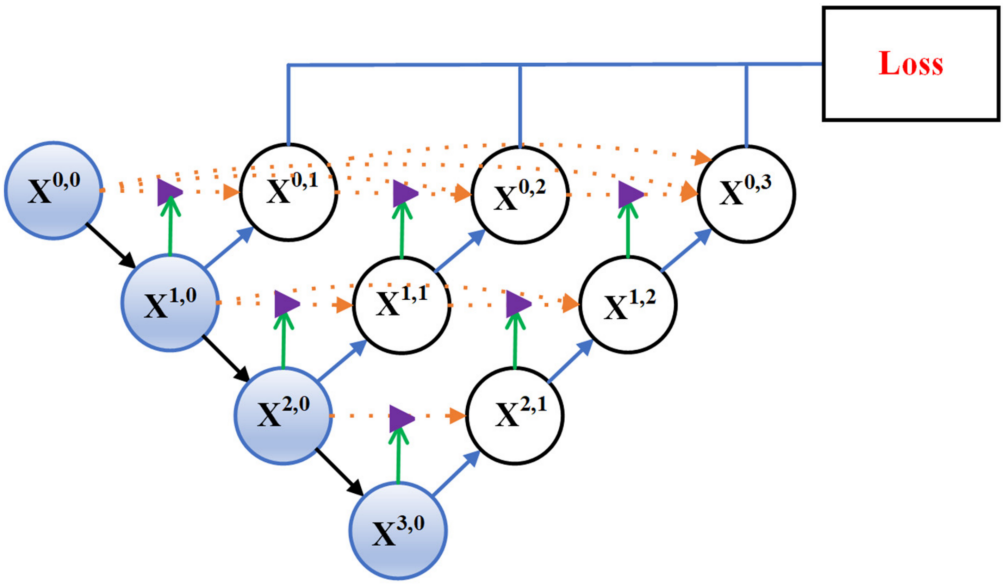


Figure 8. Residual-Attention UNet++ can be pruned to Residual-Attention UNet++ L^3 if trained with deep supervision.

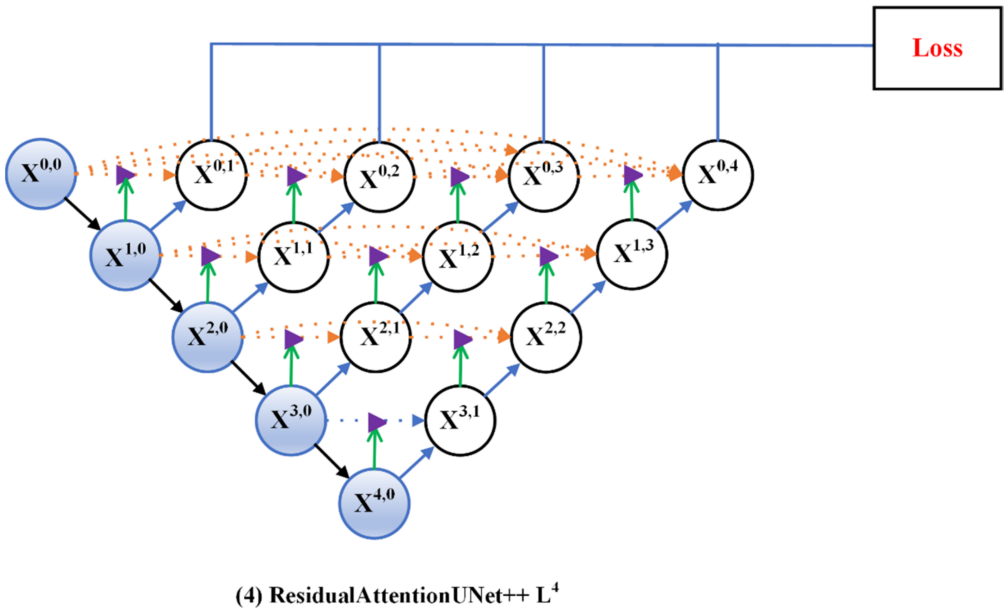


Figure 9. Residual-Attention UNet++ can be pruned to Residual-Attention UNet++ L^4 if trained with deep supervision.

4. Experiments and Results

The accurate definition of the skin canceration area, nucleus, and coronary artery boundary is very important for subsequent diagnosis and research. Therefore, to demonstrate the performance of the Residual-Attention UNet++ model, we tested it on two medical imaging datasets, which include skin cancer lesion segmentation and cell nuclei segmentation from 2D images as shown in Figure 10. The PyTorch frameworks were used on a single GPU machine with 16 GB of RAM and an NVIDIA RTX 3070 for this implementation.

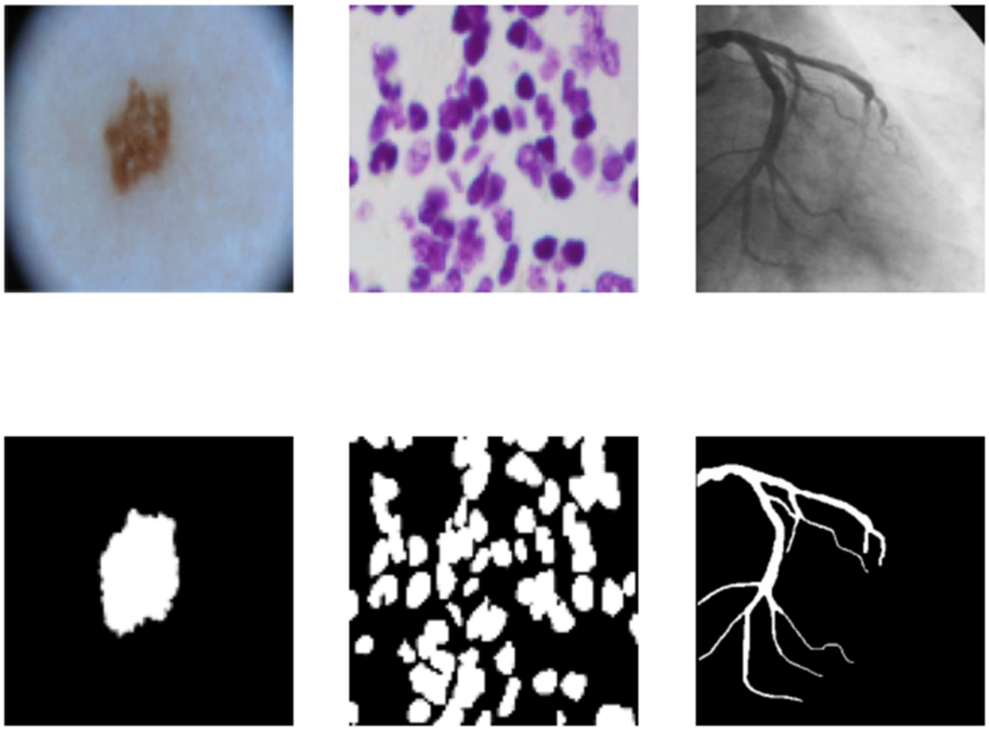


Figure 10. Medical image segmentation: skin cancer on the **left**, cell nuclei in the **middle**, and coronary artery in angiography on the **right**.

4.1. Dataset

4.1.1. Skin Cancer Segmentation

The dataset used in this work was from ISIC 2017 Challenge [34], which includes 2000 lesion images in JPEG format and 2000 corresponding binary mask images in PNG format. Among these 2000 samples, 1280 samples are for training, 320 samples for validation, and the rest for testing and we resized the images to 256 × 256 pixels in the experiment. For the target pixels, 0 and 255 represent the area outside the lesion and the area inside the lesion, respectively.

4.1.2. Cell Nuclei Segmentation

In this part, the dataset used was from The Spot Nuclei Speed Cures [35]. It consists of 670 2D samples and corresponding label images for cell nuclei segmentation. For our experiment, the training set includes 600 samples, and the other 70 images are for testing. Besides, the original images were resized to 96 × 96 pixels.

4.1.3. Coronary Artery in Angiography Segmentation

A total of 130 X-ray coronary angiography and 130 corresponding mask images in PGM format were used in this work, each angiogram being 300 × 300 pixels. The Cardiology Department of the Mexican Social Security Institute, UMAE T1-León provided the whole image database, and the ethics approval for its use in the present research, under reference R-2019-1001-078 [36]. Among these 130 samples, the training set includes 100 images, and the other 30 images are for testing.

4.2. Evaluation Metrics

In order to quantitatively analyze the experimental results, several performance metrics were considered, including F1-score, Intersection over Union (IoU), dice coef-

ficient (DC) [37], and sensitivity (SE). The calculation method of IoU and SE are shown in Equations (4) and (5) respectively.

$$IoU = \frac{TP}{TP + FN + FP} \tag{4}$$

$$SE = \frac{TP}{TP + FN} \tag{5}$$

The DC is expressed as in Equation (6).

$$DC = \frac{2|GT \cap SR|}{|GT| + |SR|} \tag{6}$$

4.3. Results

4.3.1. Skin Cancer Segmentation

We used the Adam optimizer with a learning rate of 3×10^{-4}. Besides, epoch and batch size were set to 4 and 150 respectively. The training time of the proposed model was about 10 h.

Table 1 summarizes the quantitative results for the comparison between this experiment and other methods. We tried different sets of hyperparameters for the optimization of the Residual-Attention UNet++ model. Hyperparameters tuning was conducted manually by training the model with different sets of hyperparameters and evaluating their results; Table 2 lists the hyperparameters adjusted during the experiment. The results of Residual-Attention UNet++ against other methods with respect to F1-score, SE, IoU, and DC are presented in Table 1, which shows that the proposed model achieved the highest results for the task of skin cancer segmentation. Especially, the proposed model increased the F1-score by 0.66%, the SE by 1.35%, the IoU by 0.64%, and the DC by 0.75% over UNet++.

Table 1. Experimental performance of Residual-Attention UNet++ and other methods on skin datasets.

Methods	F1-Score	SE	IoU (%)	DC (%)
FCN	0.8671	0.8748	81.45	86.58
SegNet	0.8862	0.8801	81.58	87.81
U-Net	0.8880	0.8884	81.49	87.79
Attention UNet	0.8866	0.8965	81.46	87.81
UNet++	0.8896	0.8991	81.79	87.93
Attention UNet++	0.8938	0.8942	82.21	88.54
Residual Attention UNet++	0.8955	0.9112	82.32	88.59

Table 2. Hyperparameters adjusted in the experiment.

Hyperparameter	
Batch size	64, 32, 16, 8, 4
Learning rate	0.0001, 0.0003
Optimizer	Adam, SGD
Lr scheduler	ExponentialLR, StepLR

Figure 11 shows some of the example outputs from the testing phase. As observed in Figure 11, the target lesions were segmented accurately with almost the same shape of ground truth in most cases. In addition, it can be clearly seen that the input image in the second row contains two spots; the Residual-Attention UNet++ still segmented the desired target accurately. Moreover, if we examine the third and fourth rows in Figure 11, the UNet++ and Residual-Attention UNet++ predicted two lesions and one lesion respectively, while ground truth only had one, which shows the robustness of the proposed model. In addition, Figure 12 shows the comparison between Residual-Attention UNet++ and other

methods on F1-score after multiple experiments, which further illustrates the robustness of the proposed model.

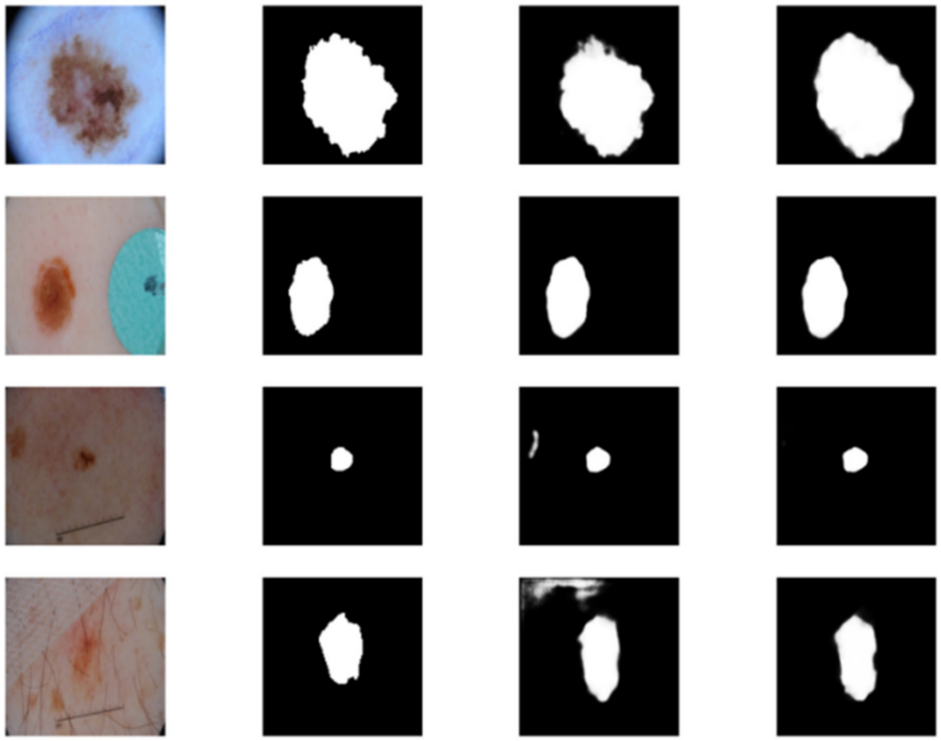


Figure 11. Qualitative assessment of UNet++ and Residual-Attention UNet++ performance on skin cancer. First column: original image, second column: ground truth, third column: UNet++'s output, last column: Residual-Attention UNet++'s output.

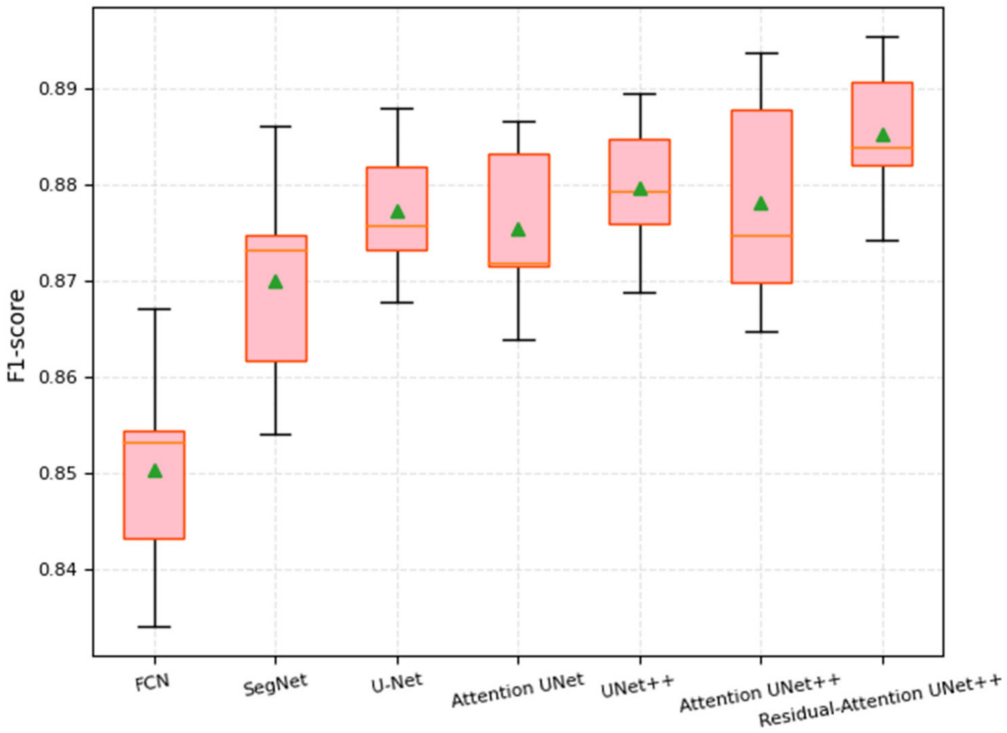


Figure 12. The comparison between Residual-Attention UNet++ and other methods on F1-score after multiple experiments for skin dataset.

4.3.2. Cell Nuclei Segmentation

In this part, Adam was selected as the optimizer, and the initial learning rate was set to 3×10^{-4}. In addition, epoch and batch size were set to 4 and 150 respectively. The training time of the proposed model was about 5 h.

Table 3 shows the summary of how well the proposed model performed against UNet++ and other existing method. As observed, Residual-Attention UNet++ improved some metrics that are critical for the semantic segmentation task, including IoU and DC, by 3.9% and 16.87% compared with the UNet++, respectively.

Table 3. Experimental performance of Residual-Attention UNet++ and other methods on cell nuclei datasets.

Methods	F1-Score	SE	IoU (%)	DC (%)
FCN	0.6260	0.7008	50.80	62.65
SegNet	0.8407	0.9051	77.59	81.26
U-Net	0.8773	0.9381	83.84	82.17
Attention UNet	0.8719	0.9362	83.51	76.49
UNet++	0.8609	0.9367	81.54	73.51
Attention UNet++	0.8771	0.9344	84.40	81.17
Residual Attention UNet++	0.8831	0.9493	87.74	85.91

It is well known that the correct definition of class boundary in the segmentation task of medical images is crucial for subsequent treatment, but it is not easy. Figure 13 shows the qualitative results for the UNet++ and Residual-Attention UNet++ models, respectively. It demonstrates that the output of the proposed approaches showed better segmentation with accurate contour. In addition, the segmented cell nuclei boundary of the proposed model was smoother and clearer than that of UNet++, which is a better match with ground truth. Moreover, Figure 14 shows the comparison between Residual-Attention UNet++ and other methods on F1-score after multiple experiments, which further illustrates the robustness of the proposed model.

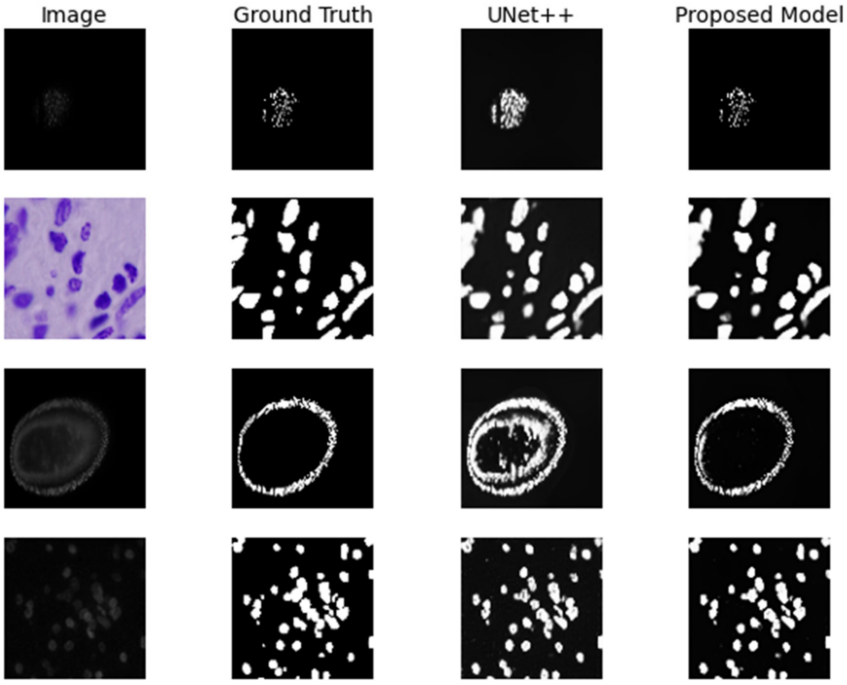


Figure 13. Qualitative assessment of the performance of UNet++ and Residual-Attention UNet++ on cell nuclei.

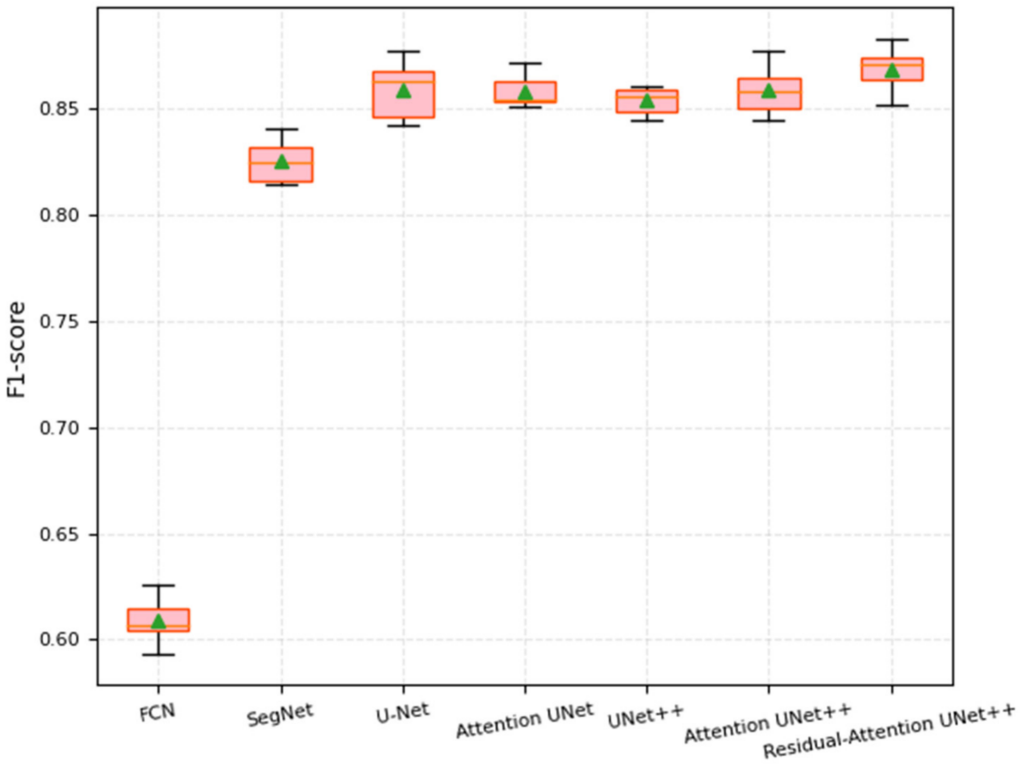


Figure 14. The comparison between Residual-Attention UNet++ and other methods on F1-score after multiple experiments for cell nuclei dataset.

4.3.3. Coronary Artery in Angiography Segmentation

We chose Adam as the optimizer, and the initial learning rate was 3×10^{-4}. As for epoch and batch size, they were set to 250 and 4 respectively. The training time of the proposed model was about 3 h.

Table 4 summarizes the quantitative results for the comparison between this experiment and other methods. As observed, in terms of SE, IoU, and DC indicators, Residual-Attention UNet++ was higher than UNet++ by, respectively, 2.66%, 1.46%, and 2.46%. In addition, compared with FCN, we can see that all evaluation indicators greatly improved, which further illustrates the superiority of the model.

Table 4. Experimental performance of Residual-Attention UNet++ and other methods on Angiography datasets.

Methods	F1-Score	SE	IoU (%)	DC (%)
FCN	0.5330	0.7250	55.12	68.34
SegNet	0.5631	0.7506	59.81	69.40
U-Net	0.5841	0.7865	64.15	69.17
Attention UNet	0.5904	0.8034	65.48	69.84
UNet++	0.5907	0.8119	65.11	70.02
Attention UNet++	0.5901	0.8135	65.51	70.47
Residual Attention UNet++	0.6110	0.8335	66.57	72.48

Figure 15 shows the qualitative results for the UNet++ and Residual-Attention UNet++ models, respectively. As observed, compared with UNet++, the definition of the coronary artery edge in the proposed model was closer to the ground truth.

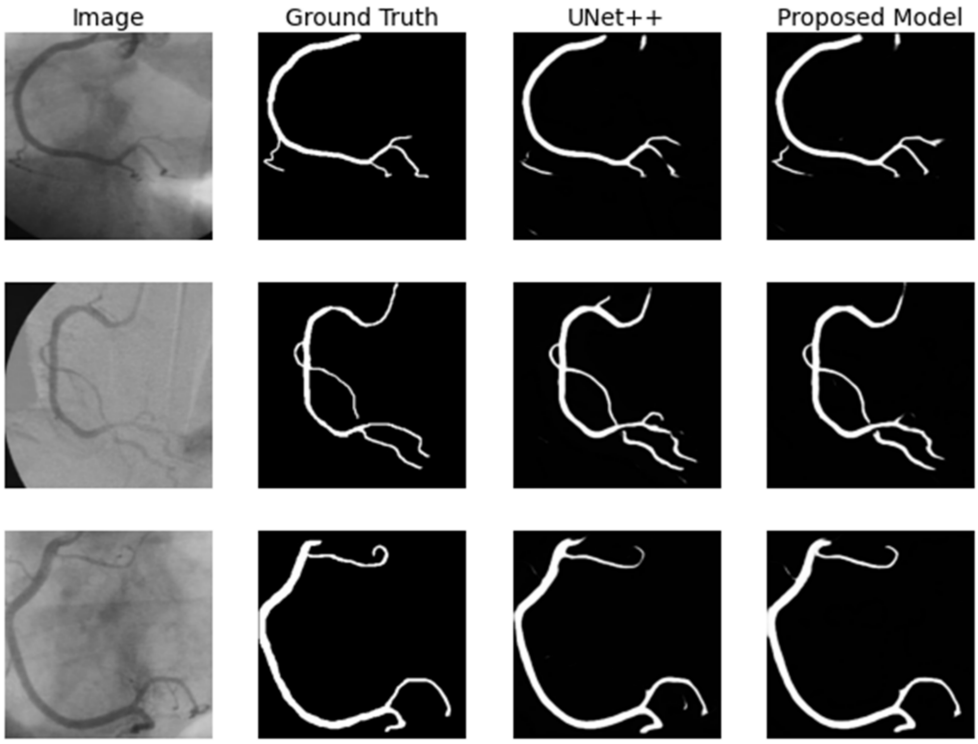


Figure 15. Qualitative assessment of the performance of UNet++ and Residual-Attention UNet++ on coronary artery in angiography.

Figure 16 shows the comparison between Residual-Attention UNet++ and other methods on IoU after multiple experiments, which further illustrates the excellent performance and robustness of the proposed model.

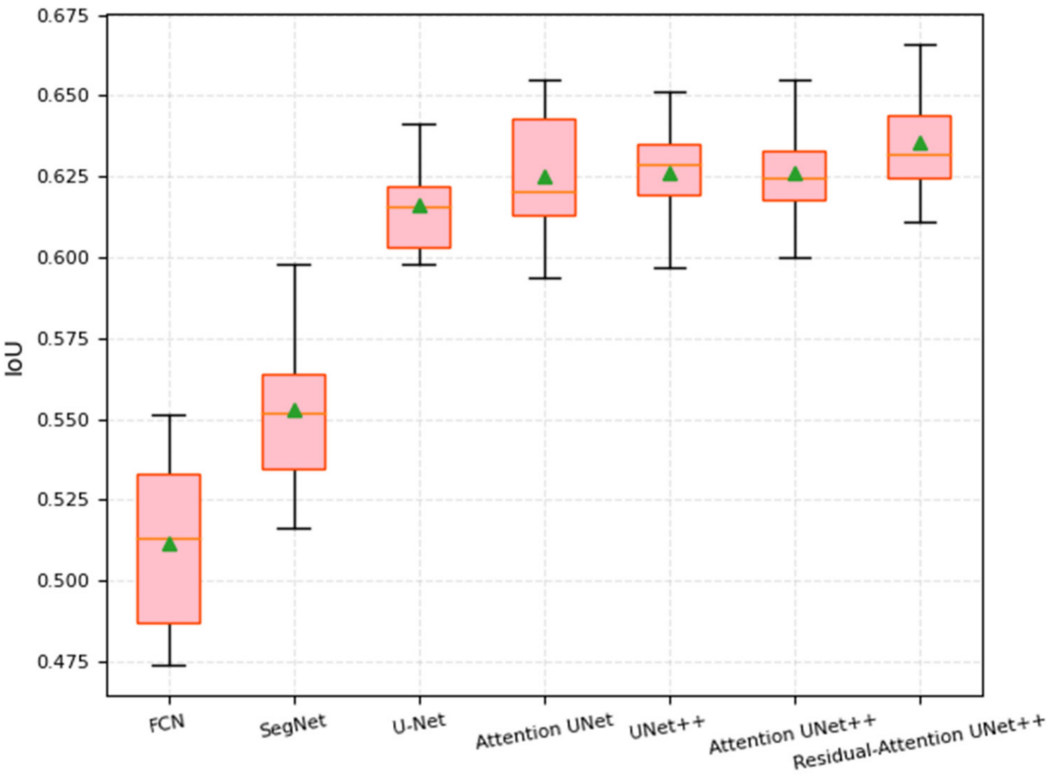


Figure 16. The comparison between Residual-Attention UNet++ and other methods on IoU after multiple experiments for angiography dataset.

4.4. Model Pruning

Figure 17 shows the Inference time, dice coefficient, and parameters of Residual-Attention UNet++ under different pruning degrees. As seen, Residual-Attention UNet++ L^3 achieved, on average, 8.696% reduction in inference time and 75.635% reduction in parameters while degrading dice coefficient by only 3.167%. As for the Residual-Attention UNet++ L^1, it achieved, on average, 15.217% reduction in inference time and 98.915% reduction in parameters while degrading the dice coefficient by 21.834%. It can be seen that pruning can not only reduce the model parameters and inference time, but also affect the segmentation performance. Therefore, it is quite crucial to choose a reasonable pruning strategy according to the actual scenario. Furthermore, considering that most deep CNN segmentation models have long inference times and require large computational resources, it makes sense to apply the pruned models to small computers and mobile devices for CAD.

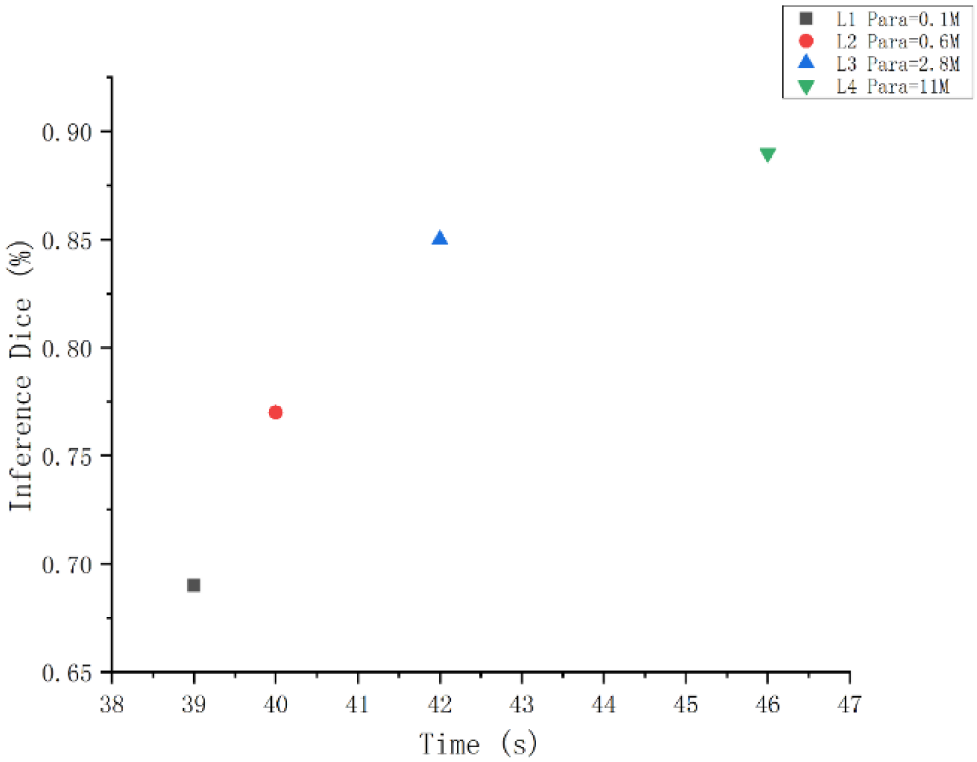


Figure 17. Inference time, dice coefficient and parameters of Residual-Attention UNet++ $L^{1\sim4}$ for the skin cancer segmentation.

5. Conclusions

In this paper, we proposed an extension of the UNet++ architecture using residual unit and attention mechanism. The proposed models are called "Residual-Attention UNet++". We used three different medical image datasets to evaluate our method. The experimental results demonstrated that the proposed Residual-Attention UNet++ model showed better performance in segmentation tasks when compared with existing methods, including the UNet++ and other models on both three datasets. In addition, with the introduction of deep supervision, the pruned Residual-Attention UNet++ enabled faster inference at the cost of minimal performance degradation. Last but not least, considering the current two-dimensional medical image segmentation, without considering the spatial dimension, more research on three-dimensional medical images will be carried out in the future.

Author Contributions: Conceptualization and methodology, Z.L. (Zan Li) and H.Z.; software, validation, Z.L. (Zan Li), Z.L. (Zhengzhen Li) and Z.R.; writing—original draft preparation, Z.L. (Zan Li) and H.Z.; review and editing, Z.L. (Zan Li) and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Young, K.D.; Utkin, V.I.; Ozguner, U. A Control Engineer's Guide to Sliding Mode Control. *IEEE Trans. Control. Syst. Technol.* **1999**, *7*, 328–342. [CrossRef]
2. Rother, C. GrabCut: Interactive foreground extraction using iterated graph cut. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
3. Davis, L.S.; Rosenfeld, A.; Weszka, J.S. Region Extraction by Averaging and Thresholding. *IEEE Trans. Syst. Man Cybern.* **1975**, *3*, 383–388. [CrossRef]
4. Senthilkumaran, N.; Rajesh, R. Edge Detection Techniques for Image Segmentation—A Survey of Soft Computing Approaches. In Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, India, 27–28 October 2009; pp. 844–846.
5. Nowozin, S.; Lampert, C.H. Structured Learning and Prediction in Computer Vision. *Found. Trends®Comput. Graph. Vis.* **2011**, *6*, 185–365. [CrossRef]
6. Sulaiman, S.N.; Isa, N.M. Adaptive fuzzy-K-means clustering algorithm for image segmentation. *IEEE Trans. Consum. Electr.* **2010**, *56*, 2661–2668. [CrossRef]
7. Ostu, N.; Nobuyuki, O.; Otsu, N. A thresholding selection method from gray level histogram. *IEEE SMC-8* **1979**, *9*, 62–66.
8. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *arXiv* **2018**, arXiv:1802.06955.
9. Xi, H.; Chen, J.; Chen, L.; Liang, H.; Wang, Q. Pyramid Residual Convolutional Neural Network based on an end-to-end model. In Proceedings of the 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA), Xi'an, China, 24–25 October 2020; pp. 154–158.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; Volume 39, pp. 3431–3440.
11. Song, W.; Zhong, B.; Sun, X. Building Corner Detection in Aerial Images with Fully Convolutional Networks. *Sensors* **2019**, *19*, 1915. [CrossRef]
12. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Comput. Sci.* **2014**, 357–361. [CrossRef]
13. Krähenbühl, P.K.V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Adv. Neural Inf. Processing Syst.* **2011**, *24*, 109–117.
14. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
15. Liu, X.; Song, L.; Liu, S.; Zhang, Y. A Review of Deep-Learning-Based Medical Image Segmentation Methods. *Sustainability* **2021**, *13*, 1224. [CrossRef]
16. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
18. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
20. Özgün, Ç.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 424–432.

21. 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; pp. 565–571.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
23. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef]
24. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; Mcdonagh, S.; Hammerla, N.Y.; Kainz, B. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
25. Yan, J.; Wang, X.; Cai, J.; Qin, Q.; Yang, H.; Wang, Q.; Cheng, Y.; Gan, T.; Jiang, H.; Deng, J.; et al. Medical image segmentation model based on triple gate MultiLayer perceptron. *Sci. Rep.* **2022**, *12*, 1–14. [CrossRef]
26. Zhou, Z.; Siddiquee, M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
29. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *arXiv* **2016**, arXiv:1605.07146.
30. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
31. Li, C.; Tan, Y.; Chen, W.; Luo, X.; Li, F. ANU-Net: Attention-based Nested U-Net to exploit full resolution features for medical image segmentation. *Comput. Graph.* **2020**, *90*, 11–20. [CrossRef]
32. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
33. Zhou, Z.; Siddiquee, M.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef] [PubMed]
34. ISIC 2017 Challenge. Available online: https://challenge2017.isic-archive.com (accessed on 11 July 2021).
35. Data Science Bowl 2018. Available online: https://www.kaggle.com/c/data-science-bowl-2018 (accessed on 14 June 2021).
36. Cervantes-Sanchez, F.; Cruz-Aceves, I.; Hernandez-Aguirre, A.; Hernandez-Gonzalez, M.A.; Solorio-Meza, S.E. Automatic Segmentation of Coronary Arteries in X-ray Angiograms using Multiscale Analysis and Artificial Neural Networks. *Appl. Sci.* **2019**, *9*, 5507. [CrossRef]
37. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]