mathematics

MDPI

Article

A Method of Lung Organ Segmentation in CT Images Based on Multiple Residual Structures and an Enhanced Spatial Attention Mechanism

Lingfei Wang, Chenghao Zhang, Yu Zhang and Jin Li *

College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; wanglingfei@hrbeu.edu.cn (L.W.); 910381219@hrbeu.edu.cn (C.Z.); zhangyu04@hrbeu.edu.cn (Y.Z.)
* Correspondence: lijin@hrbeu.edu.cn

Abstract: Accurate organ segmentation is a fundamental step in disease-assisting diagnostic systems, and the precise segmentation of lung is crucial for subsequent lesion detection. Prior to this, lung segmentation algorithms had typically segmented the entire lung tissue. However, the trachea is also essential for diagnosing lung diseases. Challenges in lung parenchyma segmentation include the limited robustness of U-Net in acquiring contextual information and the small size of the trachea being mixed up with lung, making it difficult to identify and reconstruct the lungs. To overcome these difficulties, this paper proposes three improvements to U-Net: multiple concatenation modules to enhance the network's ability to capture context, multi-scale residual learning modules to improve the model's multi-scale learning capabilities, and an enhanced gated attention mechanism to enhance the fusion of various hierarchical features. The experimental results demonstrate that our model has achieved a significant improvement in trachea segmentation compared to existing models.

Keywords: lung parenchyma segmentation; multiple concatenation module; multi-scale residual learning module; gated attention mechanism; U-Net

MSC: 92C50; 94A08

Citation: Wang, L.; Zhang, C.; Zhang, Y.; Li, J. A Method of Lung Organ Segmentation in CT Images Based on Multiple Residual Structures and an Enhanced Spatial Attention Mechanism. *Mathematics* **2023**, *11*, 4483. https://doi.org/10.3390/math11214483

Academic Editor: Ming Ma

Received: 26 September 2023
Revised: 24 October 2023
Accepted: 25 October 2023
Published: 30 October 2023

1. Introduction

In recent years, lung disease has become a significant global health concern due to factors such as declining air quality, smoking, and the COVID-19 pandemic. Computed tomography (CT) is widely used in clinical medicine for the diagnosis of lung diseases. However, subjective interpretations by radiologists can lead to misdiagnosis or missed diagnoses. Therefore, the development of a computer-aided diagnostic system is crucial. Automatic segmentation of the lung region is the first step in the diagnosis of lung disease, and its accuracy directly impacts further diagnostic analysis. Lung region segmentation, also known as lung parenchyma segmentation, involves extracting the lung parenchyma from CT images to provide a reliable basis for clinical treatment and pathological studies. This segmentation is a prerequisite for automatic quantitative diagnosis and has important research value, including the establishment of a database of geometric and statistical information on lung structure and the determination of lesion size and extent, which aid doctors in formulating accurate and timely treatment plans. Achieving high accuracy in lung parenchyma segmentation is essential for the success of a computer-aided diagnostic system.

However, lung parenchyma segmentation is a challenging task due to several factors such as differences in lung CT acquisition equipment, the complex structure of the lung tissue, and potential human interference. These factors contribute to uncertainties in the segmentation process, mainly in the following aspects: (1) Thermal/electrical noise, diverse biological tissues, and partial volume effects can cause lung CT images to be blurry. (2) The

lung textures of different diseases can make fine segmentation of lungs difficult. (3) Deep learning-based medical image segmentation algorithms require a large number of images for network training, validation, and testing. (4) Experienced radiologists are required to manually annotate the lung regions after collecting the data, and the quality of the annotated data can greatly impact the segmentation results. As a result, the collection of medical image data at low quantities can affect the accuracy of lung segmentation.

There are two main types of segmentation algorithms: traditional segmentation-based methods and deep learning-based segmentation methods. Traditional segmentation methods include edge-based, region-based, model-based, watershed-based, and clustering-based methods. Generally, the drawbacks of traditional segmentation algorithms are as follows: Edge-based segmentation methods are fast but lack the ability to extract local edge information effectively, leading to the loss of local edge information [1]. Region-based segmentation methods are sensitive to noise and are constrained by seed selection [2,3]. Model-based segmentation methods require segmented regions to satisfy specific conditions [4]. Watershed-based segmentation methods are sensitive to noise and can lead to over-segmentation [5]. The GrabCut algorithm is an interactive segmentation technique that requires users to draw a rectangular bounding box on the image to indicate approximate regions of foreground and background [6]. The algorithm iteratively refines the segmentation results based on these markings. Cluster-based segmentation methods are influenced by the initial clusters [7]. Traditional segmentation methods often involve multiple iterative steps, making it easy to get stuck in local optimal solutions during this process, resulting in incomplete pulmonary parenchyma segmentation and affecting subsequent detection.

Compared with traditional segmentation methods, deep learning-based segmentation methods offer the advantage of not requiring manual outlining of boundaries, which can reduce edge loss and enable end-to-end learning of original features. Convolutional neural networks (CNNs) are the core of deep learning-based machine vision and can effectively characterize images by directly applying visual laws. In the early days of semantic segmentation of medical images based on deep learning, deep convolutional neural networks (DCNN) [8,9] and full convolutional neural networks (FCN) [10–12] were widely used and achieved better results than traditional segmentation algorithms. However, DCNN and FCN algorithms have their own limitations. For example, DCNN's extracted pixel blocks must be classified, leading to large redundancy and slow network training. Although FCN can accept input images of any size, the up-sampled results can be blurry and lack smoothness, can be insensitive to image details, and can lack spatial consistency. Additionally, methods based on deep learning exhibit the following issues: (1) deep learning typically demands a substantial amount of data for support, with the absence of effective extensive training often leading to overfitting; (2) deep learning has the capacity to replicate content within data without understanding it and does not challenge any data or unveil concealed biases, potentially resulting in subjectivity in the generated outcomes; and (3) deep networks exhibit excessive sensitivity to alterations in images.

With the continuous development of research in the field of medical image segmentation, in 2015, the U-Net architecture [13] was proposed. U-Net improves on the FCN network architecture with a symmetric encoder–decoder structure and incorporates a skip connection structure. The encoder extracts the feature map after down-sampling to reduce data size, the decoder gradually restores the resolution by up-sampling, and the skip connection adds a concatenation operation between the encoder and decoder to fuse the shallow and deep information so that the network learns richer information. Shaziya et al. [14] used a network based on U-Net for automatic segmentation of lung parenchyma in chest X-ray images, while the U-Net was extended to a 3D network structure. Almost all subsequent network structures have been based on improvements to U-Net in medical image segmentation. The H-Dense-U-Net model designed by Li et al. [15] solved the problem of a lack of volume information due to 2D convolution and the increased cost of 3D convolution. Damseh et al. [16] proposed a new network structure called Res-U-Net, by introducing a residual structure into U-Net to solve the problem of an

inadequate feature extraction capability. Qamar et al. [17] proposed to use Dense-Net and Inception-ResNet networks in the 3D-U-Net network to form multi-scale features while keeping the spatial resolution unchanged, facilitating the fusion of shallow features with deep features. Yu et al. [18] chose deep convolution and point-by-point convolution to replace the original convolution kernel in the convolution layer of the U-Net, forming a new network structure called DB-U-Net, which greatly enhanced the network's timeliness. Gu et al. [19] designed a multi-scale network structure for lung parenchyma segmentation. These are all improvements to the U-Net backbone structure.

In recent years, more and more attention mechanisms have been proposed and applied to various network structures. For example, Hu et al. [20] proposed to generate channel attention through global pooling and squeeze-and-excitation operations to enhance important features and to suppress non-important features. Oktay et al. [21] proposed the Attention Gate module to enhance the important regions of features and to suppress the non-important regions using a region attention mechanism. Woo S et al. [22] fused the modules of the former two and proposed the CBAM module, which can enhance both spatial features and channel features. The essence of the channel attention mechanism is to learn soft attention between channels through features, and the essence of the spatial attention mechanism is to capture different types of spatial information using non-local mechanisms [23–28] either through fusion of high-level and low-level features [29,30] or by extracting multi-scale information through convolutional kernels with different receptive field sizes to enhance the feature representation of the model [18,31–34]. Although the existing methods can improve accuracy, they generally suffer from the following problems: (1) They ignore the complementary fusion of semantic information of the feature map between different levels and fail to distinguish the importance of different spaces and channels in the feature map; furthermore, there might be issues of gradient vanishing and gradient explosion [35,36]. (2) The attention mechanism is computationally intensive and has redundant parameters [32]. (3) They lack the extraction of information at different scales from the same feature map [37].

This paper proposes an improved MCRAU-Net (Multiple Concatenated, Residual, and Attention U-Net) architecture and conducts experimental validation on a lung CT dataset, segmenting the left and right lung parenchyma independently. The proposed MCRAU-Net enhances U-Net in the following ways: Firstly, a multiple concatenation module is introduced in the encoding structure to increase the network's attention to target location information. Secondly, a multiscale residual module is added to the encoder to enhance the representation of multiscale features. Finally, the gated attention mechanism is improved by positioning it after the skip connection to better aggregate high-level and low-level feature information. Through experimental comparison, MCRAU-Net not only surpasses the U-Net benchmark in terms of segmentation performance but also outperforms other mainstream networks, particularly the TransUNet model with a transformer model and the UneXt model with an MLP.

2. Related Work

2.1. U-Shaped Structure

The original U-Net model uses a symmetric encoding and decoding structure, as shown in Figure 1. If the same size convolution is used, features of the same size can be obtained for each level. Assuming N times down-sampling and up-sampling are used in the U-Net, then a total of $N+1$ layers of feature maps are generated in the U-Net structure, labelling these feature maps as $F_{Encoder1}^{C\times H\times W}$, $F_{Encoder2}^{2C\times \frac{H}{2}\times \frac{W}{2}}$, …, $F_{EncoderN+1}^{2^N C\times \frac{H}{2^N}\times \frac{W}{2^N}}$, $F_{DecoderN}^{2^{N-1}C\times \frac{H}{2^{N-1}}\times \frac{W}{2^{N-1}}}$, $F_{DecoderN-1}^{2^{N-2}C\times \frac{H}{2^{N-2}}\times \frac{W}{2^{N-2}}}$, …,$F_{Decoder1}^{C\times H\times W}$; the input image as F_{input}; and the output segmented image as $F_{output}^{Classes\times H\times W}$. Then, U-Net can be represented by Equations (1)–(5):

$$F_{Encoder1} = DC(F_{input}) \quad (1)$$

$$F_{Encoderi} = DC(DS(F_{Encoderi-1}))1 < i \leq N+1 \tag{2}$$

$$F_{DecoderN} = DC([F_{EncoderN}, US(F_{EncoderN+1})]) \tag{3}$$

$$F_{Decoderi} = DC([F_{Encoderi}, US(F_{Decoderi-1})])1 \leq i < N \tag{4}$$

$$F_{output}^{Classes \times H \times W} = conv_{1\times1}(F_{Decoder1}) \tag{5}$$

where $DC(.)$ represents a double convolution layer, the two connected layers of convolution operation, with each *convolution* operation including the convolution operation with a kernel size of 3×3, the *BatchNormalization* operation, and the *activation* operation; $DS(.)$ represents the down-sampling operation; $[,]$ represents the concatenation operation; $US(.)$ represents the up-sampling operation; $conv_{1\times1}(.)$ represents the *convolution* operation with a kernel size of 1×1; and *classes* represents the number of categories.

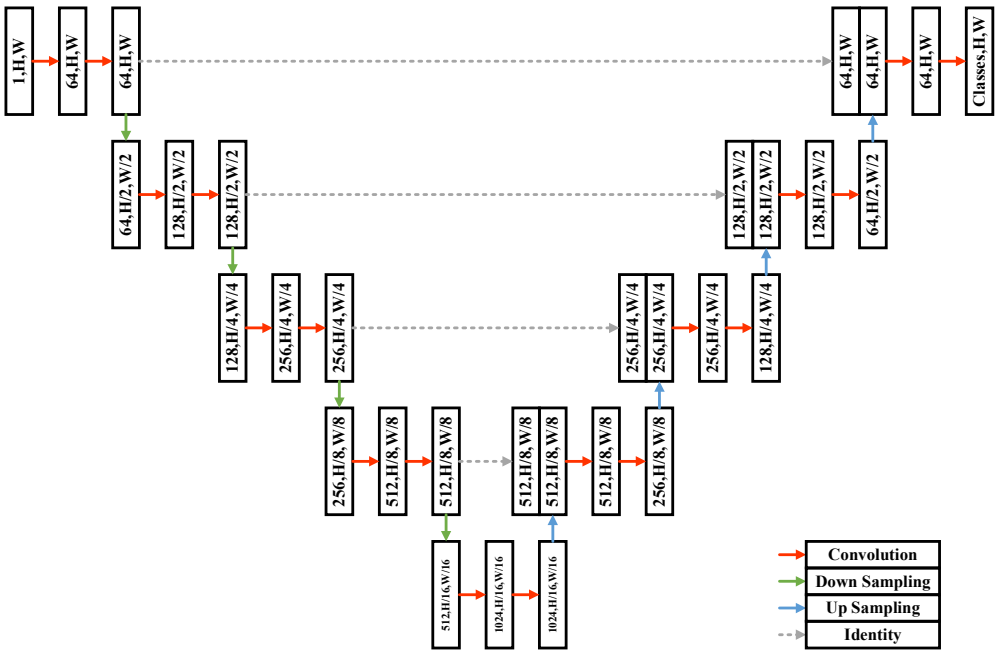

Figure 1. Structure of U-Net [13].

2.2. Residual Connections

As research on deep learning models has intensified, researchers have discovered that the depth of the model plays a critical role in its performance, with deeper networks allowing for the extraction of more complex features. However, optimizing deep networks is plagued by the problem of vanishing gradient, which can lead to the saturation of or even a reduction in network accuracy, as well as a difficulty in training due to vanishing or exploding gradients. To address this issue, DenseNet [35] and ResNet [36] were proposed, which optimize network degradation from two different perspectives. ResNet, in particular, is widely used and was introduced by Kaiming He to resolve the model degradation problem through residual learning, as depicted in Figure 2. This approach helps alleviate the difficulty in training deep CNN models. When the model needs to learn a feature of $H(x)$, it can first learn its residuals $F(x) = H(x) - x$, so that the original learned feature is actually $F(x) + x$. When the residual is 0, the stacking module is equivalent to the constant mapping operation, which does not reduce network performance. However, in practice, the residual will not be 0, allowing the stacking layer to learn new features on top of the input features, thus achieving better performance.

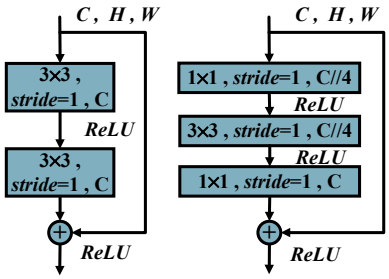

Figure 2. The two types of residual modules used in ResNet [36].

2.3. Gated Attention Mechanisms

The gated attention mechanism is a soft attention mechanism introduced in Attention U-Net [21], as illustrated in Figure 3. This mechanism effectively suppresses non-target regions in features and emphasizes salient features in specific local regions. Furthermore, it is simple to integrate into the standard U-Net network structure with low computational overhead. The attention mechanism in Attention U-Net is mathematically expressed in Equations (6)–(9):

$$AW_{Encoderi} = conv_{3\times3}^{C/2}(F_{Encoderi}) \tag{6}$$

$$AW_{Decoderi} = conv_{3\times3}^{C/2}(US(F_{Decoderi-1})) \tag{7}$$

$$attcoef_i = \sigma(conv_{3\times3}^{1}(ReLU(AW_{Encoderi} + AW_{Decoderi}))) \tag{8}$$

$$F_{Decoderi} = DC([US(F_{Decoderi-1}), F_{Encoderi} * attcoef_i]) \tag{9}$$

where $conv_{m\times m}^{n}$ represents the convolution operation; m is the size of the convolution kernel; n represents the number of channels in the output and uses different convolution kernels to calculate two different attention weight matrices $AW_{encoder}$ and $AW_{decoder}$; $US(.)$ represents the up-sampling operation, where $DC(.)$ represents the double convolution layer and represents the two connected layers of convolution operation, with each convolution operation including the convolution operation with a kernel size of 3×3, the BatchNormalization operation, and the activation operation; $[,]$ represents the concatenation operation; $ReLU(.)$ is the *ReLU* activation function; and $\sigma(.)$ represents the *Sigmoid* activation function.

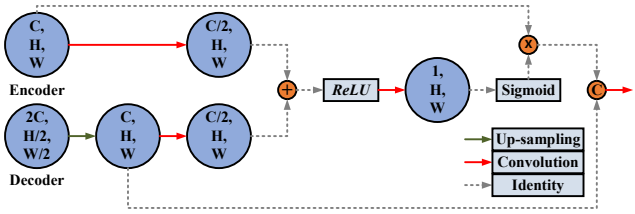

Figure 3. Schematic diagram of the gated attention mechanism [21].

3. Proposed Method

In this paper, we propose a novel algorithm called MCRAU-Net, which offers three main advantages. Firstly, MCRAU-Net incorporates more location information in the encoding layer, which increases the width of the network and enhances its sensitivity to location information. Secondly, a multi-scale residual learning module is used in the feature extraction process, which enriches the perceptual field of the network and adds multi-scale features to further enhance the feature extraction capability of the encoding layer. Finally, we introduce an improved gated attention mechanism to the skip concatenation process, which enhances the regional features of the fusion information. The attention mechanism strengthens the attention to the lung boundaries and suppresses irrelevant regions while

enhancing the sensitivity to location information and facilitating the fusion of shallow features with deep features.

3.1. Structure of Multiple Concatenation Module

Following the proposal of the U-shaped network structure, many researchers have made improvements to its coding layer, decoding layer, and splicing method. For instance, CB-Net introduced a structure that incorporates three convolutional branches fused together to expedite the fusion of deep and shallow features [38]. TransUNet uses CNN as a feature extractor for shallow features and transformer as a feature extractor for deep features, making the deep semantic features of the model more focused on global information [39]. UneXt is a lightweight semantic segmentation model that simplifies the structure of U-Net by replacing the double convolutional structure with a single convolutional structure in the encoder, reducing the number of channels in each layer, and using the *Tokenized MLP* module to fuse the global semantic information in the deep semantic features [40]. The MedT network structure uses a global branching and local branching transformer feature extractor [41]. The UCtransNet uses the transformer to fuse shallow features and deep features [37].

It has been shown that shallow features tend to represent the contour information of an image, and deep features tend to represent the semantic information of a larger perceptual range. In the task of detecting lesions in a single organ, the current dominant task framework is to first segment the organ. To this end, we have redesigned the coding layer structure of the U-Net to improve the accuracy of the detector by incorporating more positional information of the shallow features into the deeper features. This allows the network to better distinguish the positions of the left and right lungs. The improved encoder is shown in Equations (10) and (11):

$$F_{Encoder1} = DC(F_{input}) \tag{10}$$

$$F_{Encoderi} = RC([DC(R(F_{input})), \ldots, DS(F_{Encoderi-1})]) \tag{11}$$

where $RC(.)$ represents the multi-scale residual learning module; $R(.)$ represents the *Image Resize* operation, usually using quadratic interpolation; $DC(.)$ represents the two connected layers of the *convolution* operation with a kernel size of 3×3, the *BatchNormalization* operation, and the *activation* operation; $DS(.)$ indicates the down-sampling operation; and $[,]$ represents the concatenation operation.

3.2. Multi-Scale Residual Learning Module

When using residual units proposed in ResNet for image segmentation tasks, we found that the design of the residual units suffers from a lack of multi-scale learning performance. In the structure of residual unit one, two convolutions of the 3×3 kernel are used, and typically, the perceptual field of two 3×3 sized convolutions is equivalent to that of a 5×5 sized convolution kernel, which only reduces the number of parameters used for convolution. Similarly, one 3×3 convolution and two 1×1 convolutions are used in the structure of residual unit 2, which has a perceptual field of 3×3. In order to add more multi-scale features to the residual learning to suit the medical image segmentation task, we propose a new multi-scale residual module, which adds multi-scale features to the convolution and down-sampling process, respectively, incorporating features of different sizes under the perceptual fields of 3×3 and 5×5, allowing the model to learn a wider range of residuals.

In the research on GhostNet, it is found that there is a large amount of information redundancy between each convolved feature map channel [42]. Therefore, GhostNet uses the method of compressing channel features to accelerate the network training, and the compression ratio is usually 0.5. In our proposed multiscale residual learning module, as shown in Figure 4, the number of channels is compressed to half of the original number of channels for each convolution of 3×3, and the channels are restored by a convolution of

1×1 after fusing the multiscale features. Assuming that the input feature is $f \in R^{C \times H \times W}$, the details of the multiscale residual learning module are as shown in Equations (12)–(16):

$$f_{3\times3} = ReLU(BN(conv_{3\times3}^{C//2}(f))) \tag{12}$$

$$f_{5\times5} = BN(conv_{3\times3}^{C//2}(f_{3\times3})) \tag{13}$$

$$f_{fusion} = ReLU(BN(conv_{1\times1}^{C//2}(f_{3\times3})) + f_{5\times5}) \tag{14}$$

$$f_{residual} = BN(conv_{1\times1}^{C}(f_{fusion})) \tag{15}$$

$$f_{out} = ReLU(f + f_{residual}) \tag{16}$$

where $f_{3\times3}$ and $f_{5\times5}$ are the features under the receptive fields 3×3 and 5×5, respectively; f_{fusion} is a mix of features with different receptive fields; $f_{residual}$ is the residuals learned by the module; f_{out} is the features output by the module; $conv_{m\times m}^{n}$ is the convolution operation; m is the size of the convolution kernel; n is the number of channels output; c is the number of channels output by the convolution; $BN(.)$ is the *BatchNormalization* operation; and $ReLU(.)$ is the *ReLU* activation function.

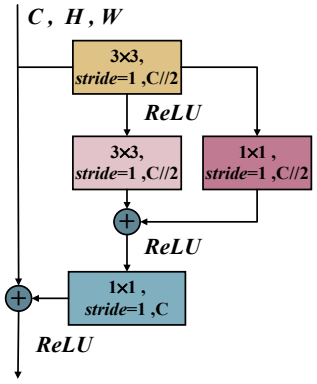

Figure 4. Multi-scale residual cells.

3.3. Improved Gated Attention Mechanism

In the original gated attention mechanism, only the regional features of the encoding layer are enhanced and the regional features of the decoding layer are ignored. Therefore, in order to make similar regions of the encoding and decoding features in the concatenation operation enhanced at the same time, the improved gated attention mechanism described in Equations (17)–(20) is used:

$$AW_{encoderi} = conv_{3\times3}^{C/2}(F_{encoderi}) \tag{17}$$

$$AW_{decoderi} = conv_{3\times3}^{C/2}(US(F_{encoderi-1})) \tag{18}$$

$$attcoef_i = \sigma(conv_{3\times3}^{1}(ReLU(AW_{encoderi} + AW_{decoderi}))) \tag{19}$$

$$F_{Decoderi} = DC([US(F_{Decoderi-1}), F_{encoderi}] * attcoef_i) \tag{20}$$

where $conv_{m*m}^{n}$ represents the convolution operation, where m is the size of the convolution kernel, n represents the number of channels in the output, and different convolution kernels are used to compute $AW_{encoder}$ and $AW_{decoder}$; $US(.)$ represents the up-sampling operation;

$DC(.)$ represents the double convolution operation; $[.]$ represents the concatenation operation; $ReLU(.)$ is the $ReLU$ activation function; and $\sigma(.)$ represents the sigmoid function. This attention mechanism enhances any similar contour features between the encoder and decoder, and its structure is shown in Figure 5.

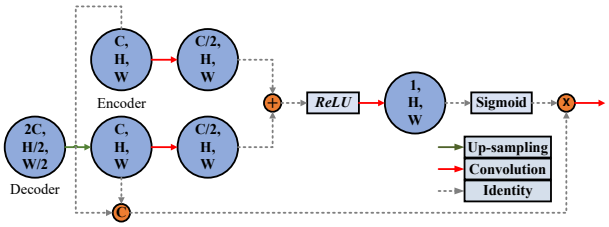

Figure 5. Improved gated attention mechanism module.

3.4. Loss Function

Mixed loss is used in this paper, where CELoss (Cross Entropy Loss) is used to solve the classification of pixels and Dice Loss is used to make the network more focused on the depiction of foreground details. The loss of the network is defined in Equation (21):

$$L_{Total} = L_{CE} + L_{Dice} \tag{21}$$

where L_{Total} represents total loss of the network, L_{CE} represents CEloss, and L_{Dice} represents Diceloss.

L_{CE} is defined in Equation (22):

$$L_{CE}(P_i, \hat{P}_i) = -\frac{1}{N}\sum_i^N (\hat{P}_i \ln(P_i)) \tag{22}$$

where P_i and $\hat{P}_i$ represent the probabilities of a foreground and background, respectively.

L_{Dice} is defined as Equation (23):

$$L_{Dice} = 1 - \frac{2\sum_i^N P_i \hat{P}_i + \varepsilon}{\sum_i^N (P_i^2 + \hat{P}_i^2) + \varepsilon} \tag{23}$$

where P_i and $\hat{P}_i$ represent the probabilities of a foreground and background, respectively; ε is usually a hyperparameter set to prevent the denominator from being 0 and is usually set as 10^{-5}.

4. Experimental Results and Discussions

To evaluate the effectiveness of the proposed method, a subset of the LUNA 2016 dataset [43] was used to segment the lung in this paper. The LUNA16 dataset is indeed derived from the LIDC-IDRI dataset, comprising a total of 888 CT scans. The dataset provides annotations for lung parenchyma and trachea. However, it is important to note that the creators of the LUNA16 dataset have indicated that the annotations for lung parenchyma should not be considered as reference standards for segmentation studies. Therefore, there is a risk of annotation errors in both lung parenchyma and trachea, as these annotations may be inaccurate or incomplete, potentially affecting the accuracy and completeness of research.

In light of this, to assess the potential risks of errors in the annotations of lung parenchyma and trachea, in the context of this paper's theme of small-sample semantic segmentation, we selected 20 cases. The CT scans of these 20 cases were independently reviewed by two radiologists from a tertiary hospital, and all the review processes were conducted using the 3D Slicer platform.

To avoid data contamination, we selected 20 cases from the dataset, out of which 12 cases were used as the training set, 4 cases were used as the validation set, and 4 cases were used as the test set, and then processed them into 2D images. In this context, "data

contamination" refers to the occurrence of different slices of the same case appearing in the training, validation, and test sets. The five-fold cross-validation method has been applied in the experiments. The evaluation metric is the average and standard deviation of a five-fold cross-validation. In the process of lung CT image processing, as the HU value of lung is around -500, we truncated the HU values of CT images to the range of $[-1200, 600]$. Subsequently, we normalized the truncated HU values, converting them into uint8 format data within the range [0, 255], and finally saved them as jpg images. The segmentation task is set to four categories: background, left lung, right lung, and trachea. The proposed MCRAU-Net is illustrated in Figure 6, and further details are provided.

The software composition of the experimental platform is as follows: Windows 11 operating system, Python 3.7 programming language, and pytorch11.2 software. The hardware platform is as follows: Intel i5-12600K CPU, 16G DDR4 4000 MHz memory, and single Nvidia 2080Ti 11G GPU. In the training process of the neural network, the parameters were optimized by an Adam optimizer, the learning rate was set from 10^{-4} to 10^{-6} by the cosine annealing rate, and the batch-size was set to 8. Twenty epochs were trained, and the weights with the smallest losses on the validation set were selected as the optimal weights. In addition, random rotation, translation, and local deformation were used to process the training set to enhance the input diversity of the data. The loss metrics during the training process are shown in Figure 7.

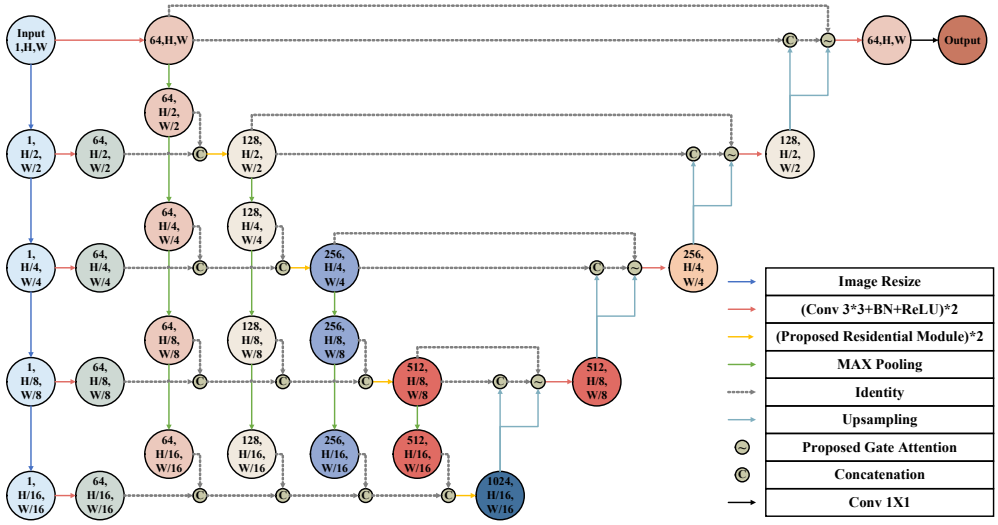

Figure 6. Schematic diagram of MCRAU-Net.

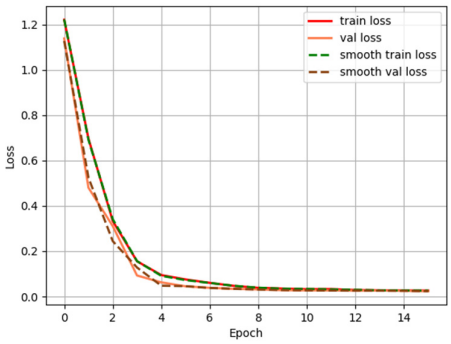

Figure 7. Schematic representation of loss during training.

4.1. Evaluation Metrics

In this paper, we use Precision, Recall, Intersection over Union, Mean Intersection over Union, Mean Pixel Accuracy, and Accuracy, which are widely used in image segmentation tasks, to verify the effectiveness of the proposed algorithm. TP, TN, FP, and FN denote true

positive, true negative, false positive, and false negative, respectively. P_p^1 represents the predicted segmentation result; P_g^1 represents the segmentation result of the ground truth; k denotes the number of categories; p_{ii} is the total number of pixels from the actual category i predicted to category i; and p_{ij} is the total number of pixels from the actual category predicted to the category j. The evaluation metrics can be formulated as follows:

$$P = \frac{TP}{TP + FP} \tag{24}$$

where P represents the probability that the outcomes predicted to be positive samples are actually positive samples.

$$R = \frac{TP}{TP + FN} \tag{25}$$

where R denotes the probability that the outcome with a true positive sample is actually a positive sample.

$$IoU = \frac{P_p^1 \cap P_g^1}{P_p^1 \cup P_g^1} \tag{26}$$

where IoU represents the ratio of the intersection of the predicted outcome for a category and the true value to the merged set, which for image segmentation, is the calculation of the intersection and merging ratio between the prediction mask and the true mask.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN + FP + TP} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_p \cap P_g}{P_p \cup P_g} \tag{27}$$

where $MIoU$ is the average ratio of the intersection in each category to the merged set.

$$MPA = \frac{1}{k+1}\frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij}} \tag{28}$$

where MPA represents the average accuracy of the predicted results.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{29}$$

$Accuracy$ represents the proportion of the total sample that was correctly classified.

In order to evaluate the effect of reconstruction from 2D images to 3D images, the proposed algorithm is evaluated by using 3D metrics commonly used for medical segmentation tasks. ${P_p}^2$ represents the predicted sample volume, ${P_g}^2$ represents the sample volume of ground truth, $S(A)$ represents the surface voxels in the A set, and $d(v, S(A))$ represents the shortest distance from any voxel to $S(A)$. The evaluation indexes can be formulated as follows:

$$DICE = \frac{2\left|P_p^2 \cap P_p^2\right|}{\left|P_p^2\right| + \left|P_p^2\right|} \tag{30}$$

where $DICE$ represents the Dice similarity coefficient: the closer it is to 1, the better the segmentation performance is proven to be.

$$VOE = \frac{2\left|P_p^2 - P_p^2\right|}{\left|P_p^2\right| + \left|P_p^2\right|} \tag{31}$$

where *VOE* represents the Volumetric Overlap Error: the lower the VOE, the better the segmentation effect.

$$RVD = \left(\frac{\left| P_p^2 \right|}{\left| P_g^2 \right|} - 1 \right) \times 100\% \tag{32}$$

where *RVD* represents the Relative Volume Difference:

$$ASD(A,B) = \frac{\left\{ \sum_{s_A \in S(A)} d(s_A \in S(B)) + \sum_{s_B \in S(B)} d(s_B \in S(A)) \right\}}{|S(A)| + |S(B)|} \tag{33}$$

where $ASD(A,B)$ represents the Average Symmetric Surface Distance:

$$\begin{aligned} MSD(A,B) = \max(& \max_{s_A \in S(A)} (d(s_A \in S(B))), \\ & \max_{s_B \in S(B)} (d(s_B \in S(A)))) \end{aligned} \tag{34}$$

where *MSD* represents the maximum symmetric surface distance, with a lower value of *MSD* indicating a better match between the two samples.

4.2. Ablation Study

This paper proposes a model, called MCRAU-Net, which includes multiple concatenation and multi-scale residual learning modules in the encoding structure to enhance network feature extraction and to obtain rich shallow semantic information. Furthermore, an improved gate attention module is added to the skip connection to enhance the characterization of features after the fusion of deep and shallow features. To evaluate the effectiveness of the proposed modules in MCRAU-Net, we conducted ablation experiments on our built dataset. We deleted the multiple concatenation, multi-scale residual learning, and gate attention modules in MCRAU-Net and performed the experiments. Firstly, we used the U-Net as the baseline network and compared its performance with the MCRU-Net, which included the multiple concatenation and multi-scale residual learning modules, to verify their effectiveness. Secondly, we added the multi-scale residual learning and improved the gate attention modules to the baseline network, forming the RAU-Net, to verify their effectiveness. Thirdly, we added the multiple concatenation and improved the attention modules to the baseline network, forming the MCAU-Net, and evaluated its validity. Finally, we verified the validity of our proposed method, MCRAU-Net, through experiments. Table 1 presents the results of our experiments.

As can be seen from Table 1, the metrics obtained for each module used in the MCRAU-Net network compared to the benchmark U-Net prove the effectiveness of the proposed method. Compared to U-Net, MCAU-Net, MCRU-Net, and MRAU-Net, all three network structures showed an increase in segmentation metrics. When the MRAU-Net, MCRU-Net, and MCAU-Net segmentation metrics were compared, the MCAU-Net metrics were better than the MCRU-Net and MRAU-Net metrics in $IOU_{Trachea}$ and $Precision_{Trachea}$. Compared to Unet, MCRAU-Net has shown an improvement of 3.8% in IOU and 3.9% in accuracy specifically within the tracheal region. After removing the attention model, the IOU and accuracy of the tracheal region decreased by 1.8% and 1.9%, respectively. After eliminating the multiple concatenation structure, the IOU and accuracy of the tracheal region decreased by 2.2% and 1.7%, respectively. After removing the residual connections, the IOU and accuracy of the tracheal region decreased by 3.2% and 2.8%, respectively. Therefore, the most effective module for network performance improvement is the gate attention module, followed by the multiple concatenation module and, finally, the proposed multi-scale residual learning module. Compared with the three network structures of MCRU-Net, MCAU-Net, and MRAU-Net, the segmentation metrics of MCRAU-Net are better than

those of MCAU-Net, MCRU-Net, and MRAU-Net. The effectiveness of the algorithm proposed in this paper is proven.

Table 1. Results of ablation experiments (%, average ± standard deviation).

	U-Net	MCRU-Net	MRAU-Net	MCAU-Net	MCRAU-Net
IOULeft-Lung	98.43 ± 0.22	98.1 ± 0.18	98.32 ± 0.18	98.23 ± 0.19	98.49 ± 0.16
IOURight-Lung	98.49 ± 0.13	98.59 ± 0.17	98.51 ± 0.17	98.47 ± 0.16	98.52 ± 0.18
IOUTrachea	80.74 ± 1.61	81.37 ± 1.71	82.37 ± 1.68	82.64 ± 1.82	84.53 ± 1.77
mIOU	94.6 ± 0.45	94.41 ± 0.46	94.71 ± 0.43	94.78 ± 0.45	95.29 ± 0.42
RecallLeft-Lung	99.18 ± 0.18	98.98 ± 0.23	99.02 ± 0.19	98.98 ± 0.22	99.26 ± 0.21
RecallRight-Lung	99.29 ± 0.18	99.2 ± 0.18	99.1 ± 0.18	99.23 ± 0.19	99.31 ± 0.18
RecallTrachea	95.32 ± 0.82	95.0 ± 1.01	94.58 ± 0.98	94.83 ± 1.01	95.54 ± 1.14
PrecisionLeft-Lung	99.23 ± 0.18	99.1 ± 0.25	99.29 ± 0.30	99.24 ± 0.22	99.28 ± 0.34
PrecisionRight-Lung	99.19 ± 0.12	99.38 ± 0.13	99.4 ± 0.18	99.23 ± 0.15	99.22 ± 0.13
PrecisionTrachea	84.13 ± 1.53	85.16 ± 1.52	86.34 ± 1.48	86.55 ± 1.44	88 ± 1.4
MPA	98.42 ± 0.26	98.47 ± 0.28	98.15 ± 0.32	98.23 ± 0.33	98.48 ± 0.36
ACC	99.65 ± 0.14	99.76 ± 0.1	99.77 ± 0.11	99.76 ± 0.09	99.78 ± 0.04

4.3. Comparison Experiments

MCRAU-Net was compared with U-Net, Vgg16-UNet, Resnet50-UNet, U-Next, and TransUNet. It is important to note that all networks were trained from scratch, and no transfer learning information was utilized. Table 2 shows the segmentation metrics corresponding to the different network structures, and compared with the six mainstream networks U-Net, Vgg16-UNet, Resnet50-UNet, UneXt and TransUNet, the proposed network showed a significant improvement, especially in Precision and IOU for the Trachea, the range of which is from 1.9% to 9%. The improvement in segmentation metrics represents an enhancement in network performance, compared with other networks, stemming from the MCRAU-Net's new encoding and decoding structure, which enhances not only the network's feature extraction capability but also the network's recovery capability. Therefore, the performance of MCRAU-Net is better than other mainstream structures and the new network can be well used for semantic segmentation of lung parenchymal CT images.

Table 2. Cross-sectional comparison of MCRAU-Net with mainstream segmentation algorithms (%, average ± standard deviation).

	U-Net	Vgg16-UNet	Resnet50-UNet	UneXt	TransUNet	UCTransNet	MCRAU-Net
IOULeft-Lung	98.43 ± 0.22	98.14 ± 0.34	98.38 ± 0.32	98.1 ± 0.21	98.47 ± 0.17	98.52 ± 0.17	98.49 ± 0.16
IOURight-Lung	98.49 ± 0.13	98.29 ± 0.14	98.47 ± 0.33	98.19 ± 0.25	98.4 ± 0.19	98.6 ± 0.22	98.52 ± 0.18
IOUTrachea	80.74 ± 1.61	79.1 ± 3.56	80.04 ± 2.93	75.26 ± 5.42	82.04 ± 2.08	82.91 ± 2.28	84.53 ± 1.77
mIOU	94.6 ± 0.45	93.75 ± 0.99	94.14 ± 0.77	92.73 ± 1.41	94.59 ± 0.51	94.88 ± 0.61	95.29 ± 0.42
RecallLeft-Lung	99.18 ± 0.18	99.07 ± 0.14	99.23 ± 0.34	98.92 ± 0.22	99.15 ± 0.19	99.22 ± 0.19	99.26 ± 0.21
RecallRight-Lung	99.29 ± 0.18	99.13 ± 0.17	99.19 ± 0.11	99 ± 0.35	99.32 ± 0.17	99.3 ± 0.18	99.31 ± 0.18
RecallTrachea	95.32 ± 0.82	95.75 ± 1.04	96.42 ± 1.68	94.74 ± 1.17	95.67 ± 1.08	95.71 ± 0.92	95.54 ± 1.14
PrecisionLeft-Lung	99.23 ± 0.18	99.05 ± 0.31	99.23 ± 0.12	99.16 ± 0.28	99.3 ± 0.13	99.26 ± 0.11	99.28 ± 0.34
PrecisionRight-Lung	99.19 ± 0.12	99.15 ± 0.19	99.03 ± 0.43	99.17 ± 0.31	99.07 ± 0.23	99.21 ± 0.13	99.22 ± 0.13
PrecisionTrachea	84.13 ± 1.53	79.87 ± 6.56	82.87 ± 3.39	78.62 ± 6	85.16 ± 2.45	85.91 ± 2.27	88 ± 1.4
MPA	98.42 ± 0.26	98.44 ± 0.28	98.04 ± 0.92	98.1 ± 0.26	98.48 ± 0.32	98.51 ± 0.31	98.48 ± 0.36
ACC	99.65 ± 0.14	99.66 ± 0.02	99.69 ± 0.02	99.63 ± 0.05	99.69 ± 0.03	99.72 ± 0.02	99.78 ± 0.04
Parameters	34.52 M	24.89 M	43.93 M	1.47 M	105.32 M	65.6 M	48.5 M

When the CT dataset was reconstructed in 3D, the 2D data corresponded to the area of the lung and the 3D data corresponded to the volume of the lung. To further demonstrate the superiority of the MCRAU-Net network, validation was performed with 3D segmentation metrics. The 2D data obtained from U-Net, Vgg16-UNet, Resnet50-UNet, UneXt, TransUNet, UCTransNet, and MCRAU-Net were converted into 3D data, and the specific 3D segmentation metrics were compared as shown in Table 3. In the 3D

segmentation metrics, the metrics corresponding to the MCRAU-Net network were higher than the other networks in terms of the DICE, IOU, VOE, and ASSD metrics, and the good performance of the MCRAU-Net network was also reflected in the segmentation metric RVD. Therefore, the superior performance of the 3D segmentation metrics also demonstrates that the proposed algorithm can be well used for semantic segmentation of lung parenchymal CT images.

Table 3. Comparison of various indicators after 3D reconstruction of segmented lung images (%, average ± standard deviation).

	U-Net	UneXt	TransUNet	UCTransNet	MCRAU-Net
DICE	97.67 ± 0.25	97.86 ± 0.23	96.69 ± 0.25	97.9 ± 0.20	98.01 ± 0.18
IOU	95.54 ± 0.26	95.61 ± 0.32	94.14 ± 0.29	95.78 ± 0.24	96.17 ± 0.22
VOE	0.0435 ± 0.045	0.0389 ± 0.057	0.057 ± 0.049	0.0432 ± 0.043	0.0372 ± 0.031
RVD	0.0079 ± 0.0038	0.0068 ± 0.002	0.0265 ± 0.0031	0.0151 ± 0.0029	0.0053 ± 0.0026
ASSD	0.439 ± 0.18	0.288 ± 0.15	1.06 ± 0.32	0.2946 ± 0.19	0.2278 ± 0.12
MSD	83.64 ± 8.86	68.96 ± 10.32	86.28 ± 15.32	78.58 ± 9.68	63.07 ± 7.82

4.4. Visual Comparison of Experimental Results

In Figure 8, subfigures (a) and (b) represent the original image and the gold standard of the lung, and from subfigure (c) to (i) are the results of Vgg16-UNet, U-Net, Resnet50-UNet, UneXt, TransUNet, and MCRAU-Net segmentation, respectively. Several networks have some drawbacks, with under-segmentation and over-segmentation problems.

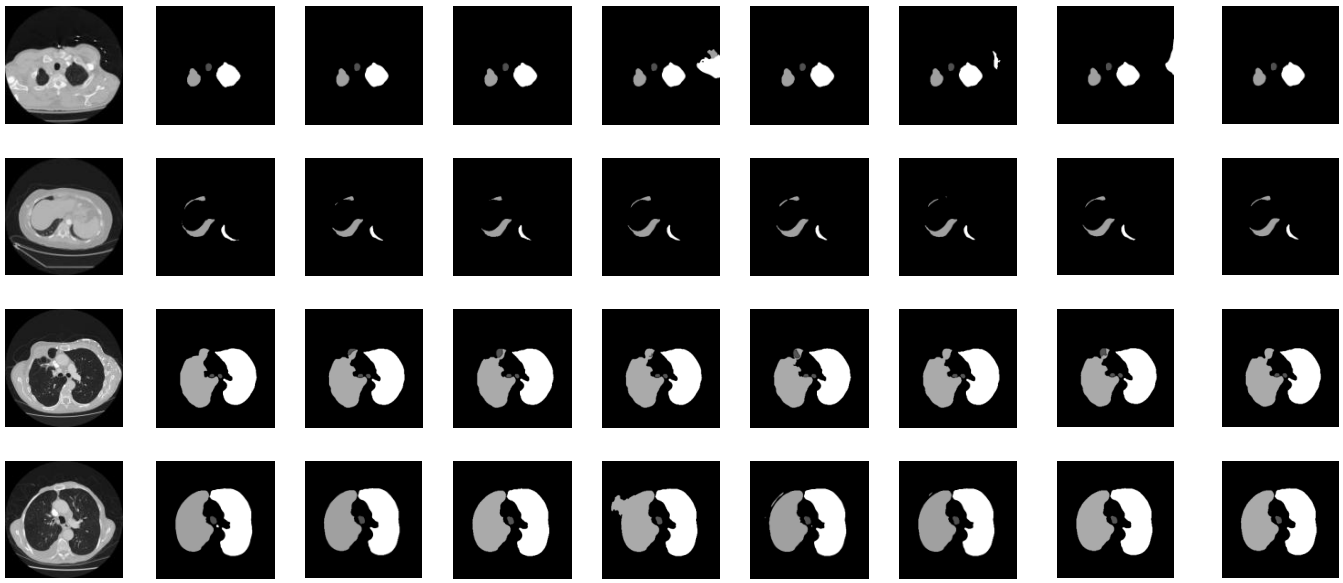

Figure 8. Visual comparison of segmentation results of different algorithms. (**a**) Original CT Image. (**b**) Gold Standard. (**c**) Resnet50-UNet. (**d**) Vgg16-Unet. (**e**) U-Net. (**f**) UneXt. (**g**) TransUNet. (**h**) UCTransNet. (**i**) MRACU-Net.

In the first row of images, U-Net, TransUNet, and UCTransNet all have the issue of misclassifying parts of the background as lung tissue, with U-Net exhibiting the most prominent segmentation errors. In the second row of images, ResNet50-UNet, Vgg16-UNet, and TransUNet all exhibit under-segmentation, with insufficient segmentation of the lung tissue in the left lung. In the third row of images, ResNet50-UNet, Vgg16-UNet, UneXt, and UCTransNet misclassify the holes in the left lung as trachea. In the fourth row of images, U-Net, UneXt, and TransUNet all have the issue of over-segmentation, misclassifying parts of the background as left lung. Based on these observations, we

can conclude that the feature extraction capabilities of Vgg16-UNet, U-Net, ResNet50-UNet, UneXt, and TransUNet are limited, with incomplete fusion of shallow and deep information. In contrast, our proposed MCRAU-Net effectively addresses the above-mentioned limitations and accurately segments the edge regions of the lung, outperforming other mainstream networks.

The encoding structure of our network model utilizes the multiple concatenation module and multi-scale residual module to enhance multi-scale features and to improve the feature acquisition capability. This strengthens the ability of the network to extract both boundary and location information. In the decoding phase, we employ an improved attention mechanism module and skip connections to enhance the attention to both boundary and semantic information. This promotes the fusion of the two types of features and strengthens the reduction capability of the decoding part of the network. As a result, our network is capable of reducing under-segmentation and over-segmentation in the segmentation process.

4.5. Heatmap Analysis

The feature map of each layer of MCRAU-Net is generated as a heatmap, which is able to represent the importance of each position for that class. As can be seen from Figure 9, we can intuitively understand the operation of each convolutional layer in MCRAU-Net. In the encoding part, each convolutional layer contains the boundary information of the lung parenchyma; in the decoding part, each convolutional layer contains the location information of the lung, especially the last layer of MCRAU-Net (Decoder1), which contains rich edge information. And, the feature heat maps of the last layer of Vgg16-UNet, U-Net, Resnet50-UNet, UneXt, TransUNet, UCTransNet, and MCRAU-Net were extracted, as shown in Figure 10. As can be seen, the focus on the lung region of MCRAU-Net is significantly better than that of the other mainstream networks, while the lung region has the darkest color. MCRAU-Net captures the semantic information of the lung well and effectively avoids over-segmentation and under-segmentation, demonstrating that MCRAU-Net can be better used for semantic segmentation of the lung parenchyma.

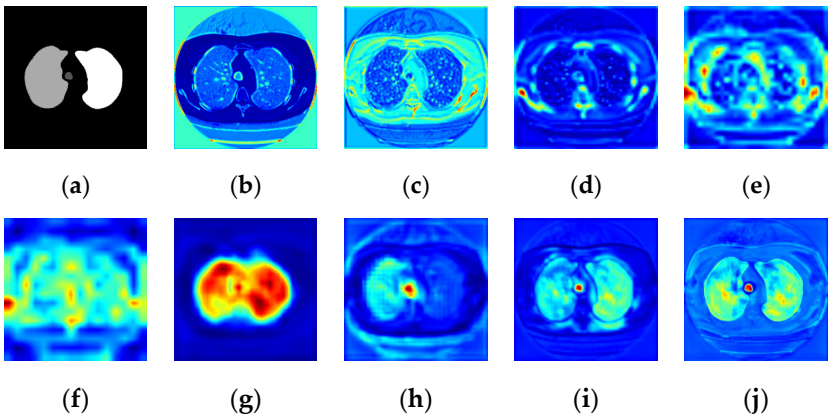

Figure 9. Heat map analysis of MCRAU-Net characteristics at each level. (**a**) Gold Standard. (**b**) Feature visualization results of Encoder layer 1. (**c**) Feature visualization results of Encoder layer 2. (**d**) Feature visualization results of Encoder layer 3. (**e**) Feature visualization results of Encoder layer 4. (**f**) Feature visualization results of Encoder layer 5. (**g**) Feature visualization results of Decoder layer 4. (**h**) Feature visualization results of Decoder layer 3. (**i**) Feature visualization results of Decoder layer 2. (**j**) Feature visualization results of Decoder layer 1.

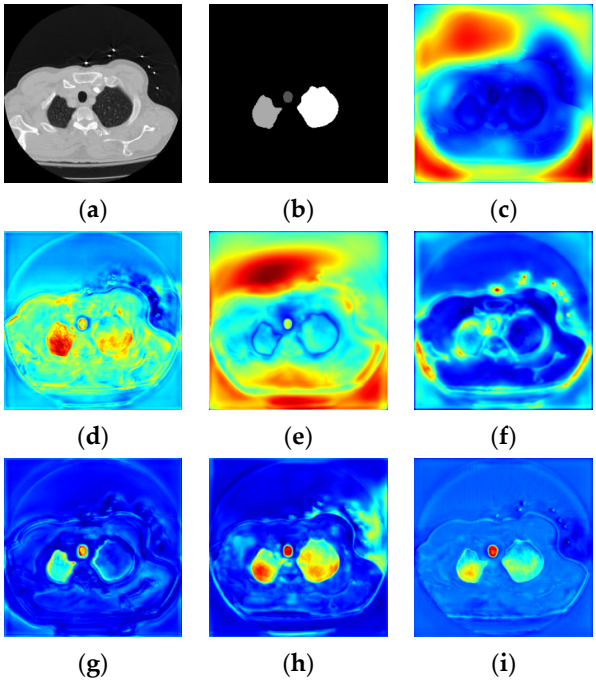

Figure 10. Heatmap analysis of the last layer of features for each algorithm. (**a**) Original CT images. (**b**) Gold Standard. (**c**) Feature visualization results of Vgg16-UNet. (**d**) Feature visualization results of U-Net. (**e**) Feature visualization results of Resnet50-UNet. (**f**) Feature visualization results of UneXt. (**g**) Feature visualization results of TransUNet. (**h**) Feature visualization results of UCTransNet. (**i**) Feature visualization results of MCRAU-Net.

5. Conclusions

Given the challenges posed by the complex lung structure, uneven gray scale, partial volume effect of CT images, and limited robustness of U-Net in obtaining context information, this paper proposes a new method called MCRAU-Net for lung parenchyma segmentation in CT images; one of the results is shown in Figure 11. MCRAU-Net introduces a multiple concatenation module in the encoding phase to enhance the network's ability to acquire location information. It also incorporates multi-scale residual learning modules to improve the feature extraction capability and an improved attention module to fuse high-level and low-level features. To optimize the training process, we use a hybrid loss function consisting of CELoss and DiceLoss, with CELoss addressing the classification task and DiceLoss emphasizing foreground details. The results demonstrate that our proposed model can effectively avoid redundant and irrelevant information transmission.

Accurate lung parenchyma segmentation is a critical component in developing an auxiliary diagnosis system for lung diseases. It also impacts further diagnosis and treatment of lung diseases. By analyzing lung segmentation results, clinicians can obtain essential information, including the volume, shape, and position of a patient's lungs, laying a solid foundation for further treatment and surgical planning. The existing deep learning-based lung segmentation algorithms, such as those in references [44,45], are lung segmentation algorithms designed for X-ray images. They treat the left lung and right lung as a single entity for segmentation, without a finer-grained distinction of the lung regions. Reference [46] focused on the segmentation of lung CT images. They utilized the Densenet network architecture; however, they also did not perform a finer-grained segmentation of the lung airways. In reference [47], the U-Net model was employed to differentiate between the left lung, right lung, and heart. The target regions in this case were quite large, and no modifications to the U-Net architecture were necessary. In the future, we plan to integrate and deploy our segmentation model on relevant medical image delineation platforms to improve the accuracy and efficiency of computer-aided diagnosis and treatment.

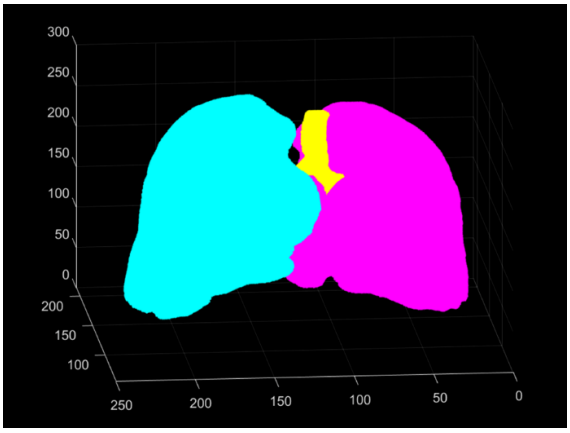

Figure 11. The results of the CT image segmentation and reconstruction. In the picture, the blue mask is the left lung, the purple mask is the right lung, and the yellow mark is the trachea.

Accurate segmentation of lung parenchyma and the trachea holds significant clinical importance. Firstly, the segmentation of lung parenchyma and the trachea can provide more precise medical imaging information, allowing for a more accurate assessment of bronchial narrowing and lesion extent, thereby aiding in treatment planning and surgical procedures. Secondly, in pulmonary surgeries, the precise segmentation of lung parenchyma and trachea helps surgeons better preserve lung tissue, reducing surgical trauma and complications. Additionally, scientific research on the interactions involving connectivity, filling, stabilization, and nutrition between lung parenchyma and trachea enhances our understanding of lung structure and function, offering insights into the prevention and treatment of lung diseases.

In terms of future research directions, several aspects can be considered: (1) improve algorithm performance by developing more efficient and accurate segmentation algorithms for different types and qualities of medical images to enhance the precision and speed of lung parenchyma and trachea segmentation; (2) consider physiological motion characteristics by thoroughly considering the physiological motion characteristics of the lungs, such as respiratory motion and cardiac pulsation, in the algorithm design to mitigate the impact of motion artifacts on segmentation results; (3) expand the application areas by exploring applications beyond segmentation, including quantitative analysis and functional assessments of lung parenchyma and trachea, providing valuable clinical information; and (4) establish public datasets by fostering academic progress and algorithm optimization to create public datasets for lung parenchyma and trachea segmentation.

Author Contributions: L.W. and J.L. designed the research study, L.W., C.Z. and Y.Z. wrote the manuscript together. All authors contributed to editorial changes in the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The datasets generated and analyzed during the current study are available in the [LUNA16] repository at https://luna16.grand-challenge.org/home/. The access date is 6 August 2023.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Gupta, D.; Anand, R.S. A hybrid edge-based segmentation approach for ultrasound medical images. *Biomed. Signal Process. Control* **2017**, *31*, 116–126. [CrossRef]
2. Mazouzi, S.; Guessoum, Z. A fast and fully distributed method for region-based image segmentation. *J. Real-Time Image Process.* **2021**, *18*, 793–806. [CrossRef]
3. Badshah, N.; Atta, H.; Shah, S.I.A.; Attaullah, S.; Ullah, M. New local region based model for the segmentation of medical images. *IEEE Access* **2020**, *8*, 175035–175053. [CrossRef]
4. Xiang, D.; Bagci, U.; Jin, C.; Shi, F.; Zhu, W.; Yao, J.; Sonka, M.; Chen, X. CorteXpert: A model-based method for automatic renal cortex segmentation. *Med. Image Anal.* **2017**, *42*, 257–273. [CrossRef]
5. Liang, Y.; Fu, J. Watershed algorithm for medical image segmentation based on morphology and total variation model. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 1954019. [CrossRef]
6. Lee, S.; Lee, A.; Hong, M. Cardiac CT Image Segmentation for Deep Learning–Based Coronary Calcium Detection Using K-Means Clustering and Grabcut Algorithm. *Comput. Syst. Sci. Eng.* **2023**, *46*, 2543–2554. [CrossRef]
7. Chen, J.; Zheng, H.; Lin, X.; Yang, M. A novel image segmentation method based on fast density clustering algorithm. *Eng. Appl. Artif. Intell.* **2018**, *73*, 92–110. [CrossRef]
8. Chen, Y.; Wang, Y.; Hu, F.; Wang, D. A lung dense deep convolution neural network for robust lung parenchyma segmentation. *IEEE Access* **2020**, *8*, 93527–93547. [CrossRef]
9. Park, B.; Park, H.; Lee, S.M.; Seo, J.B.; Kim, N. Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks. *J. Digit. Imaging* **2019**, *32*, 1019–1026. [CrossRef]
10. An, F.P.; Liu, X.W. Medical image segmentation algorithm based on feedback mechanism convolutional neural network. *Biomed. Signal Process. Control* **2019**, *53*, 101589.
11. Li, J.; Chen, H.; Li, Y.; Peng, Y. A novel network based on densely connected fully convolutional networks for segmentation of lung tumors on multi-modal MR images. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing, Dublin, Ireland, 16–18 October 2019; pp. 1–5.
12. Lu, Y.; Lu, G.M.; Li, J.X.; Zhang, D. Fully shared convolutional neural networks. *Neural Comput. Appl.* **2021**, *33*, 8635–8648. [CrossRef]
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
14. Shaziya, H.; Shyamala, K.; Zaheer, R. Automatic lung segmentation on thoracic CT scans using U-net convolutional network. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 643–647.
15. Li, X.M.; Chen, H.; Qi, X.J.; 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] [PubMed]
16. Damseh, R.; Cheriet, F.; Lesage, F. Fully convolutional DenseNets for segmentation of microvessels in two-photon microscopy. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 June 2018; pp. 661–665.
17. Qamar, S.; Jin, H.; Zheng, R.; Ahamd, P.; Usama, M. A variant form of 3D-UNet for infant brain segmentation. *Future Gener. Comput. Syst.* **2020**, *108*, 613–623. [CrossRef]
18. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1857–1866.
19. 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] [PubMed]
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
21. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. 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.
24. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
25. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. A2-nets: Double attention networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, 3–8 December 2018.
26. Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving convolutional networks with self-calibrated convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10096–10105.

27. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3024–3033.
28. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
29. Tay, C.P.; Roy, S.; Yap, K.H. Aanet: Attribute attention network for person re-identifications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7134–7143.
30. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3139–3148.
31. Fang, X.; Yan, P. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Trans. Med. Imaging* **2020**, *39*, 3619–3629. [CrossRef]
32. Fu, H.; Cheng, J.; Xu, Y.; Wong, D.W.K.; Liu, J.; Cao, X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans. Med. Imaging* **2018**, *37*, 1597–1605. [CrossRef]
33. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
34. Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; Yang, X. Cascaded context pyramid for full-resolution 3D semantic scene completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7801–7810.
35. Huang, G.; Liu, S.; Van, L.; Weinberger, K.Q. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2752–2761.
36. 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.
37. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2441–2449. [CrossRef]
38. Liu, Y.; Wang, Y.; Wang, S.; Liang, T.; Zhao, Q.; Tang, Z.; Ling, H. Cbnet: A novel composite backbone network architecture for object detection. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11653–11660. [CrossRef]
39. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
40. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022*; Part V; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 23–33.
41. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–1 October 2021*; Part I 24; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 36–46.
42. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
43. Setio, A.A.A.; Traverso, A.; De Bel, T.; Berens, M.S.; Van Den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M.E.; Geurts, B.; et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **2017**, *42*, 1–13. [CrossRef] [PubMed]
44. Cao, F.; Zhao, H. Automatic Lung Segmentation Algorithm on Chest X-ray Images Based on Fusion Variational Auto-Encoder and Three-Terminal Attention Mechanism. *Symmetry* **2021**, *13*, 814. [CrossRef]
45. Kim, M.; Lee, B.-D. Automatic Lung Segmentation on Chest X-rays Using Self-Attention Deep Neural Network. *Sensors* **2021**, *21*, 369. [CrossRef] [PubMed]
46. Jalali, Y.; Fateh, M.; Rezvani, M.; Abolghasemi, V.; Anisi, M.H. ResBCDU-Net: A Deep Learning Framework for Lung CT Image Segmentation. *Sensors* **2021**, *21*, 268. [CrossRef] [PubMed]
47. Mehta, A.; Lehman, M.; Ramachandran, P. Autosegmentation of lung computed tomography datasets using deep learning U-Net architecture. *J. Cancer Res. Ther.* **2023**, *19*, 289–298. [PubMed]