



Article

TransCotANet: A Lung Field Image Segmentation Network with Multidimensional Global Feature Dynamic Aggregation

Xuebin Xu ^{1,2,3}, Muyu Wang ^{1,2,3} , Dehua Liu ^{1,2,3}, Meng Lei ^{1,2,3}, Jun Fu ^{1,2,3}  and Yang Jia ^{1,2,3,*}

- ¹ School of Computer Science and Technology, Xi'an University of Posts & Telecommunications, Xi'an 710121, China; xuxuebin@xupt.edu.cn (X.X.); wmy961017@stu.xupt.edu.cn (M.W.); liudehua@stu.xupt.edu.cn (D.L.); leimeng@stu.xupt.edu.cn (M.L.); fujun_learner@stu.xupt.edu.cn (J.F.)
- ² Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts & Telecommunications, Xi'an 710121, China
- ³ Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, China
- * Correspondence: jia yang@xupt.edu.cn

Abstract: Chest X-ray (CXR) images can be used to diagnose a variety of lung diseases, such as tuberculosis, pneumonia, and lung cancer. However, the variation in lung morphology due to differences in age, gender, and the severity of pathology makes high-precision lung segmentation a challenging task. Traditional segmentation networks, such as U-Net, have become the standard architecture and have achieved remarkable results in lung field image segmentation tasks. However, because traditional convolutional operations can only explicitly capture local semantic information, it is difficult to obtain global semantic information, resulting in difficult performance in terms of accuracy requirements in medical practical applications. In recent years, the introduction of Transformer technology to natural language processing has achieved great success in the field of computer vision. In this paper, a new network architecture called TransCotANet is proposed. The network architecture is based on the U-Net architecture with convolutional neural networks (CNNs) as the backbone and extracts global semantic information through symmetric cross-layer connections in the encoder structure, where the encoder stage includes an upsampling module to improve the resolution of the feature map, and uses the dynamic aggregation module CotA to dynamically aggregate multi-scale feature maps and finally obtain more accurate segmentation results. The experimental results show that the method outperformed other methods for lung field image segmentation datasets.

Keywords: lung field image segmentation; transformer; dynamic aggregation module; TransCotANet



Citation: Xu, X.; Wang, M.; Liu, D.; Lei, M.; Fu, J.; Jia, Y. TransCotANet: A Lung Field Image Segmentation Network with Multidimensional Global Feature Dynamic Aggregation. *Symmetry* **2023**, *15*, 1480. <https://doi.org/10.3390/sym15081480>

Academic Editors: João M. F. Rodrigues and Jan Egger

Received: 5 May 2023
Revised: 29 May 2023
Accepted: 6 June 2023
Published: 26 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

CXR is widely used as a major and important medical imaging technique for the diagnosis of various lung diseases in clinical treatment and preliminary screening. Medical image segmentation, on the other hand, helps doctors to perform accurate diagnoses by annotating complex medical image (such as X-ray films, magnetic resonance imaging (MRI) and computed tomography (CT) scans, etc.) data to discover and extract the desired object region of interest [1,2]. Lung field segmentation is very important for the diagnosis of lung diseases, such as pneumonia [3] and tuberculosis [4,5]. Traditional CXR medical image annotation methods rely on manual annotation by clinicians, and although CXRs annotated by clinicians may be highly accurate, this method is usually labor-intensive and costly. Therefore, we need a high-precision, deep-learning-based medical image segmentation method to replace the traditional manual annotation method to make the analysis of medical images more automated, fast, and accurate. With the continuous development in the field of computer vision, convolutional neural networks (CNNs) play a crucial role in the field of medical image segmentation. Continued research advances in neural network techniques for deep learning have provided new ideas for medical image segmentation, such as the

full convolutional neural network (FCN), which directly applies end-to-end convolutional networks to image segmentation tasks and has become a landmark, pioneering work in the field of semantic segmentation [6]. In recent studies, the U-Net [7] model has become a more suitable choice. This model uses a set of symmetric encoder–decoder structures connected across layers and further optimizes the processing flow by multi-scale fusion to better preserve detailed information. This approach has been widely used in the field of medical image segmentation, such as providing great help for the accurate segmentation of computed tomography (CT) images, which can effectively locate and identify different organ regions and generate clear medical meaning [7–9].

In research in the field of computer vision, convolutional neural networks (CNNs) have exhibited excellent performance and potential for a wide range of applications. However, due to the local connectivity properties of convolutional operations, each neuron is only connected to neurons in the adjacent region of the previous layer, which leads to limitations in the model's ability to process global information and an explicit limitation in long-term semantic interactions, while it is difficult to learn global semantic information. As a result, this structure often exhibits translational invariance, i.e., when the object orientation or location changes, the corresponding neurons may not be activated correctly, thus affecting the recognition results. To address this limitation, a CNN-based self-attentive mechanism for feature usage has been proposed [10]. Owing to the remarkable success of Transformer in the field of natural language processing (NLP), it is able to globally model the correlation between input sequences and better capture long-range dependencies by calculating the relative importance between all positions within the input sequence. This technique is implemented through a self-attentive mechanism and has been successfully applied to many NLP tasks, such as machine translation, text classification, and speech recognition [11]. Compared with previous CNN approaches, Transformer has shown powerful capabilities in global relevance modeling and excellent transferability in pre-training tasks with large-scale data [10,12]. In many semantic segmentation tasks, Swin-Transformer has shown powerful feature extraction capabilities and superior performance and generalization capabilities by using cross-layer connectivity and cross-attention mechanisms for information transfer and feature fusion, with great success in various segmentation tasks [13]. Motivated by Swin-Transformer, the proposal of Swin-Unet [14], with Swin-Transformer as its backbone network and U-Net as the network structure, is able to reduce the computation and number of parameters with guaranteed segmentation accuracy compared with the traditional U-Net network. UCTransNet [15] proposes a more efficient method for fusing features, currently employing deep learning in, for example, the federal neural network CCT and an adaptive complementary modal analysis (CCA) module to fuse the multi-scale feature information of medical images. However, owing to the shrinking path design of the U-Net structure, some features of small targets gradually lose visibility or are lost at deeper layers as the number of layers increases, which is still challenging for the semantic segmentation of small targets in medical images, and the improvement direction still needs to be explored to further enhance its performance.

In this paper, we present a study on dynamic feature aggregation. We find that the contextual feature extraction of UCTransNet does not work well on small medical datasets. On the other hand, owing to its inadequate feature extraction, some pixel-level feature information is lost, which is especially obvious in low-resolution medical images, and the segmentation accuracy is further affected due to the loss of subtle boundary features and insufficient feature extraction of contextual information.

To this end, we propose a deep learning model called TransCotANet, which contains a CotA module for the dynamic aggregation of input features and the filtering of noise and unnecessary information. By using contextual information to achieve dynamic feature extraction and fusion, it retains the advantages of Transformer and U-Net jump structure dynamic feature fusion, but also enhances visual performance and removes unnecessary information and noise, which is more conducive to the high-precision segmentation of medical images. Numerous experiments have shown that our proposed method has

advantages over various other research results in three datasets from the Japanese Society of Radiological Technology (JSRT) [16], Montgomery County (MC) [17], and Shenzhen [18].

2. Related Work

As the field of computer vision continues to evolve, numerous remarkable studies have been dedicated to combining deep learning with medical image segmentation. In this paper, we will discuss the progress related to this research approach around three aspects.

2.1. Convolutional Neural Lung Field Segmentation Network

Traditional lung image segmentation tasks often result in poor segmentation accuracy due to the blurred edges of lung regions and differences in gender and age between individuals, while manual annotation is labor-intensive. In recent years, with the development of neural networks, a series of excellent deep learning methods has emerged for the automatic segmentation of lung images, significantly improving the accuracy and efficiency of medical image segmentation. For example, Ngo et al. [19] proposed a new fusion of a distance regularization level set method and a convolutional neural network model that could effectively handle the ambiguous regions appearing at the lung edges and had good segmentation performance for dense lung tissue and lung texture. Sheng Change et al. [20] added to the traditional convolutional encoder–decoder structure Jump connection and feedback mechanism to improve the adaptability to factors such as dense texture, ambient lighting, and different postures, employing data enhancement and Dropout techniques and introducing specific learning rate adjustment strategies to improve the training effect. Souza et al. [21] designed a deep neural network-based automatic lung segmentation and reconstruction method using a two-stage structure that combined a convolutional neural network (CNN) with a conditional generative adversarial network (GAN) for the coarse segmentation of lung regions and reducing a large amount of detailed information, respectively. Saïdy et al. [22] proposed a deep learning and morphological knowledge-based lung segmentation method to solve the fragmentation problem in lung images during training, minimizing the loss function to optimize the model. Fan et al. [23] proposed a COVID-19 lung CT infection segmentation network called Inf-Net, which used both reverse the attention mechanism and explicit edge attention mechanism.

2.2. Transformer Combined with CNN Networks

Researchers have attempted to integrate Transformer model encoders and self-attention mechanisms into CNNs through pixel global interaction modeling based on feature maps. Such a scheme aims to enhance the expressiveness and accuracy of different medical image segmentation tasks and improve the performance and practical value of deep learning methods. For example, Chen et al. [24] designed a Transformer encoder and U-Net decoder structure and introduced a global self-attention mechanism. Hu et al. [14] proposed to use Swin-Transformer as the backbone, built in the form of a U-Net network structure, using a self-attention mechanism, introducing multi-scale feature fusion and local feature fusion mechanisms. Valanarasu et al. [25] proposed a gating-based axial attention mechanism to overcome the problem of using a small number of samples. Unlike these approaches, our proposed TransCotANet network adds a CotA (dynamic feature aggregation) module to the encoder and the dynamic aggregation of contextual features is performed through the CotA module.

2.3. U-Net

The U-Net model constructs a symmetric encoder and decoder architecture in which the presence of jump connections enables the pixel-level segmentation prediction of images, aiming to achieve the accurate segmentation of biomedical images through convolutional neural network techniques. Various research works have used different modules and attention mechanisms to improve and optimize U-Net, thus increasing the accuracy of deep learning methods in medical image segmentation. For example, Zhou et al. [15]

proposed a novel medical image segmentation architecture based on nested and dense jump connections, in which the jump connections were designed to reduce the gap between the feature maps of the encoder and decoder subnetworks and fuse the high-resolution features in the encoder and the corresponding semantics in the decoder, thus gradually enriching the feature representation. Such an approach aimed to improve the information acquisition and object recognition of complex biomedical images by deep learning models. Oktay et al. [26] proposed the integration of the attention mechanism gate module in the jump connection of U-Net. Gao et al. [27] proposed the application of the self-attention module to both the encoder and decoder in order to capture image with minimal overhead. Xu et al. [28] proposed an additional extension path based on the U-Net model and built a corresponding supervised signal, which could obtain better results for medical image segmentation by the double supervision technique. This method uses multi-level feature fusion and full convolution operations in a deep learning framework, and joint training through an encoder–decoder structure to improve the performance and optimization of the model for biomedical image segmentation tasks.

3. Materials and Methods

In this section, we first elaborate on the overall network structure and then introduce the proposed CotA module. The following is a brief overview of our proposed lung field segmentation method, which first divides the image into a number of patches as coded inputs and then decodes them back to the original spatial resolution to map the features to the pixel-level output image after obtaining the corresponding tensor information following several convolution and pooling operations. UCTransNet uses the CCT and CCA modules for adaptive multi-scale feature information fusion, which can dynamically assign the extracted multi-scale feature information, but it only uses the original U-Net network structure to extract features, which will ignore some key information for small medical image datasets. Different from this approach, our method proposes a CotA module to dynamically aggregate image feature information, which expands the sensing field and allows better access to key information, thus enriching the multi-scale semantic feature information. We have been conducting this research since August 2022 at the BDAI Big Data Lab, Xi'an University of Posts and Telecommunications.

3.1. TransCotANet

TransCotANet Overview. First, input a medical image $X \in \mathbb{R}^{H \times W \times C}$ with height H , width W , and number of channels C . $H \times W$ denotes the actual spatial resolution. The objective of the medical image segmentation task is to be able to predict the semantic mask map at the corresponding pixel level with size $H \times W$. Our proposed TransCotANet framework is shown in Figure 1. (Part (A) shows the structure of the CotA module and part (B) shows the overall structure of TransCotANet). We aimed to utilize the CotA module in order to expand the perceptual field, perceive the input contextual information, improve the feature extraction of contextual information, enhance the aggregation of multidimensional global features, and further improve the quality of semantic segmentation. TransCotANet has a CotA module to greatly capture the adjacent information features and improve the characterization of encoder semantics. In addition, the CotA module used in TransCotANet aims to aggregate static contextual information and dynamic contextual information using multi-level complementarity to obtain dynamically perceived different semantic information and aggregate the contextual information of adjacent keys for multi-scale feature prediction fusion. In the TransCotANet structure, we added the CotA module after the E_1 , E_2 , E_3 , and E_5 jump connection layers in the encoder to expand the perceptual field and pass the information features after each stage of aggregation to the next layer to obtain richer and more accurate contextual information features, which are then passed to the four outputs, \hat{O}_1 , \hat{O}_2 , \hat{O}_3 , and \hat{O}_4 , by CCT [15] and then decode the reduced feature map by upsampling and decoder volume layer feature D_1 , D_2 , D_3 , and D_4 connection.

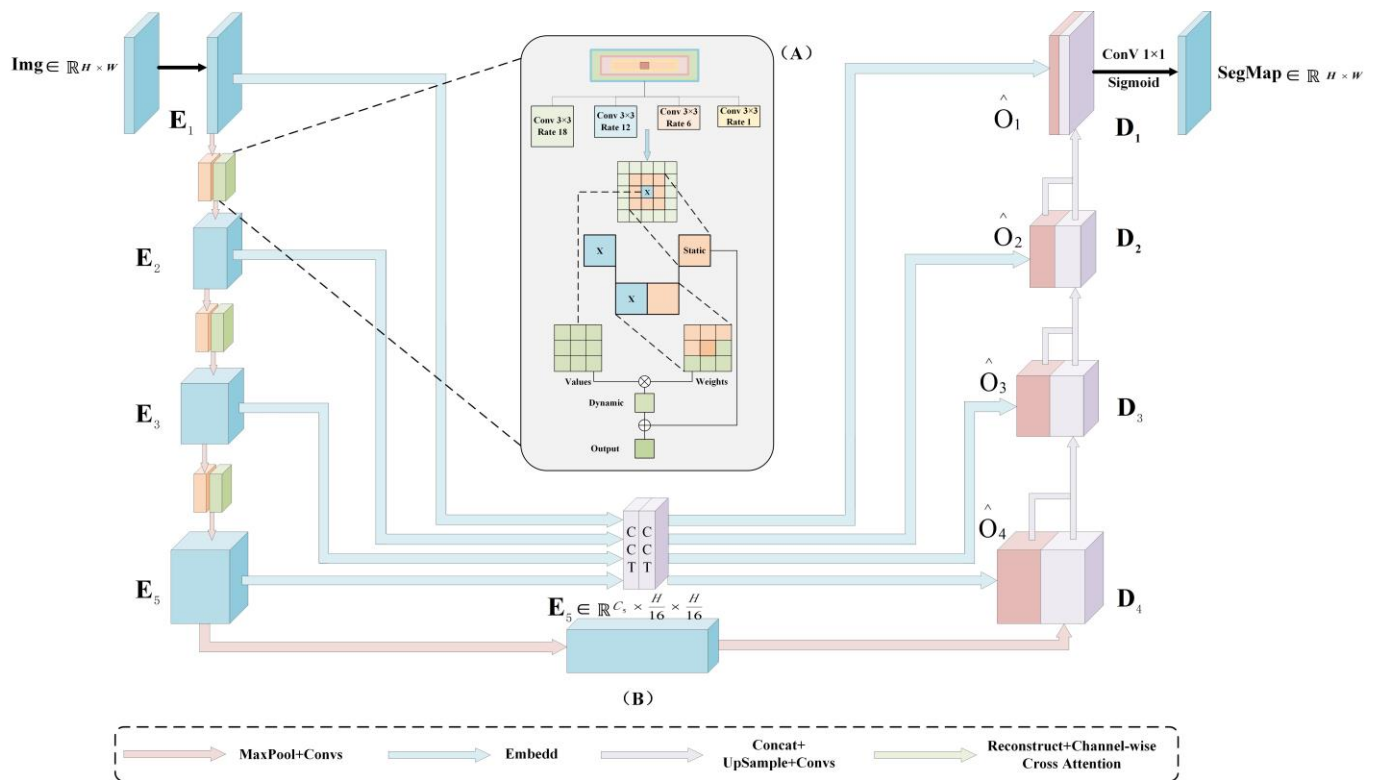


Figure 1. Demonstration of the overall framework. (A) Structure diagram of the CotA module; (B) structure diagram of the proposed TransCotANet network.

3.2. CotA Module

CotA module overall structure. The structure of the CotA module is shown in Figure 2.

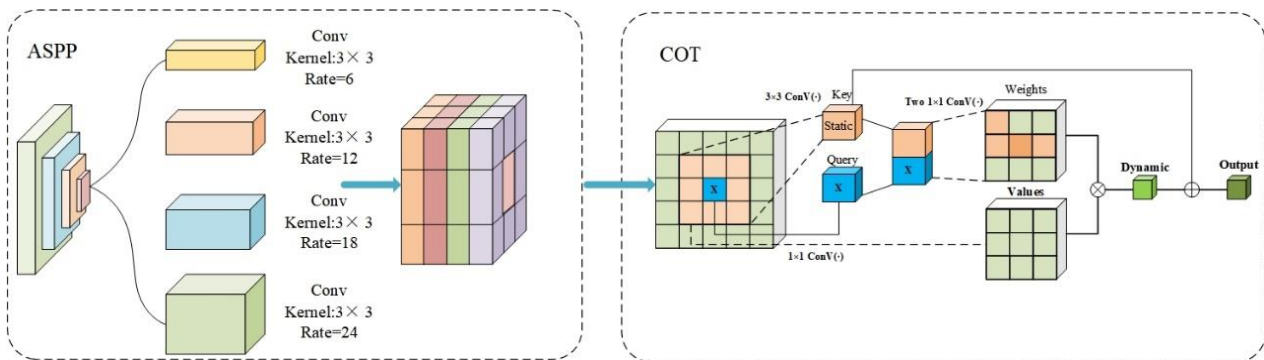


Figure 2. CotA module structure.

In the CotA module, two parts are included; the first part is a void space convolutional pooling pyramid (ASPP) [29] and the second part is COTAttention [15]. We found that the U-Net network structure, in which the encoder consisted of multiple convolutional and pooling layers, was used to extract features from the input image, but the convolutional layers used in U-Net had a fixed perceptual field size, which meant that important contextual information may have been lost when processing larger images. In addition, U-Net extracts features mainly by local convolution and pooling operations, which may lead to some global features being ignored, thus affecting the accuracy of segmentation. We use the improved CotA module of ASPP for the feature extraction of semantic segmented images during feature extraction. We convolved the $H \times W \times C$ feature map after one convolution, with height H , width W , and number of channels C . We achieved different perceptual fields by convolving cavities with four different sampling rates to obtain a feature map with

larger perceptual fields. The sampling rate of the convolution module is set to 1, 6, 12, and 18 in four different parallel convolution layers, and the convolution kernel is always set to 3×3 to keep it constant. Finally, the obtained feature maps are stitched together to capture the multi-scale information, and the number of channels of the feature maps is reduced by a convolution layer fed into the global pooling layer to obtain a feature map of size $H \times W \times C$ and dimension 1×1 . The 2D feature map 1×1 of $X \in \mathbb{R}^{H \times W \times C}$ is taken as the input and embedded into the W_q , W_k , and W_v matrix with queries, keys, and values defined as $Q = X$, $K = X$ and $V = XW_v$ and respectively. The first convolution of $K \times K$ groups for all neighboring secret keys of $K \times K$ makes full use of the contextual information between different K and mines the contextual information on the spatial scale. The upper and lower keys $M^1 \in \mathbb{R}^{H \times W \times C}$ reflect the static contextual information after the convolution of the adjacent keys 3×3 , and the contextual attention matrix A is obtained from two consecutive 1×1 convolution layers (W_ε and W_φ where W_ε contains the activation function ReLU and W_φ does not) after cascading the information of query Q with K^1 :

$$A = [M^1, Q]W_\varepsilon W_\varphi \quad (1)$$

In each attention head, based on query Q and contextual information features, static M^1 is mined to guide enhanced self-attention learning and then all values (V) are integrated to compute the attentional feature map M^2 based on the contextual attention matrix A obtained above:

$$M^2 = V \otimes A \quad (2)$$

where \otimes is represented as a multiplication operation of pairs of confusion matrices. We fuse the static contextual information M^1 and dynamic contextual information M^2 as the output. The weight values generated by COT are multiplied with the original individual features and used as the input features for the next layer of convolution.

CotA module implementation process. The input medical image $X \in \mathbb{R}^{H \times W \times C}$, with height H , width W , and number of channels C . After one convolution, four parallel convolution kernels are passed into the 3×3 convolution layer, and the sampling rate is set to 1, 6, 12, and 18 to obtain four different features, G_1 , G_2 , G_3 , and G_4 . We fuse the static contextual information M^1 and dynamic contextual information M^2 in COTAttention to obtain the weights K and features F , K , which are multiplied with the original input features to obtain the corresponding weighted feature values E :

$$E_n = [KG_1; KG_2; KG_3; KG_4] \quad (3)$$

P is separately stitched with the adaptive mean global average pooled F to obtain the weighted global feature T :

$$T = E_n + F \quad (4)$$

The feature maps of each channel are compressed to 1×1 so as to extract the features of each channel and then obtain the global features; finally, the feature maps are sampled from 1×1 back to the original size as the output. The overall global feature dynamic aggregation is:

$$O = X \in \mathbb{R}^{H \times W \times C} (F + A[KG_1 + KG_2 + KG_3 + KG_4]) \quad (5)$$

The input of the medical segmentation image is $X \in \mathbb{R}^{H \times W \times C}$, the output is O , and the global mean pooling is A .

4. Experimental Results

In this section, we first introduce the datasets of three lung field regions, JSRT, MC, and Shenzhen. Secondly, we present the implementation details of the experimental setup, based on which we performed some experiments to validate the effectiveness of our

proposed method, and compare it with various other methods. Finally, we design multiple sets of ablation experiments to analyze our proposed method and performance.

4.1. Experimental Dataset

JSRT dataset [16]. One of the datasets we used was created in 1998 by the Japanese Society of Radiological Technology in collaboration with the Japanese Radiological Society (JRS). It contains 247 PACXR images of the human chest region. Each CXR line slice is a single-channel grayscale image with a color depth of 8 and has a resolution of 2048×2048 ; 154 have pulmonary nodules and 93 are normal. In order to ensure the universality of the dataset and prevent training overfitting, we used two simple three-fold random rotations and three-fold flipped data enhancement strategies, and inserted the original image data into the expanded dataset together with the training to obtain the original seven-fold dataset number and 1729 images. We used 1400 images as training cases, 140 images as test cases, and 140 images as validation cases. In the JSRT dataset, we used the segmented chest slice SCR dataset [30], because the lung segmentation mask in it corresponded to the JSRT dataset.

Montgomery dataset [17]. Another dataset was created by the Montgomery County, Maryland, Department of Health and Human Services collection in the United States. It contains 138 radiograph images of the human chest region. Each CXR line slice is a single-channel grayscale image with a color bit depth of 8 and a resolution of 4892×4020 or 4020×4892 . In this dataset, 80 images are from healthy cases and 58 are from pulmonary nodules cases. We used one of the lung segmentation masks as a basic fact, labeled under the supervision of a professional radiologist and provided by Candemir et al. [29]. Similarly, we used two simple five-fold machine rotations and five-fold flip data enhancement strategies, and inserted the original image data into the expanded dataset together with the training; the dataset was expanded to eleven times the original, with 1518 images. We used 1210 images as training cases, 110 images as validation cases, and 198 images as test cases.

Shenzhen dataset [18]. This dataset is a collaboration between the Third People's Hospital of Shenzhen, Guangdong, China, and the National Library of Medicine of Maryland, USA. It contains 662 radiographic images of the human chest region; each CXR radiograph is a single-channel grayscale image with a color depth of 8 and an average resolution of 326 radiographs healthy subjects and 336 from patients with tuberculosis. We used this dataset supervised and labeled by professional radiologists, which was provided by Jaeger et al. [18]. For the Shenzhen dataset, we used simple three-fold machine rotations and a simple three-fold flip data enhancement strategy, and trained the original image data together in the expanded dataset, which was expanded to seven times the original size, containing 4634 images. We use 3710 images as training cases, 336 images as validation cases, and 336 images as test cases. Table 1 shows the CXR and its specification summary of the X-ray film images of the three data sets after data enhancement, and the original and enhanced example images of the three data sets are shown in Figures 3–5.

Table 1. Summary of the JSRT, MC, and Shenzhen datasets.

Dataset Name	Number of Image Pairs			Number of Image Pairs
	Train	Val	Test	
JSRT [16]	1400	140	140	Image format: PNG Image size: 2048×2048
MC [17]	1210	110	198	Image format: PNG Image size: 4892×4020
Shenzhen [18]	3710	336	336	Image format: PNG Image size: 3000×3000

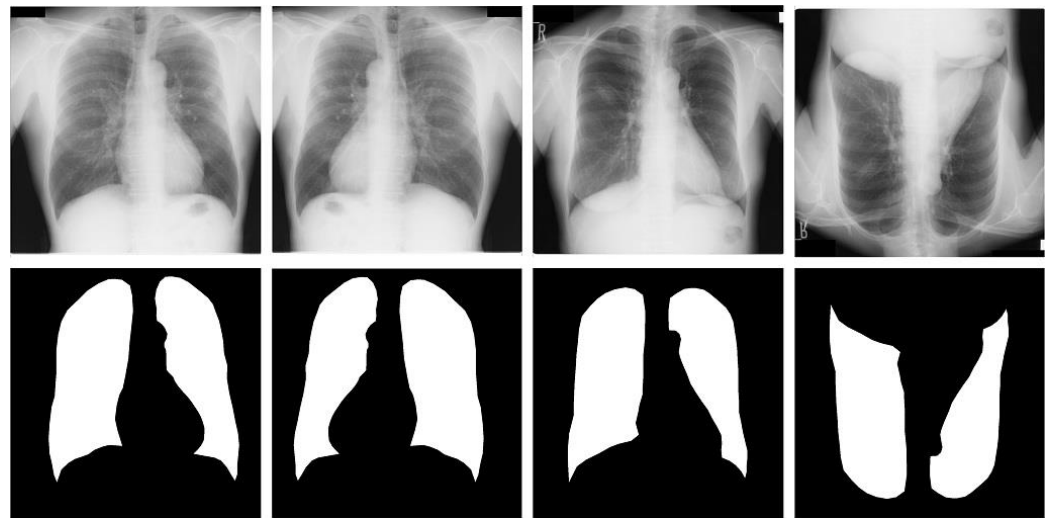


Figure 3. Images and masks of the JSRT dataset.

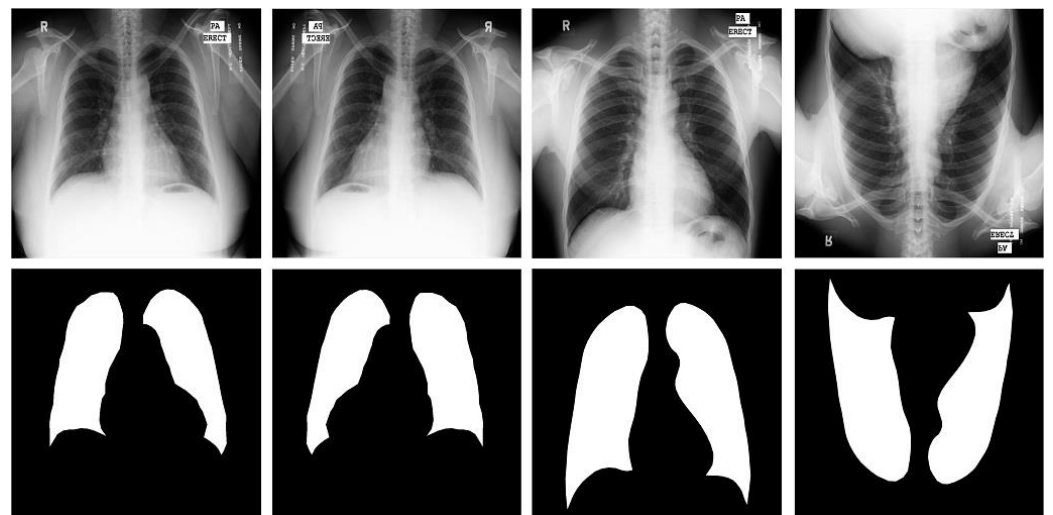


Figure 4. Images and masks of the MC dataset.

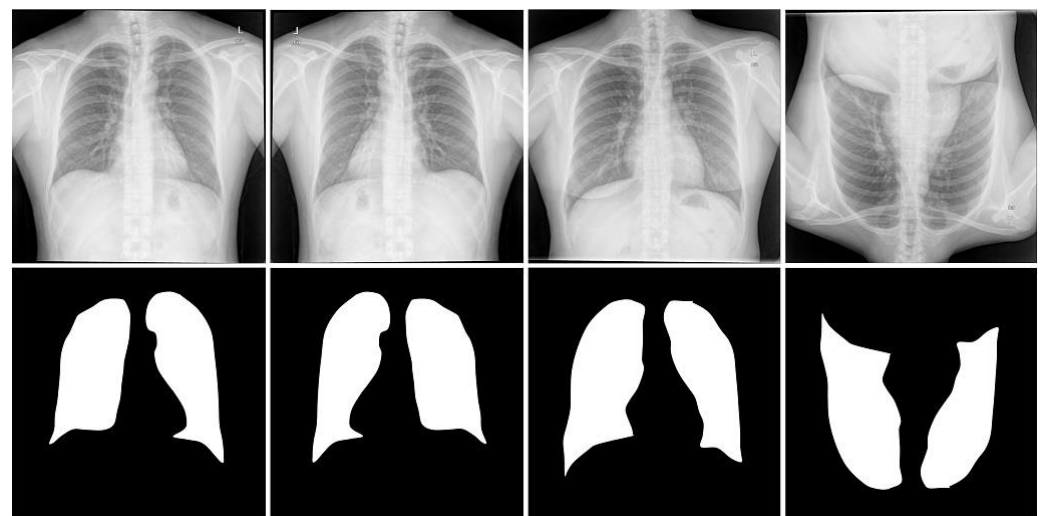


Figure 5. Images and masks of the Shenzhen dataset.

4.2. Details of Implementation

We trained and tested the JSRT dataset, the Montgomery dataset, and the Shenzhen dataset with both the input resolution and patches at 224×224 and 16, respectively. For the learning rate, we trained and tested our model using the Adam optimizer, setting the initial resolution to 0.001. For the use of loss functions, we chose cross-entropy loss and sieve loss as our loss functions for training our network. Three five-fold cross-validations were performed and the mean and standard deviation were obtained, highlighting that we were more convincing in these three small datasets of medical images. For each dataset, we performed 300 epochs to train our model. According to the results of the study, our method outperformed other comparable methods. We chose Dice [31] (DSC) as the main evaluation metric to assess the similarity between the factual and predicted masks. In addition, we added four evaluation metrics, the Jaccard index (JC), Accuracy (ACC), Specificity (TNR), and Sensitivity (TPR), to calculate and quantitatively evaluate the segmentation performance at the pixel level.

$$\begin{aligned}
 \text{Dice} &= \frac{2 \times TP}{2 \times TP + FP + FN} \\
 \text{Jaccard} &= \frac{TP}{TP + FP + FN} \\
 \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
 \text{Specificity} &= \frac{TN}{FP + TN} \\
 \text{Sensitivity} &= \frac{TP}{TP + FN}
 \end{aligned} \tag{6}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative, respectively. All experiments were run using the NVIDIA Tesla V100 SXM2 16G server.

4.3. Comparison with Other Methods

In this subsection, we comprehensively evaluate the performance of our proposed method on three datasets, JSRT, MC, and Shenzhen, and compare it with other state-of-the-art methods.

4.3.1. Experimental Evaluation of JSRT Dataset

To demonstrate the segmentation performance of our proposed TransCotANet, we compared it with other state-of-the-art techniques. We evaluated TransCotANet with three different types of methods. These include three U-Net-based methods: U-Net [7], Dense-UNet [32], and U-Net++ [33]; one other network-based method: LF-Net [34]; and four state-of-the-art Transformer-based segmentation methods: Attention U-Net [35], Swin-UNet [14], TransAttUNet [36], and UCTransNet [15]. To demonstrate the effectiveness of our proposed CotA module, we conducted experiments using the initial code and settings they released. We used cross-entropy loss and sieve loss as our loss functions. The results of our experiments on the JSRT dataset are reported in Table 2, where the best results are shown in bold.

Table 2. Comparison with the state-of-the-art segmentation methods on the JSRT dataset. The black, bold font in the table represents the best results.

Method	Year	DSC%	JC%	ACC%	TNR%	TPR%
U-Net [7]	2015	96.17	97.71	98.21	-	94.94
U-Net++ [33]	2018	97.84	95.80	98.93	-	99.28
Attention U-Net [35]	2019	97.59	95.31	98.81	-	98.82
Dense-Unet [32]	2020	97.60	95.30	-	98.80	97.90
LF-Net [34]	2021	97.20	98.16	99.52	99.79	98.95
Swin-Unet [14]	2021	97.67	95.48	98.71	-	95.42
TransAttUnet [36]	2021	98.88	97.82	99.41	-	98.88
UCTransNet [15]	2022	98.32	97.63	99.37	99.78	98.20
TransCotANet	-	99.03	98.76	99.14	99.79	98.32

From the experimental results, first, we can see that our proposed TransCotANet achieved the best evaluation results for the DSC, JC, and TNR coefficients. This improvement demonstrates the overall effectiveness of our CotA module design in medical lung field image segmentation accuracy. Second, our TransCotANet segmentation outperformed other techniques compared with various previous baselines, improving the mean DSC and JC to 99.03% and 98.76%, respectively, especially in U-Net (96.17%), with a 2.96% improvement in the DSC score. We compared the proposed TransCotANet with the latest Transformer-based techniques TransAttUnet and Swin-Unet, both of which achieved better evaluation results, reflecting our ability to improve the quality of the detailed segmentation of medical lung field images. The DSC score was improved by 0.71% compared with the baseline based on the UCTransNet network, demonstrating that our proposed CotA module effectively aggregated multi-scale semantic information and enriched the ability of contextual semantic feature information in the network. Finally, our proposed TransCotANet achieved the highest DSC and JC scores in lung field segmentation, which demonstrated that our designed CotA module dynamically perceived the aggregation of multi-scale contextual information and showed a strong ability to learn high-level and low-level semantic feature details.

4.3.2. Experimental Evaluation of MC Dataset

Our proposed TransCotANet was applied to the MC dataset and evaluated in combination with other recent techniques. We evaluated TransCotANet with two different types of methods. These included three U-Net-based methods: U-Net [7], improved U-Net [37], and Dense-Unet [31]; and four other network-based methods: Modification in FCN [38], SEDUCM [32], Atrous Convolutions [39], AlexNet, and ResNet2 [22]; and one state-of-the-art Transformer-based segmentation method: UCTransNet [15]. The results of our experiments on the MC dataset are reported in Table 3, where the best results are shown in bold.

Table 3. Comparison with the state-of-the-art segmentation methods on the MC dataset. The black, bold font in the table represents the best results.

Method	Year	DSC%	JC%	ACC%	TNR%	TPR%
U-Net [7]	2015	96.17	97.71	98.21	-	94.94
SEDUCM [32]	2017	95.60	93.50	-	-	-
Atrous Convolutions [39]	2017	96.40	94.10	-	-	-
AlexNet and ResNet [22]	2018	94.00	88.00	96.90	96.70	97.50
Modification in FCN [38]	2018	91.74	97.84	-	-	-
Dense-Unet [31]	2020	97.90	95.90	-	99.20	98.10
improved U-Net [37]	2022	97.70	95.50	98.90	99.30	97.50
UCTransNet [15]	2022	96.78	93.70	97.68	98.86	97.16
TransCotANet	-	98.02	97.89	98.91	99.32	97.36

The experimental results in Table 3 show that our proposed TransCotANet outperformed various state-of-the-art techniques on the MC datasets in terms of comprehensive evaluation, which reflected the strong generalization of our model. First, the DSC score of our proposed TransCotANet was improved by 1.92% compared with the U-Net baseline, which reflected that our method still achieved a high performance improvement for the MC datasets. Second, our model showed a significant improvement in all scores compared with various other baselines, consistently outperforming state-of-the-art techniques.

Finally, the addition of our proposed CotA module to TransCotANet compared with the UCTransNet baseline resulted in a significant increase in the DSC and JC scores, by 1.28% and 4.47%, respectively, and the CotA module achieved a powerful dynamic aggregation of contextual semantic features, reaching 98.02%. The highest DSC and JC scores of 98.02% and 97.89% were achieved for the MC dataset, which fully demonstrated the effectiveness of our method.

4.3.3. Experimental Evaluation of Shenzhen Dataset

Similarly, we performed a comprehensive study on the Shenzhen dataset and evaluated it in combination with other state-of-the-art technologies. We evaluated TransCotANet with two different types of methods. Three U-Net-based methods were included: U-Net [7], MPDC DDLA U-Net [40], and MultiResUNet [41]; along with four other network-based methods: Deeplabv3 [29], AG-net [42], CNN+Neural Net [43], and LF-Net [34]; and one state-of-the-art Transformer-based segmentation method: UCTransNet [15]. The results of our experiments on the Shenzhen dataset are reported in Table 4, where the best results are shown in bold.

Table 4. Comparison with the state-of-the-art segmentation methods on the Shenzhen dataset. The black, bold font in the table represents the best results.

Method	Year	DSC%	JC%	ACC%	TNR%	TPR%
U-Net [7]	2015	95.80	92.20	-	-	-
Deeplabv3 [29]	2017	95.80	92.20	-	-	-
AG-net [42]	2019	96.10	92.50	-	-	-
CNN+Neural Net [43]	2019	87.00	-	93.00	-	-
MultiResUNet [41]	2019	96.00	92.40	-	-	-
LF-Net [34]	2021	90.55	95.86	-	98.55	97.67
MPDC DDLA U-Net [40]	2021	96.70	92.90	98.31	-	-
UCTransNet [15]	2022	96.78	92.91	98.02	98.93	96.30
TransCotANet	-	97.66	94.41	98.46	99.35	97.78

The experimental results in Table 4 show that our proposed TransCotANet outperformed various state-of-the-art techniques on the Shenzhen dataset in terms of comprehensive evaluation, which reflected the superiority of our model. First, the DSC score of our proposed TransCotANet was improved by 1.9% compared with the U-Net baseline, which reflected that our approach achieved high performance improvement on the Shenzhen dataset. Second, our model showed significant improvement in all scores compared with various other baselines. Finally, the addition of our proposed CotA module to the TransCotANet compared with the UCTransNet baseline resulted in DSC and JC scores that improved to 97.66% and 94.41%, 0.8% and 1.6% over those of the original network, respectively, and the CotA module revealed a strong dynamic aggregation of contextual semantic features in the Shenzhen dataset. The highest DSC score of 97.66% was achieved for image segmentation, which fully demonstrated that our proposed model, which still maintained superior performance on multiple datasets, exhibited strong generalization learning capability.

To consider the generalizability and robustness of the model, we employed two testing methods: 1. Training on the Montgomery dataset (M) and testing on the Shenzhen dataset (S). 2. Training on the Shenzhen dataset (S) and testing on the Montgomery dataset (M) [44].

We evaluated the performance of the model by the DSC coefficients and JC coefficients in a comprehensive manner. The comprehensive evaluation of the cross-testing experiments on the two datasets is shown in Table 5.

Table 5. Results of the combined evaluation of the Montgomery dataset (M) and the Shenzhen dataset (S) cross-test. The black, bold font in the table represents the best results.

Test DSC/JC	U-Net	Res-U-Net	BCDU-Net	Incep-Res-U-Net	R2U-Net	Att-R2U-Net	DEFU-Net	UCTransNet	TransCotANet
Train: S	0.771	0.816	0.767	0.898	0.866	0.866	0.915	0.916	0.923
Test: M	0.774	0.819	0.908	0.902	0.871	0.871	0.916	0.894	0.919
Train: M	0.856	0.664	0.909	0.890	0.912	0.907	0.923	0.920	0.932
Test: S	0.857	0.665	0.909	0.892	0.914	0.910	0.923	0.903	0.927

From Table 5, we can see that the proposed TransCotANet achieved the best results in both the DSC and JC coefficient combined evaluation, which reflected the robustness and universality of the proposed TransCotANet for medical image segmentation tasks.

4.4. Ablation Studies

In this subsection, first, we perform a multiset ablation experimental study on three datasets, JSRT, MC, and Shenzhen, and then we perform a comprehensive evaluation analysis on the two different datasets.

4.4.1. Comprehensive Evaluation of the JSRT Dataset

To comprehensively evaluate the CotA modules of our proposed TransCotANet model, as well as their effectiveness, we conducted various ablation studies on the JSRT datasets. We inserted the proposed CotA modules into our network with different strategies to comprehensively evaluate the effectiveness of our proposed CotA modules by changing the positions of the three modules in our TransCotANet architectural encoder.

Effect of insertion position. We designed four ablation experiments by placing CotA modules after E_1 , E_2 , and E_3 separately and all together to verify the validity of our provided method (E_1 , E_2 , and E_3 represent the corresponding modules in the TransCotANet framework). Figure 6 summarizes the average DSC and JC segmentation performance of the four ablation experiments.

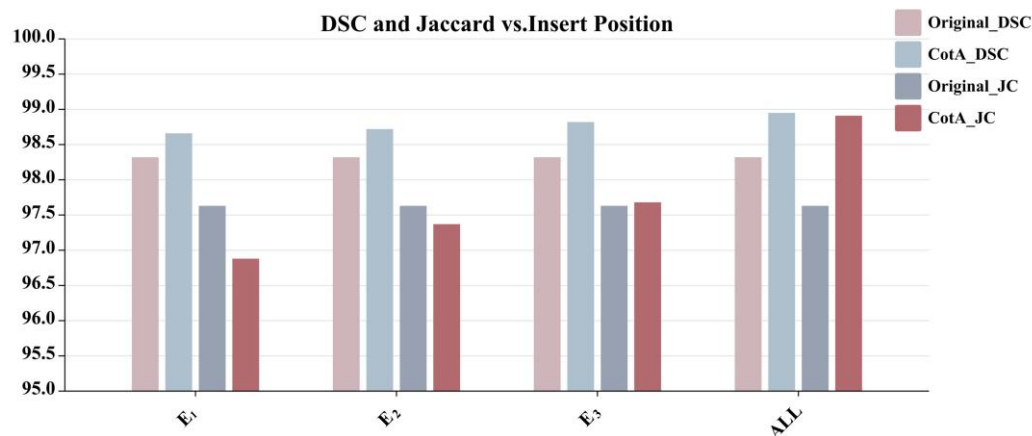


Figure 6. An ablation study on the four insertion positions of the CotA module in TransCotANet on the JSRT dataset.

The purpose of this ablation was to test the effect of adding CotA modules at different positions to our proposed TransCotANet. When we inserted the CotA module only at E_1 and E_2 , we found that the DSC coefficients were consistently better than the original UCTransNet, but the JC coefficients were slightly lower than the initial ones. We considered that it may have been due to the CotA module in the shallow layer of the encoder, which increased the complexity of the model; it may have over-fit the training data, resulting

in predictions on the test data that did not match the actual results, thus making the JC coefficient lower. After inserting E_3 , we found that both Dice coefficients and JC coefficients were higher than the accuracy in the initial network. After we inserted the CotA module together into E_1 , E_2 , and E_3 , we found that the DSC coefficient and JC evaluation scores were improved to 99.03% and 98.76%, respectively. The above experimental results initially demonstrated the superiority of the designed CotA module. Obviously, we added the CotA module to the encoder to effectively expand the perceptual field, dynamically perceive the multi-scale feature information fusion, and further improve the comprehensive effect of semantic segmentation.

4.4.2. Comprehensive Evaluation Analysis of the MC Dataset

To take into account the generalization and robustness of the model, we also used the same ablation experimental strategy for the MC datasets for the comprehensive evaluation of the model as a whole. That is, the effectiveness of our proposed CotA module was evaluated comprehensively by changing the positions of the three modules in our TransCotANet architecture encoder. Figure 7 summarizes the average DSC and JC segmentation performance of the four ablation experiments on the MC dataset.

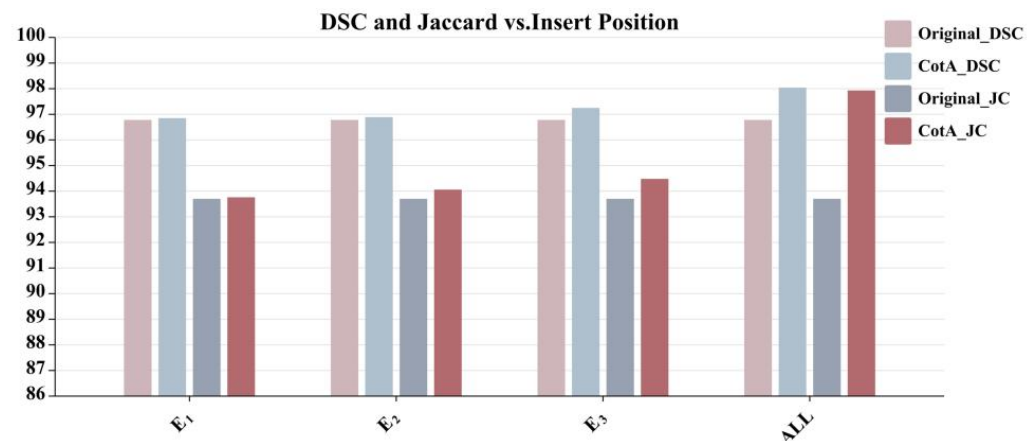


Figure 7. An ablation study on four insertion positions of the CotA module in TransCotANet on the MC dataset.

As can be seen in Figure 7, after only inserting the CotA module into E_1 , we found that the DSC coefficient slightly increased, but the JC coefficient slightly decreased, probably due to the shallow encoder increasing the model depth, which over-fit the training data, resulting in prediction results on the test data that did not match the actual results. After inserting the CotA modules into E_1 , E_2 , and E_3 in our proposed TransCotANet, the DSC coefficient scores slightly improved and the JC coefficient slightly improved. After inserting all CotA modules into E_1 , E_2 , and E_3 , we found that the DSC and JC coefficients rose significantly to 98.02% and 97.89%, respectively. The above experiments demonstrate the powerful contextual semantic-feature-learning ability of our proposed TransCotANet model for different datasets and fully demonstrate the effectiveness of our designed CotA module in medical lung field image segmentation tasks.

4.4.3. Comprehensive Evaluation Analysis of the Shenzhen Dataset

Similarly, we used the same ablation experimental strategy for the Shenzhen dataset as that for the previous two datasets to perform the overall comprehensive evaluation of the model. That is, the effectiveness of our proposed CotA module was evaluated comprehensively by changing the positions of the three modules in our TransCotANet architecture encoder. Figure 8 summarizes the average DSC and JC segmentation performances in the four ablation experiments on the Shenzhen dataset.

As can be seen in Figure 8, after inserting the CotA modules into our proposed TransCotANet into E_1 , E_2 , and E_3 , the DSC coefficient score slightly increased and the JC coefficient increased slightly. After inserting all CotA modules into E_1 , E_2 , and E_3 , we found that the DSC and JC coefficients rose significantly to 97.66% and 94.41%, respectively. The above experiments demonstrate the powerful contextual semantic-feature-learning ability of our proposed TransCotANet model for different datasets and fully demonstrate the effectiveness of our designed CotA module in medical lung field image segmentation tasks.

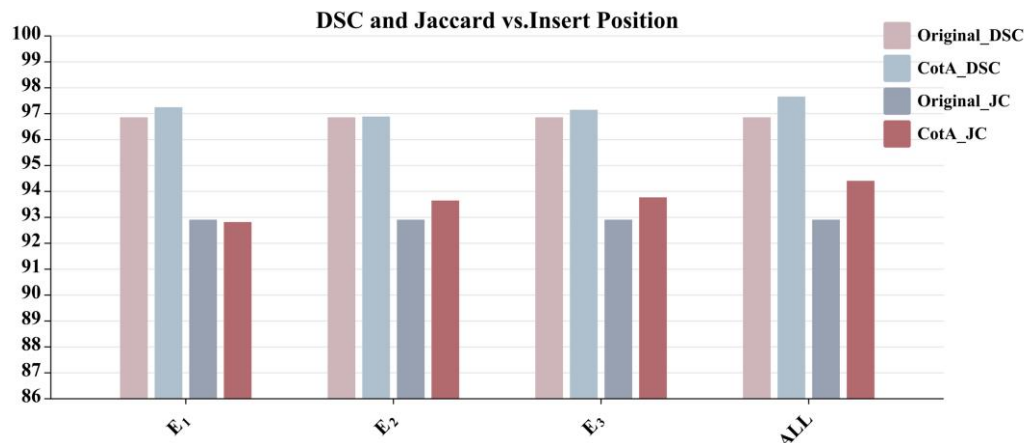


Figure 8. An ablation study on four insertion positions of the CotA module in TransCotANet on the Shenzhen dataset.

4.5. Model Visualization and Discussion

Visualization. We present our findings by comparing the qualitative results on the JSRT dataset, MC dataset, and Shenzhen dataset, as shown in Figure 9. In the figure, the squares represent the comparison of the difference areas of segmentation results for different data sets.

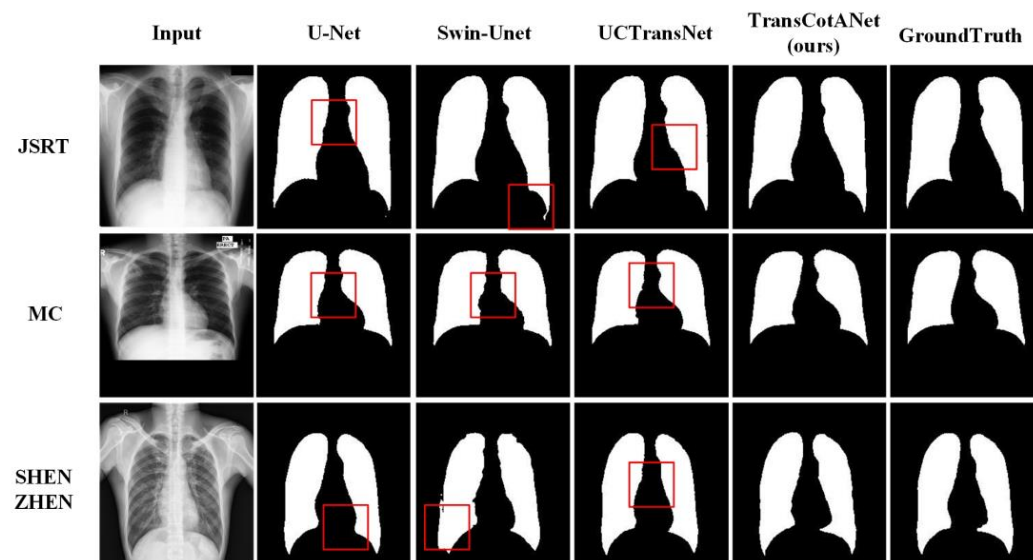


Figure 9. Quantitative results of lung field segmentation on the JSRT, MC, and Shenzhen datasets.

We used the CNN-based method U-Net, the Transformer-based Swin-Unet, and UCTransNet, and to verify their ability to recognize semantic features, we visualized the feature maps for each decoder stage. Looking at the segmentation results, we can find that these methods may have suffered from over-segmentation or lossy segmentation. For the first experiment, there was under-segmentation in the U-Net network: from the above figure, we can see that, in the upper region of the lung field of the JSRT image and

the middle region of the lung field of the MC image, the segmented edges are missing compared with the true values, and the lower region of the lung field of the Shenzhen image also shows lossy segmentation, which was due to the fact that the downsampling process weakened the detail information and the upsampling process made it difficult to fully recover the original features and spatial information. Even though the U-Net network employs cross-layer connections to preserve certain low-level features, this mechanism still cannot solve the lossy segmentation problem well because of the limitations imposed by the characteristics of the network itself and the multi-level information transfer. Comparatively, in the second experiment, the Swin-Unet network showed over-segmentation: in the above figure, we can see that over-segmentation occurs in the lower-right region of the JSRT lung field image, the middle region of the MC lung field image, and the lower-left region of the Shenzhen lung field image, which was due to the poor generalization ability of the model, and the associated region was segmented into multiple unrelated parts. This led to information loss and redundancy, which decreased the performance and efficiency of the model on lung medical image processing and increased the false positive rate and uncertainty of the likelihood analysis. In addition, in the third experiment, the UCTransNet network made an error in lung prediction and did not detect its central location correctly. We can see that there was a slight over-segmentation in the right region of the JSRT lung field image, the central region of the MC lung field image, and the central region of the Shenzhen lung field image, which may have been due to the inability of the model to obtain enough semantic features to fully convey the information. As shown in the figure, in the last experiment, our proposed TransCotANet, on the other hand, was able to fully and dynamically perceive the contextual multi-scale semantic feature information and was able to correctly segment the left and right lung images with more reliable results and clearer boundaries. Therefore, it can be concluded that our method could not only effectively learn semantic information for medical image segmentation, but also improve the segmentation performance and ensure the accuracy of the segmentation results.

Discussion. Although our proposed method can dynamically aggregate contextual semantic information to obtain better accuracy in medical image segmentation, it still has two limitations. We found that the method we designed was not effective in segmenting image data with strong interference and a small lung field region. As shown in Figure 10, these were the visualized images with different noise disturbances obtained from each of the three datasets, JSRT, MC, and Shenzhen. We can see that, in the lung field region, the area where the lung edges were connected to the organs and bones often contained white noise, which caused great interference to our segmentation and led to a decrease in the segmentation accuracy of individual data, thus affecting the overall segmentation accuracy. In the first column of the dataset of images from the JSRT, we can clearly see that in the lower-right region of the lung field that there was a large area of white noise, and this phenomenon may have been due to the unclear chest region image caused by different acquisition devices or occlusion; the white noise had very small differences with the bones, which easily led to the under-segmentation phenomenon of the model and thus affected the overall performance in the comprehensive assessment. In the second column of the MC dataset images, we can see that the white noise was concentrated in the spine region, which may have been caused by the distance of the patient from the sampling device; due to this interference, it made it difficult for the model to define the boundaries of the lung field region, which was more prone to the under-segmentation phenomenon. In the third column of the Shenzhen dataset images, we found strong white noise overall, and the organs in the middle of the chest region (heart, etc.) also produced strong interference, which caused the overall segmentation performance to degrade. In contrast, the segmentation accuracy was maintained for the rightmost image with lower interference. In addition, our proposed method, due to the high network complexity, was slow for high-resolution processing, and the semantic segmentation network performed pixel-level classification of the input image, which was slow and may not have been suitable for application scenarios with high real-time requirements. In the future, we will design a network that can analyze accurate

edge information while obtaining more semantic information, and prune and optimize the network backbone to maintain high accuracy while making the network more lightweight. In the figure, the red circles represent the areas where the data set received interference.

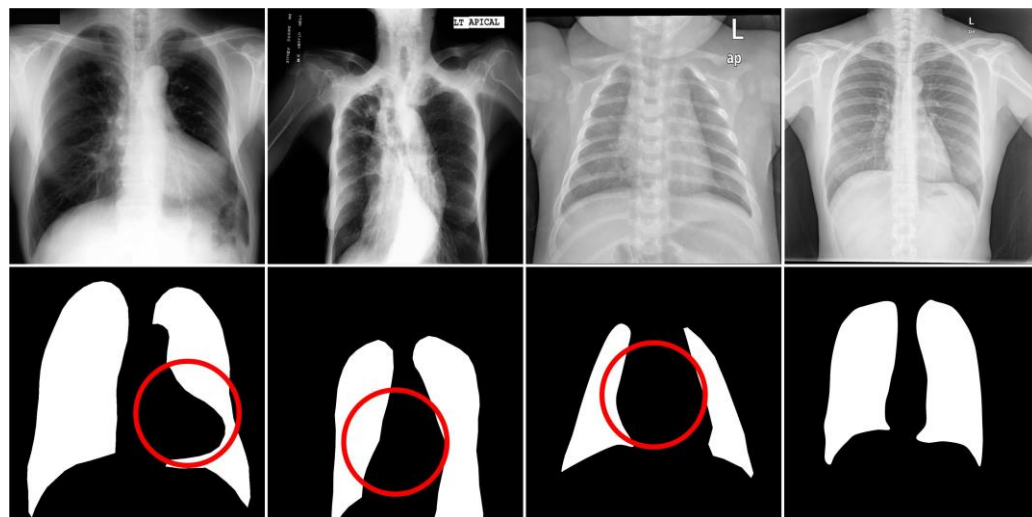


Figure 10. Segmentation noise interference visualization images for JSRT, MC, and Shenzhen datasets.

5. Conclusions

In this paper, we propose a new module combining contextual attention ASPP, called the CotA module, which can dynamically perceive multi-scale feature information and efficiently capture adjacent contextual information to improve the quality of medical image segmentation. In detail, the dynamic perception module can make full use of the adjacent contextual information through the neighboring interactions between semantic features in the encoder and the multi-scale image features. Meanwhile, the CotA module can effectively expand the perceptual field to greatly capture adjacent information features and improve the semantic representation of the encoder. Compared with previous state-of-the-art techniques, the proposed TransCotANet greatly benefits from the dependency of adjacent features and the capability for the dynamic aggregation of multi-scale information of the CotA module, which ensures consistent representation with semantic features. We effectively improved the information capture capability in the encoder, enriched the semantic feature information, and alleviated the problem of insufficient granularity caused by insufficient feature extraction in the traditional U-Net network architecture. Extensive experimental results demonstrate that our proposed TransCotANet achieved consistent performance improvement in medical image segmentation. Our method achieved DSC and JC coefficients of 99.03% and 98.76%, 98.02% and 97.89%, and 97.66% and 94.41% on the JSRT dataset, MC dataset, and Shenzhen dataset, respectively, which were better than those of other advanced methods.

Author Contributions: Conceptualization, X.X.; Investigation, X.X. and M.W.; Methodology, X.X. and M.W.; Writing—original draft, X.X. and M.W.; Validation, D.L.; Formal analysis, M.L.; Visualization, J.F.; Data curation, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China, No. 61673316, and the Department of Education Shaanxi Province, China, under Grant 16JK1697.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The JSRT, MC, and Shenzhen datasets are openly available at <http://imgcom.jsrt.or.jp/minijsrtdb/> (accessed on 20 May 2023), <https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html#tuberculosis-image-data-sets> (accessed on 20 May 2023), and <https://data.lhncbc.nlm.nih.gov/>

nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html (accessed on 20 May 2023), respectively.

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

References

1. Roy, K.; Banik, D.; Bhattacharjee, D.; Krejcar, O.; Kollmann, C. LwMLA-NET: A lightweight multi-level attention-based network for segmentation of COVID-19 lungs abnormalities from CT images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5007813. [[CrossRef](#)]
2. Yang, X.; Wei, Q.; Zhang, C.; Zhou, K.; Kong, L.; Jiang, W. Colon polyp detection and segmentation based on improved MRCNN. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 4501710. [[CrossRef](#)]
3. Santosh, K.; Antani, S. Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities? *IEEE Trans. Med. Imaging* **2017**, *37*, 1168–1177. [[CrossRef](#)] [[PubMed](#)]
4. Vajda, S.; Karargyris, A.; Jaeger, S.; Santosh, K.; Candemir, S.; Xue, Z.; Antani, S.; Thoma, G. Feature selection for automatic tuberculosis screening in frontal chest radiographs. *J. Med. Syst.* **2018**, *42*, 146. [[CrossRef](#)]
5. Karargyris, A.; Siegelman, J.; Tzortzis, D.; Jaeger, S.; Candemir, S.; Xue, Z.; Santosh, K.; Vajda, S.; Antani, S.; Folio, L. Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays. *Int. J. Comput. Assist. Radiol. Surg.* **2016**, *11*, 99–106. [[CrossRef](#)]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Part III 18, pp. 234–241.
8. 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](#)]
9. Yu, Q.; Xie, L.; Wang, Y.; Zhou, Y.; Fishman, E.K.; Yuille, A.L. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8280–8289.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2661.
11. Patel, K.; Bur, A.M.; Li, F.; Wang, G. Aggregating global features into local vision transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 1141–1147.
12. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
14. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2023; pp. 205–218.
15. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *1*, 3552. [[CrossRef](#)]
16. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.-I.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.* **2000**, *174*, 71–74. [[CrossRef](#)]
17. Candemir, S.; Jaeger, S.; Palaniappan, K.; Musco, J.P.; Singh, R.K.; Xue, Z.; Karargyris, A.; Antani, S.; Thoma, G.; McDonald, C.J. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* **2013**, *33*, 577–590. [[CrossRef](#)]
18. Jaeger, S.; Karargyris, A.; Candemir, S.; Folio, L.; Siegelman, J.; Callaghan, F.; Xue, Z.; Palaniappan, K.; Singh, R.K.; Antani, S. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* **2013**, *33*, 233–245. [[CrossRef](#)] [[PubMed](#)]
19. Ngo, T.A.; Carneiro, G. Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 25–27 September 2015; pp. 2140–2143.
20. Chang, C.-S.; Lin, J.-F.; Lee, M.-C.; Palm, C. Semantic lung segmentation using convolutional neural networks. In Proceedings of the Bildverarbeitung für die Medizin 2020: Algorithmen–Systeme–Anwendungen Workshops, Berlin, Germany, 15–17 March 2020; pp. 75–80.
21. Souza, J.C.; Diniz, J.O.B.; Ferreira, J.L.; da Silva, G.L.F.; Silva, A.C.; de Paiva, A.C. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Comput. Methods Programs Biomed.* **2019**, *177*, 285–296. [[CrossRef](#)] [[PubMed](#)]

22. Saidy, L.; Lee, C.-C. Chest X-ray image segmentation using encoder-decoder convolutional network. In Proceedings of the 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 19–21 May 2018; pp. 1–2.
23. Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2626–2637. [[CrossRef](#)]
24. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
25. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 2–5 September 2021; pp. 36–46.
26. 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.
27. Gao, Y.; Zhou, M.; Metaxas, D.N. Utnet: A hybrid transformer architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 8–12 October 2021; pp. 61–71.
28. Xu, Z.; Liu, S.; Yuan, D.; Wang, L.; Chen, J.; Lukasiewicz, T.; Fu, Z.; Zhang, R. ω -net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution. *Neurocomputing* **2022**, *500*, 177–190. [[CrossRef](#)]
29. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
30. Van Ginneken, B.; Stegmann, M.B.; Loog, M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. *Med. Image Anal.* **2006**, *10*, 19–40. [[CrossRef](#)]
31. Yahyatabar, M.; Jouvret, P.; Cheriet, F. Dense-Unet: A light model for lung fields segmentation in Chest X-Ray images. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 1242–1245.
32. Yang, W.; Liu, Y.; Lin, L.; Yun, Z.; Lu, Z.; Feng, Q.; Chen, W. Lung field segmentation in chest radiographs from boundary maps by a structured edge detector. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 842–851. [[CrossRef](#)]
33. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Granada, Spain, 8–12 October 2018; pp. 3–11.
34. Singh, A.; Lall, B.; Panigrahi, B.K.; Agrawal, A.; Agrawal, A.; Thangakunam, B.; Christopher, D. Deep LF-Net: Semantic lung segmentation from Indian chest radiographs including severely unhealthy images. *Biomed. Signal Process. Control* **2021**, *68*, 102666. [[CrossRef](#)]
35. Ma, H.-J.; Ledward, D. High pressure/thermal treatment effects on the texture of beef muscle. *Meat Sci.* **2004**, *68*, 347–355. [[CrossRef](#)] [[PubMed](#)]
36. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *arXiv* **2021**, arXiv:2107.05274.
37. Liu, W.; Luo, J.; Yang, Y.; Wang, W.; Deng, J.; Yu, L. Automatic lung segmentation in chest X-ray images using improved U-Net. *Sci. Rep.* **2022**, *12*, 8649. [[CrossRef](#)]
38. Hooda, R.; Mittal, A.; Sofat, S. An efficient variant of fully-convolutional network for segmenting lung fields from chest radiographs. *Wirel. Pers. Commun.* **2018**, *101*, 1559–1579. [[CrossRef](#)]
39. Hwang, S.; Park, S. Accurate lung segmentation via network-wise training of convolutional networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Québec City, QC, Canada, 8–12 October 2017; pp. 92–99.
40. Ma, L.; Hou, X.; Gong, Z. *Multi-Path Aggregation U-Net for Lung Segmentation in Chest Radiographs*; PREPRINT (Version 1) available at Research Square; Research Square: Durham, NC, USA, 2021.
41. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [[CrossRef](#)]
42. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [[CrossRef](#)]
43. Huynh, H.T.; Anh, V.N.N. A deep learning method for lung segmentation on large size chest X-ray image. In Proceedings of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Danang, Vietnam, 20–22 March 2019; pp. 1–5.
44. Zhang, L.; Liu, A.; Xiao, J.; Taylor, P. Dual encoder fusion u-net (defu-net) for cross-manufacturer chest X-ray segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9333–9339.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.