Automated U-Net-ConvMixer Attention method for Lung Segmentation

Farah OUBELKAS
Laboratoire de Recherche Informatique, Réseaux, Mobilité et Modélisation
Hassan First University of Settat
Faculté Sciences et Technique
Settat, Morocco
f.oubelkas@uhp.ac.ma

Lahcen MOUMOUN
Laboratoire de Recherche Informatique, Réseaux, Mobilité et Modélisation
Hassan First University of Settat
Faculté Sciences et Technique
Settat, Morocco
lahcen.moumoun@uhp.ac.ma

Abdellah JAMALI
Laboratoire de Recherche Informatique, Réseaux, Mobilité et Modélisation
Hassan First University of Settat
Faculté Sciences et Technique
Settat, Morocco
abdellah.jamali@uhp.ac.ma

***Abstract*— Accurate lung segmentation in chest X-rays is vital for diagnosing various pulmonary pathologies. While U-Net architectures and their derivatives have achieved success in medical applications, their local convolution operations inherently limit their ability to capture global contextual information. In this work, we present a novel ConvMixer-based model for lung segmentation. Inspired by the ConvMixer architecture, this model effectively extracts both local and global features from lung images. To improve our segmentation results, we proposed a post-processing step in order to eliminate weakly contributing features from the segmentation. We evaluate our model on two publicly available chest X-ray datasets, Shenzhen and Montgomery, demonstrating superior performance compared to state-of-the-art segmentation methods. Notably, our final model achieves an accuracy of 97.52% and an IoU of 92.71%. These results suggest the proposed ConvMixer-based model as a promising approach for lung segmentation with the potential to contribute to improved diagnosis of various lung diseases.**

Keywords—Lung segmentation, ConvMixer, U-Net, multi-scale attention, Encoder-Decoder

I. INTRODUCTION

Pinpointing the lungs in medical images, known as lung segmentation, plays a critical role in various healthcare practices. From detecting lung cancer and diagnosing diseases to planning treatments, accurate segmentation is essential. It forms the foundation for interpreting medical imaging correctly and ultimately aids in effective diagnosis and treatment of lung conditions.

Accurately separating the lungs in medical images, known as lung segmentation, presents several challenges: (1) variable image quality, that means that imaging modalities, patient conditions, and even imaging parameters can significantly impact lung scan quality, making consistent segmentation difficult. Technological systems known as computer-aided detection (CAD) that assist physicians in interpreting medical images but achieving uniform image quality remains a challenge [1]. (2) anatomical diversity; In other words, the inherent variability in lung size and shape, both between individuals and within the same person across scans, complicates the development of universal segmentation algorithms [2]. (3) disease influences; Pathologies like lung nodules or tumours [3] further complicate segmentation by introducing irregularities in the lung boundary, potentially leading to misinterpretations like false positive or negative results.

In this paper, we propose a new, fully convolutional architecture for lung segmentation, drawing inspiration from the ConvMixer model's approach. The proposed system efficiently combines two key elements: ConvMixer for capturing global image context and a multi-scale attention gate for selectively suppressing irrelevant features while amplifying crucial ones. This dual approach aims to enhance segmentation performance and mitigate the influence of image noise and artifacts.

This work marks a significant step forward in lung segmentation using the ConvMixer model. We overcome limitations of previous methods by: Firstly, introducing a novel ConvMixer-based approach; Our method surpasses the state-of-the-art, utilizing the U-Net with the ConvMixer model's efficiency and global context capture for accurate lung detection and segmentation. Secondly, we integrate a mask attention module to extract context across different image scales, significantly improving feature representation and segmentation accuracy. Thirdly, extensive evaluations on diverse publicly available datasets showcase the method's potential, achieving highly promising performance compared to recent research in handling complex scenarios such as the presence of conditions like tuberculosis and nodules, and variations in lung shape due to age and gender.

The structure of the paper is as follows: The subsequent sections delve deeper into related research, the proposed method's details (including the mask attention module), experimental setup, and a thorough analysis of the results. We conclude by summarizing these contributions and highlighting potential future directions.

II. RELATED WORK

Lung segmentation, the process of extracting lung regions from medical images, plays a crucial role in diagnosing and treating various pulmonary diseases. Deep learning has emerged as a powerful tool for this task, with numerous architectures offering diverse approaches. Lung segmentation research has explored the potential of combining multiple imaging modalities, like CT and MRI, to achieve improved accuracy [4]. While these studies demonstrate that leveraging diverse data sources can outperform single-modality approaches, they often require massive datasets and significant computational resources. This presents challenges in terms of data acquisition, storage, and processing power, limiting their widespread adoption.

To solve these limitations, deep learning has been widely used in medical research for image processing, including various types of CNN models such as 2D, 2.5D, and 3D [5] .Murugappan et al. proposed an efficient lung segmentation method for CT scans using deep learning [6], highlighting an alternative approach that tackles the limitations of multimodality methods. Their work demonstrates the effectiveness of fine-tuning pre-trained models, achieving both improved segmentation accuracy and faster convergence. This strategy offers a promising direction by utilizing existing deep learning models and tailoring them to specific tasks, reducing the need for vast amounts of data and computational power.

There are some approaches based on the encoder-decoder architecture that was used for lung segmentation. U-Net [7] is a versatile encoder-decoder structure for semantic segmentation. The heart of U-Net's effectiveness lies in its ingenious skip connections. These connections serve as information highways, seamlessly integrating high-level contextual understanding with fine-grained image details. This synergy of scales empowers U-Net to achieve unmatched segmentation accuracy, enabling more precise diagnoses and improved healthcare outcomes.

Its variants like Attention U-Net [8] leverage attention mechanisms to capture long-range dependencies, achieving high accuracy in lung segmentation tasks. However, their reliance on numerous convolutional layers can lead to computational inefficiency. DeepRes U-Net [9] is also another variant that incorporates residual connections within the U-Net architecture, addressing the vanishing gradient problem and improving performance, especially for deep networks. Dense Attention U-Net [10] utilizes dense attention blocks within the U-Net decoder, further enhancing feature representation and boundary localization, leading to superior segmentation results.

Recent years have seen a surge in Transformer-based networks applied to medical image segmentation [11]. Chen et al. [12] introduced TransUNet that extracts local and global information from the input image. While TransUNet achieved impressive results on diverse medical image datasets, exceeding the performance of other segmentation models, it needs substantial computational resources and heavily relies on large training data. SegFormer [13], a transformer-based architecture that employs hierarchical feature fusion and local attention, achieving state-of-the-art performance in various segmentation tasks was proposed. While powerful, its computational cost due to transformer layers might be a limitation, especially for resource-constrained environments. In 2023, ViT-Seg [14] was proposed. It explores pure transformer architectures for medical image segmentation, showcasing their potential in capturing long-range dependencies.

One promising approach is the use of ConvMixer-based-models. Trockman et al. proposed the ConvMixer [15] which have recently gained attention in the computer vision community for their ability to capture global features of images.

Our proposed ConvMixer-based approach offers a unique balance of efficiency, global context capture, and accuracy specifically tailored for lung segmentation in 2D X-ray images. By comparing its performance with the aforementioned works and highlighting its unique advantages, we aim to demonstrate its significance and potential in this field. Additionally, we suggest exploring the integration of attention mechanisms within our architecture to potentially improve feature representation and boundary localization, drawing inspiration from Dense Attention U-Net's approach.

III. PROPOSED METHOD

A. Network architecture

1) Encoder-Decoder architecture

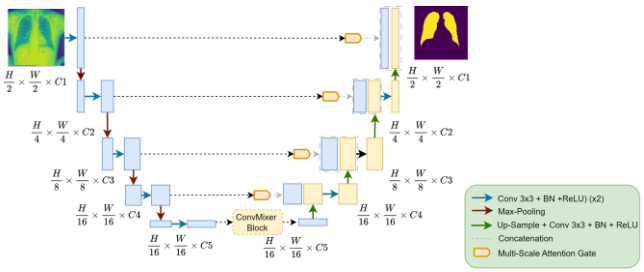

Fig 1: Overview of the proposed architecture

To solve the previously mentioned challenges, we present a novel lung segmentation architecture that harnesses the combined power of encoder-decoder networks and attention mechanisms for accurate delineation of lung regions. As illustrated in Figure 1, the proposed architecture employs both an encoder-decoder network and an attention gate mechanism. This design enables the model to capture both large-scale, significant features and, to a lesser extent, prioritize smaller features. The attention gate is strategically placed after each decoding layer to selectively focus on crucial areas in the segmentation features while suppressing irrelevant details. This combination of modules has shown effectiveness in tackling various lung segmentation challenges.

The proposed architecture consists of two primary stages: an encoder and a decoder. The input image is passed through the encoder where the model leverages conventional convolutions to extract high level semantic information. The encoder follows the typical architecture of a convolutional neural network. It comprises five convolutional levels arranged in a top-bottom fashion. Each level includes two ordinary convolution blocks and a down-sampling operation. More specifically, each ordinary convolution block consists of a convolution layer, a batch normalization layer, and a ReLU activation function, all of which utilize a 3x3 kernel size, 1 stride, and 1 padding. Down-sampling within the encoder is achieved through a max pooling operation with a 2x2 window size.

The extracted features are then processed by the ConvMixer model, which consists of a sequence of N ConvMixer layers. Each ConvMixer layer utilizes a two-step process:

1- Depthwise convolution: This involves applying grouped convolution, where the number of groups equals the number of channels (h) in the feature maps.

2- Pointwise convolution: This step employs a 1x1 kernel size to further refine the features.

Following each convolution, an activation function (GELU) and batch normalization are applied to enhance the model's learning capability:

$$F(X) = BN\left(\sigma\left(Conv(X, G1)\right)\right) + X \quad (1)$$

Where F(X) represents the output feature map, σ represents the GELU activation function, BN denotes batch normalization and Conv(X,G1) represents the convolution operation (pointwise convolution or depthwise convolution) applied to X with kernel G1.

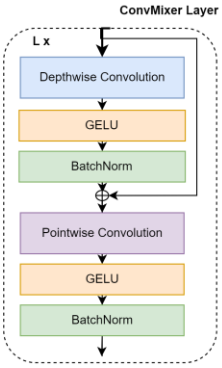

Fig 1. ConvMixer Block

In the decoder stage, the features extracted from the multi-level feature aggregation are fused with the corresponding up-sampled features, resulting in precise localization. In every step in the decoder stage, we do the upsampling of the feature map followed by a convolution layer, a batch normalization layer and ReLU activation function.

The decoder stage refines the extracted features for precise localization of the lung region. This is achieved by strategically combining them with progressively upsampled information. With each step of the decoder, the following operations take place: (1) Upsampling: The feature map is enlarged spatially to match the resolution of the output image. This increases the level of detail and enables precise localization. (2) Convolution layer: This layer extracts relevant features from the upsampled and combined feature maps. (3) Batch normalization layer: This step improves the stability of the training process by normalizing the outputs of the previous layer across mini-batches. (3) ReLU activation function: This introduces non-linearity into the model, allowing it to learn complex relationships between the features and ultimately improve the segmentation accuracy.

2) Attention mask

Previous approaches leveraging attention mechanisms have shown success in various segmentation tasks. However, a frequent limitation is their reliance on global pooling operations (average or max) to generate spatial attention weights. These techniques essentially average or take the maximum value across all feature channels, potentially neglecting crucial local information and leading to context-blindness.

To address this issue, we propose a novel mask attention module with skip-connections (see Figure 3 for its structure). This module incorporates a supervised learning branch to enhance the model's ability to understand contextual information at various scales. The input to the mask attention module is the feature map (f) obtained from the decoder at a specific stage. To effectively capture features at different resolutions, we employ three different convolution operations with varying receptive fields (coverage areas).

Firstly, Pointwise convolution; This convolution focuses solely on the feature channels without considering spatial relationship. This allows the module to capture global context and understand the overall feature distribution. Secondly, we use regular convolution with a 3x3 kernel, stride of 1, and padding of 1 to capture local spatial dependencies within a small neighborhood. This enables the module to understand the relationships between features in close proximity. Finally, The dilated convolution with a 3x3 kernel, stride of 1, padding of 2, and a dilation rate of 2 is used to capture context and dependencies over a larger spatial extent. The increased dilation rate allows the receptive field to expand without increasing the number of parameters in the filter, enabling the module to capture long-range dependencies within the feature map.

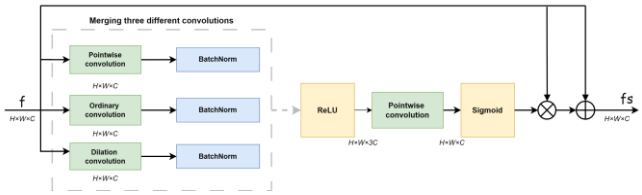

Fig 2: Multi-scale attention gate

Following each convolutional layer, a batch normalization step is applied to enhance the model's stability and training process. The feature maps generated by these diverse convolution operations are then combined to create a richer representation encompassing features at various scales. This allows the model to effectively understand contextual information at various levels, leading to improved segmentation accuracy.

$$h_C = \begin{pmatrix} \sigma_1(Concat\{BN\{PointwiseConv(h)\}, \\ BN\{OrdinaryConv(h)\}, \\ BN\{DilationConv(h)\}\}) \end{pmatrix} \quad (2)$$

$$h_s = h \times \sigma_2\left(PointwiseConv(h_c)\right) + h \quad (3)$$

Where h is the encoding features, h_c represents the concatenated features and h_s symbolize the output feature from the attention gate. The used activation functions are: σ_1 and σ_2 . σ_1 is ReLU and σ_2 is Sigmoid.

IV. EXPERIMENTAL RESULTS

A. Dataset

In this work, we evaluated our proposed method on two publicly available datasets [16] commonly used in chest X-ray analysis: Shenzhen and Montgomery. These datasets all contain chest X-ray images along with corresponding masks, encompassing a variety of abnormalities such as effusions and miliary patterns (as illustrated in Table 1).

TABLE 1: SHENZHEN AND MONTGOMERY DATASET DISTRIBUTION

Dataset	Disease	Total	Training	Test
Shenzen	Tuberculois	336	235	101
	Nontuberculosis	326	228	98
Montgomery	Tuberculosis	58	41	17
	Nontuberculosis	80	56	24

The Shenzhen dataset [16] consists of 662 chest X-ray images with corresponding binary masks. The images are

categorized as showing normal conditions or tuberculosis disease. Half from healthy patients and half from those with tuberculosis. You can see examples of images from the Shenzhen dataset in Figure 4. The Montgomery dataset [16] comprises 138 CXR images, divided into two groups: 80 from healthy individuals and 58 from patients diagnosed with tuberculosis. These images boast high resolution (either 4020 x 4892 or 4892 x 4020 pixels) and utilize a 12-bit grayscale system. You can see examples of images from the Montgomery dataset in Figure 5. While maintaining high resolution, the images in this dataset exhibit varying sizes and employ an 8-bit grayscale system.

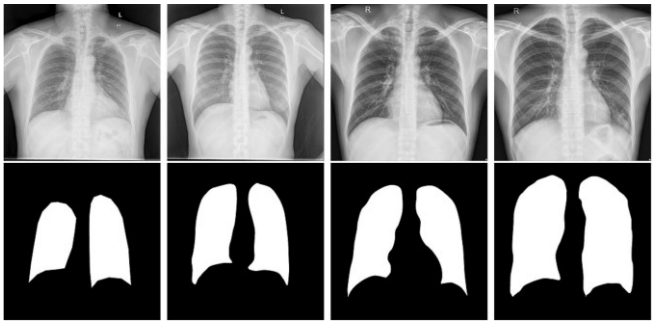

Fig.4: Training dataset sample with the corresponding ground truth for Shenzhen Dataset

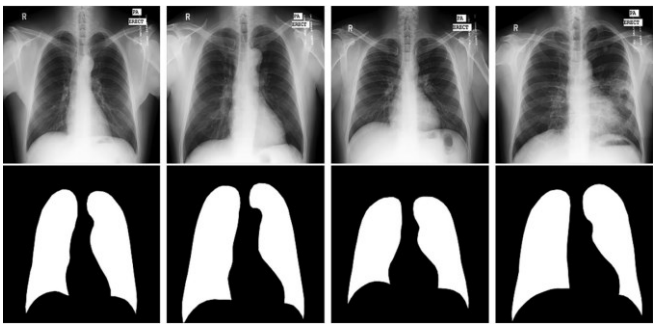

Fig. 5: Training dataset sample with the corresponding ground truth for Montgomery Dataset

Our initial approach involved training the model on the Shenzhen dataset and evaluating its performance on the Montgomery dataset. However, upon closer inspection, we discovered significant differences in the characteristics of the two datasets. Notably, the Shenzhen dataset contained images with a lower lung-to-image ratio compared to those in the Montgomery dataset. This inconsistency could potentially lead to biased or inaccurate results if the model were trained solely on the Shenzhen data and applied to the Montgomery data. To address this challenge, we opted to load and combine the datasets separately, ensuring a more homogenous training base for the model.

B. Implementation

Our experiments began by converting the images to PNG format as they were in JPEG and DICOM formats. We also standardize the input images to a uniform size of 512x512 pixels. To extract the object's outline, we applied a contour detection technique to the segmentation mask. The outline thickness was set to 8 pixels. To prevent the model from adapting too closely to the training data and potentially performing poorly on unseen data (overfitting), we employed online data augmentation during training. This involved randomly flipping the images horizontally and vertically. We optimized the learning process using the Adam optimizer, initially setting the learning rate to 0.0001 and the momentum to 0.9. We trained the model in batches of 8 images at a time. To avoid excessively long training times, we limited the maximum number of training cycles to 80.

The loss function utilized is the combination of Binary Cross Entropy and Dice where y is the ground truth target map and $\tilde{y}$ is the predicted map:

$$L(y,\tilde{y}) = 0.5\, BCE(\tilde{y},y) + Dice\,(\tilde{y},y) \qquad (4)$$

During the testing phase, while the model could generate segmentation results at various scales, we used the 512x512 segmentation output for final performance evaluation on a given chest X-ray image.

To evaluate the model's performance, we employed a standard technique called "random train-test split" on the combined dataset (combining data from Shenzhen and Montgomery). This approach involves randomly shuffling all the data points and then dividing them into two sets: a training set (typically 70% of the data) used to train the model and a testing set (typically 30% of the data) used to assess the model's ability to generalize to unseen data.

C. Evaluation metrics

In order to comprehensively evaluate our segmentation model's performance, we employed a combination of five established metrics, going beyond the commonly used Intersection over Union (IoU). This approach provides a more nuanced understanding of the model's strengths and weaknesses. Here's a breakdown of the metrics used where TP is true positives, FN is false negatives and FP is false positives:

- Intersection over Union (IoU): This metric reflects the overlap between predicted and actual segmentation masks:

$$IoU = \frac{Intersection\ Area}{Union\ Area}$$

- Recall: This metric focuses on the model's ability to identify all relevant pixels within the object of interest:

$$Recall = \frac{TP}{TP + FN}$$

- Precision: Contrary to recall, precision emphasizes the model's accuracy in predicting positive pixels:

$$Precision = \frac{TP}{TP + FP}$$

- F1-Score: This metric combines the strengths of recall and precision into a single score, providing a balanced view of the model's performance:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Accuracy: This metric represents the overall correctness of the model's predictions, encompassing both true positives and negatives:

$$Accuracy = \frac{Correct\ predictions}{All\ predictions}$$

V. RESULTS AND DISCUSSION

Applying the segmentation model directly to unprocessed images yielded several challenges. First, the model incorrectly classified non-lung regions with similar grayscale intensities to lungs as part of the lung region. Second, the model failed to accurately segment the entire lung area, missing a small portion within the actual lung boundaries.

To address the issues encountered during segmentation, we implemented a two-step post-processing approach. Firstly, we utilized connected-domain analysis to remove erroneously identified regions that did not truly belong to the lung. This technique helps identify and eliminate isolated pixels or small clusters that are mistakenly classified as lung tissue due to their similar grayscale intensity. Secondly, to recover the missing section of the entire lung area, we employed an ecological opening operation. This morphological operation aims to remove small objects (like the missing lung area) while preserving the overall shape and connectivity of the larger object (the lung itself). The efficacy of these post-processing steps is demonstrated in the Figure 6.

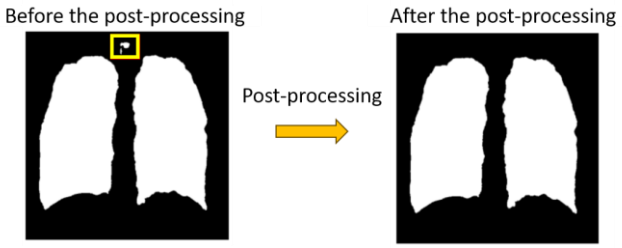

Fig. 6: Comparison between the before the post-processing and after. The yellow box is an irrelevant feature.

The experimental results of our proposed method in comparison with the state-of-arts are illustrated in the Table 2 below.

TABLE 2: COMPARISON OF THE PROPOSED APPROACH WITH THE RELATED WORKS

Model	IoU (%)	Recall (%)	Precision (%)	F1 (%)	Accuracy (%)
U-Net [7]	70.2	80.62	83.71	80.82	96.81
TransUnet [12]	67.02	75.61	82.34	76.38	95.87
Attention U-Net [8]	70.51	82.23	83.71	80.25	96.92
Our Proposed approach	**92.71**	**93.14**	**93.05**	**91.58**	**97.52**

Our model outperforms others in segmentation, as shown in Table 2. It achieves the highest performance, with a 22% improvement in IoU and a 10% increase in F1 score. While TransUnet needs substantial training data and struggles with limited datasets, our method excels by considering both global and local contexts, effectively identifying task-relevant features. This combined analysis of global and local information, along with the ability to pinpoint crucial features, underpins the strengths of our proposed model.

Figure 7 and Figure 8 display results of segmentation of our proposed approach where red represents the prediction results and the green represents the ground truth. We can see clearly that our model produces precise and detailed lesion regions and shapes regardless if the lung is healthy or unhealthy.

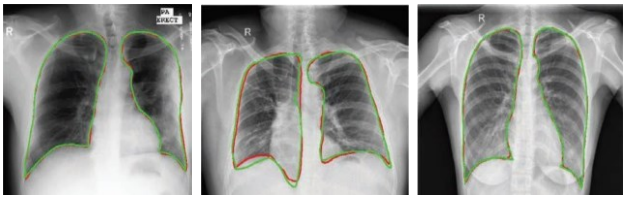

Fig. 7: Ground Truth (Green) VS prediction results (Red) – Normal lung on the left, others are unhealthy lungs

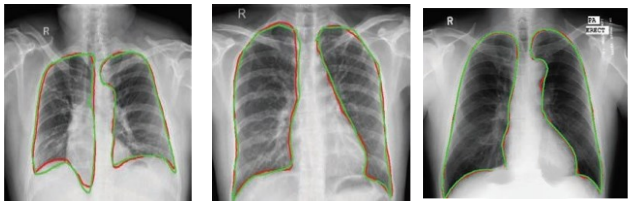

Fig. 8: Ground Truth (Green) VS prediction results (Red) - Normal lung on the left, others are unhealthy lungs

VI. CONCLUSION

Lung segmentation, a cornerstone of medical image analysis, faces challenges like imprecise boundaries, lesion-induced artifacts, and limited multiscale information utilization. This paper proposes a novel framework for lung segmentation, addressing these issues with an encoder-decoder structure and two innovative modules. The ConvMixer module captures intricate patterns, dependencies, and spatial information within feature maps, potentially enhancing performance. Additionally, a mask attention module with skip connections empowers the model to grasp contextual information across various scales. Finally, a post-processing step was added to refine the predicted results.

Evaluations on Shenzhen and Montgomery datasets demonstrate that our framework outperforms U-Net based models and state-of-the-art methods in lung segmentation. Future endeavours include: (1) extending the model's application to other medical image segmentation tasks and (2) developing an end-to-end deep learning model for simultaneous lung segmentation and disease classification.

VII. REFERENCES

[1] L. Xi, K. Li, R. Yang and L.-S. Geng, "Review of Deep Learning Based Automatic Segmentation for Lung Cancer Radiotherapy," *Frontiers Media,* vol. 11, 2021.

[2] L. J. Isaksson, P. Summers, F. Mastroleo, G. Marvaso, G. Corrao, M. G. Vincini, M. Zaffaroni, F. Ceci, G. Petralia, R. Orecchia and B. A. Jereczek-Fossa, "Automatic Segmentation with Deep Learning in Radiotherapy," *Multidisciplinary Digital Publishing Institute,* vol. 15, no. 17, pp. 4389-4389, 2023.

[3] R. Zainab, K. Bangul, A. Saad, K. Samiullah and I. Md Shohidul , "Lung Tumor Image Segmentation from Computer Tomography Images Using MobileNetV2 and Transfer Learning," *Multidisciplinary Digital Publishing Institute,* vol. 10, no. 8, pp. 981-981, 2023.

[4] L. e. a. Lenchik, "Automated segmentation of tissues using CT and MRI," in *Academic radiology*, 2019.

[5] X. Liu, L. Song, S. Liu and Y. Zhang, *A Review of Deep-Learning-Based Medical Image Segmentation Methods,* 2024.

[6] M. B. A. P. N. e. a. Murugappan, "Automated semantic lung segmentation in chest CT images using deep neural network," *Neural Comput & Applic,* 2023.

[7] P. F. a. T. B. Olaf Ronneberger, "U-net: Convolutional networks for biomedical image," *International Conference on Medical image computing and computer-assisted intervention. Springer,* p. 234–241, 2015.

[8] J. S. Ozan Oktay, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999,* 2018.

[9] M. Z. e. a. Alom, "Recurrent residual U-Net for medical image segmentation," *Journal of Medical Imaging,* pp. 014006-014006, 2019.

[10] Z. e. a. Zhang, "DENSE-INception U-net for medical image segmentation.," *Computer methods and programs in biomedicine 192,* 2020 : 105395.

[11] N. S. N. P. Ashish Vaswani, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[12] Y. L. Jieneng Chen, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv : 2102.04306,* 2021.

[13] W. Z. X. L. P. L. S. S. R. A. G. X. L. J. M. a. S. P. nwei Xie, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *Advances in Neural Information Processing Systems (NeurIPS),* 2021.

[14] W. Z. X. L. Y. L. a. S. P. Zhengxin Li, "SegViT: Semantic Segmentation with Plain Vision Transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] A. T. a. J. Z. Kolter, "Patches are all you need?," *arXiv preprint,* 2022.

[16] C. S. Jaeger S, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant Imaging Med Surg,* pp. 2223-4292, 2014.

[17] M. Z. e. a. Alom, "Recurrent residual U-Net for medical image segmentation.," *Journal of Medical Imaging,* pp. 014006-014006, 2019.

[18] Y. Z. H. J. Chen Lingdong, "Development of lung segmentation method in x-ray images of children based on TransResUNet," *Frontiers in Radiology,* vol. 3, 2023.

[19] Z. M. M. R. S. N. T. Zhou, "Unet++: A nested u-net architecture for medical image segmentation," *Deep learning in medical image analysis and multi-modal learning for clinical decision support, Springer,* pp. 3-11, 2018.

[20] X. Z. S. R. a. J. S. Kaiming He, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.