TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 1006 – 1020
© TÜBİTAK
doi:10.55730/1300-0632.4031

Focal modulation network for lung segmentation in chest X-ray images

Şaban Öztürk[1,2,3*], **Tolga Çukur**[2,3,4]

[1]Department of Electrical and Electronics Engineering , Faculty of Engineering, Amasya University, Amasya, Turkiye
[2]Department of Electrical and Electronics Engineering , Faculty of Engineering , Bilkent University, Ankara, Turkiye
[3]National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkiye
[4]Neuroscience Program, Bilkent University, Ankara , Turkiye

Received: 16.06.2023 • **Accepted/Published Online:** 09.08.2023 • **Final Version:** 27.10.2023

Abstract: Segmentation of lung regions is of key importance for the automatic analysis of Chest X-Ray (CXR) images, which have a vital role in the detection of various pulmonary diseases. Precise identification of lung regions is the basic prerequisite for disease diagnosis and treatment planning. However, achieving precise lung segmentation poses significant challenges due to factors such as variations in anatomical shape and size, the presence of strong edges at the rib cage and clavicle, and overlapping anatomical structures resulting from diverse diseases. Although commonly considered as the de-facto standard in medical image segmentation, the convolutional UNet architecture and its variants fall short in addressing these challenges, primarily due to the limited ability to model long-range dependencies between image features. While vision transformers equipped with self-attention mechanisms excel at capturing long-range relationships, either a coarse-grained global self-attention or a fine-grained local self-attention is typically adopted for segmentation tasks on high-resolution images to alleviate quadratic computational cost at the expense of performance loss. This paper introduces a focal modulation UNet model (FMN-UNet) to enhance segmentation performance by effectively aggregating fine-grained local and coarse-grained global relations at a reasonable computational cost. FMN-UNet first encodes CXR images via a convolutional encoder to suppress background regions and extract latent feature maps at a relatively modest resolution. FMN-UNet then leverages global and local attention mechanisms to model contextual relationships across the images. These contextual feature maps are convolutionally decoded to produce segmentation masks. The segmentation performance of FMN-UNet is compared against state-of-the-art methods on three public CXR datasets (JSRT, Montgomery, and Shenzhen). Experiments in each dataset demonstrate the superior performance of FMN-UNet against baselines.

Key words: Focal modulation, lung segmentation, chest x-ray, transformer, attention

1. Introduction

Pulmonary diseases, notably tuberculosis, and lung cancer remain prominent causes of global mortality despite the potential for effective interventions following early detection [1]. Most pulmonary nodules are detected accidentally on chest X-ray (CXR) images acquired for other assessments, as the characteristic symptoms associated with pulmonary disease are typically not evident until later stages. While CXR is a pervasive diagnostic modality given its low operational cost and widespread availability compared to other modalities such as computer tomography (CT) and magnetic resonance imaging (MRI) [2], detection of abnormalities on

[*]Correspondence: saban.ozturk@amasya.edu.tr

CXR images is still a challenging task [3]. In particular, a high level of expertise and focus is required to identify relatively smaller nodules under the immense diagnostic workload for radiologists [4]. The globally increasing workload of radiologists is anticipated to pose a significant diagnostic challenge in the long run, not only in low- but also high-income countries [5]. Therefore, having an automated computer-aided diagnostic (CAD) system to read CXR images is highly desirable to effectively reduce processing times and improve diagnostic accuracy, especially in the early stages of the disease where the imaging markers of progression might require more careful inspection [6].

To achieve accurate assessment and diagnosis of pulmonary diseases, a proficient CAD system should possess the ability to automatically detect and identify regions exhibiting pathological abnormalities [7]. Characteristic structural properties of the lungs, including their shape, size, and area, form the primary focus of investigation for CXR-CAD systems, as they offer crucial insights into pulmonary disease progression. An important stage in CAD systems is the localization of the lung tissues in CXR images by segregation of background regions, such that the focus can be directed to disease-related regions [8]. This lung segmentation procedure is a challenging task due to several reasons. First, there are variations in the shape and size of the lung and pathological abnormalities based on factors such as gender, age, genetic differences, and heart size. Second, there are shape distortions caused by severe lung disease or opacities and consolidations resulting from various infections. Third, overlapping structures can be present in lung regions due to projection-based imaging in CXR. Lastly, foreign objects such as cardiac pacemakers and other implanted devices can be present [9]. These factors present a significant challenge in the delineation of the lung boundaries and they can compromise segmentation performance [10].

Over the years, numerous important approaches have been proposed in the literature to improve performance in automatic lung segmentation based on traditional and learning-based techniques. Traditional segmentation methods typically rely on hand-crafted features extracted via predefined image filters. A group of methods has focused on edge detection for segmentation. Saad et al. [11] employ the Canny edge detection filter and morphological operations, Xu et al. [12] utilize a principal component analysis-based active shape model for global edge detection, and Liu et al. [13] apply a multi-scale Sobel operator based approach for edge detection. Other studies have focused on clustering to achieve segmentation. Ahmad et al. [14] employed the Fuzzy C-Means (FCM) clustering approach, and Sangamithraa and Govindaraju [15] utilized the EK-mean clustering approach. Intensity-thresholding methods have also been proposed to delineate lung tissue. Leader et al. [16] employed a slice-based pixel-value thresholding method, while Annangi et al. [17] utilized the Otsu thresholding method. Candemir et al. [18] employed a nonrigid registration-driven method for the segmentation of retrieval-based patient-specific lung model segmentation. Traditional methods for lung segmentation in CXR images commonly depend on a limited set of hand-crafted features that fail to capture a comprehensive representation of the image distribution [19].

To address shortcomings of traditional methods, recent studies have instead focused on learning-based methods for lung segmentation [20]. In particular, various deep learning architectures have been proposed including fully-connected [23, 24], convolutional [21, 22, 25–27, 31, 37, 39], adversarial [28–30], recurrent [9, 32, 34], and graphical models [33]. Yet, the majority of the studies in the literature adopt UNet-based architectures following an hourglass structure that offers a good compromise between performance and efficiency [35]. The vanilla UNet model is based on the convolutional encoder and decoder stages with skip connections. As such, it shows limited sensitivity to long-range context between distant image features, and limited adaptation to sample-specific image features. Several enhanced designs have been proposed to boost segmentation performance further. Arora et al. [36] incorporated channel and spatial attention blocks. Ghali and Akhloufi [38] incorporated attention gates between the deconvolutional layers of the UNet decoder stage. Cao and Zhao [40] introduced a three-terminal attention mechanism that combines the channel and spatial attention modules. While attention-augmented UNet models typically show improved adaptation to sample-specific features, they can still suffer from limited capture of long-range context. Recently, vision transformer (ViT) architectures have been adopted to capture long-range context between image patches [41, 44]. Chen et al. [42] introduced the TransUNet architecture by integrating the ViT between the encoder and decoder stages of UNet. Chen et al. [43] enhanced the TransUNet architecture by incorporating ViT into an already attention-augmented UNet. Although ViT modules excel in capturing long-range relationships in CXR images, their practical use is limited by the quadratic computational complexity with respect to input image size. As such, ViT-based methods commonly adopt either a coarse-grained global self-attention mechanism with limited resolution or a fine-grained local self-attention mechanism with limited coverage. In turn, this compromises either the spatial precision or the contextual sensitivity of transformer modules.

To address the abovementioned limitations, here we propose a novel segmentation method named FMN-UNet that combines convolutional modules with a focal modulation network (FMN) [45]. By leveraging a focal attention mechanism as opposed to the regular self-attention mechanism in ViT, FMN-UNet effectively captures the relationships between both fine-grained local and coarse-grained global features while alleviating computational costs. Receiving as input a high-resolution CXR image, FMN-UNet projects the input through a convolutional encoder stage to extract a relatively compact latent representation. An FMN stage further processes encoded feature maps at the lowest and highest resolutions to capture both global and local contextual relationships between image features. Finally, the feature maps are projected through a convolutional decoder stage to produce segmentation masks. To the best of our knowledge, this is the first study that investigates the potential of focal modulations combined with UNet for CXR lung segmentation. The main contributions of FMN-UNet can be summarized as follows:

- FMN-UNet leverages focal attention mechanisms at multiple resolutions to effectively capture both fine-grained local and coarse-grained global visual features.
- The focal attention mechanism is incorporated into a convolutional encoder-decoder architecture, ensuring computational efficiency without compromising performance.
- FMN-UNet is demonstrated on three public CXR datasets to showcase its superiority over state-of-the-art lung segmentation methods.

The rest of this paper is structured as follows: Section 2 provides notations and details of the FMN-Unet architecture. Section 3 encompasses information about the datasets, model implementation details, evaluation metrics, and competing methods. Section 4 presents the ablation study and comparative analyses with other methods. Lastly, Section 5 discusses the implications of our findings.

2. Methodology

2.1. Notations

Given a CXR image repository denoted by $D = \{x_n, y_n\}_{n=1}^{N}$, where N represents the total number of images in D, $x_n \in \mathbb{R}^{H\text{x}W\text{x}C}$ denotes the input CXR image, $y_n \in \{0,1\}^{H\text{x}W}$ represents the lung segmentation mask, HxW indicates the image resolution, and C denotes the number of channels in the input image. The main objective of the FMN-UNet architecture is to achieve a high-performance lung segmentation task by generating a detailed segmentation mask directly from the input CXR image in an end-to-end manner, $f_{FMN-UNet} : x_n \in \mathbb{R}^{H\text{x}W\text{x}C} \rightarrow \hat{y}_n \in \{0,1\}^{H\text{x}W}$, where $\hat{y}_n$ represents predicted output mask. To fulfill this objective, the FMN-UNet architecture can be dissected into three primary components: the encoder stage (f_{enc}), the FMN stage (f_{FMN}), and the decoder stage (f_{dec}). In the encoder stage, features are extracted from the CXR image and mapped down, $f_{enc} : x_n \in \mathbb{R}^{H\text{x}W\text{x}C} \rightarrow m_n \in \mathbb{R}^{P\text{x}P\text{x}C_m}$, where P denotes map dimensions and C_m denotes the number of map channel. The FMN stage reveals the fine-grained local and coarse-grained global relationships of the feature map, $f_{FMN} : m_n \in \mathbb{R}^{P\text{x}P\text{x}C_m} \rightarrow \hat{m}_n \in \mathbb{R}^{P\text{x}P\text{x}C_m}$. Finally, decoder stage creates a predicted output lung mask, $f_{dec} : \hat{m}_n \in \mathbb{R}^{P\text{x}P\text{x}C_m} \rightarrow \hat{y}_n \in \{0,1\}^{H\text{x}W}$.

2.2. FMN-UNet

In this study, we proposed the FMN-UNet model for improved performance and efficiency in lung segmentation tasks. Inspired by UNet, FMN-UNet uses convolutional encoder and decoder stages with progressively lowered and elevated spatial resolution to follow an hourglass shape, and it uses skip connections to propagate encoded features maps at each resolution to the decoder layers at corresponding levels. To improve the capture of contextual representations, FMN-UNet incorporates an FMN stage in between the encoder and decoder stages. The FMN stage comprises single FMN stages between the encoder and decoder stages at the lowest and highest resolutions. FMN stages leverage a focal attention mechanism to capture contextual features without introducing a substantial computational burden. Figure 1 illustrates the basic architecture of FMN-UNet and its submodules. The remaining parts of this section describe in detail the encoder, FMN, and decoder stages of FMN-UNet along with its loss function.

The encoder stage of the FMN-UNet architecture serves to embed the input CRX image into a relatively lower dimensional latent representation, which is expected to suppress noise and other artifactual features irrelevant to the segmentation task. The encoder stage comprises four down-sampling blocks, with each block composed of two convolution layers, one batch normalization layer, one dropout layer, and one max pooling layer. The down-sampled feature map, $m_n^{P\text{x}P\text{x}C_m}$, for an input sample, $x_n^{H\text{x}W\text{x}C}$, is calculated as in Equation 1, $m_n = f_{enc}(x_n)$.

$$m_n = P_{max}\left(Drop\left(BN\left(Conv\left(Conv\left(x_n\right)\right)\right)\right)\right) \tag{1}$$

where P_{max} represents the max pooling operator, $Drop$ denotes the dropout operation, BN represents batch normalization, and $Conv$ denotes the 2D convolution operation. To reveal the feature map's fine-grained local and coarse-grained global feature relationships, m_n is first split into tokens, m_{nj} with J tokens. Then the focal modulation mechanism is used to compute contextual representations as described in Equation 2.

$$\hat{m}_n = q\left(m_{nj}\right) \odot h\left(s\left(j, m_n\right)\right) \tag{2}$$

where $q(.)$ denotes a query projection function, $h(.)$ is a linear layer to obtain the modulator, $s(.)$ represents context aggregation function and $\odot$ is the element-wise multiplication. $s(.)$ can be expressed using spatial- and level-aware gating weights G, and linear layer Z with focal level l as in $\sum_{l=1}^{L+1} g_j^l \cdot z_j^l$. To calculate G and Z:

$$G^l = GeLU\left(Z^{l-1}\right) \tag{3}$$

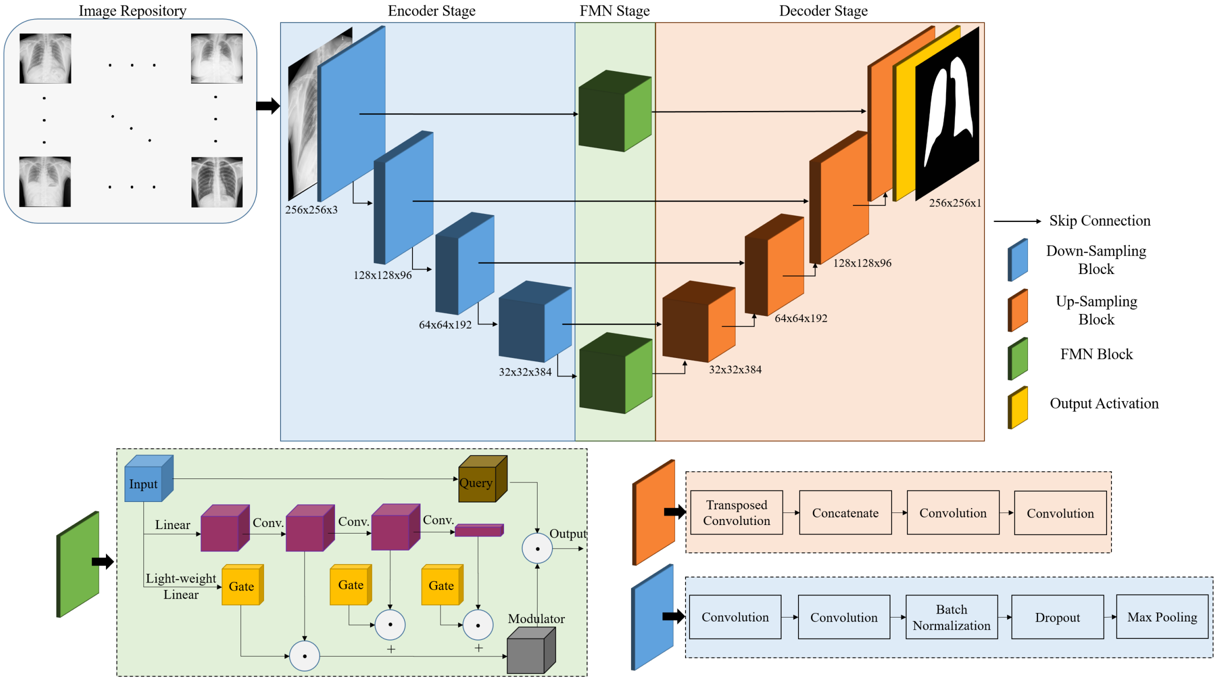

Figure 1. The basic structure of the proposed FMN-UNet model for lung segmentation. The encoder stage is represented in blue and consists of four convolutional down-sampling blocks. The FMN stage is shown in green and consists of a single FMN stage inserted between feature maps at the highest and lowest resolution levels. The decoder stage is depicted in orange and consists of four convolutional up-sampling blocks. The layers within each block are detailed in the bottom part of the figure.

$$Z^l = GeLU\left(DWConv\left(Z^{l-1}\right)\right) \tag{4}$$

where $DWConv$ is a depth-wise convolution operator and the $GeLU$ activation function is used. Equation 2 can be rewritten at the token level as in Equation 5.

$$\hat{m}_n = q\left(m_{nj}\right) \odot h\left(\sum_{l=1}^{L+1} g_j^l \cdot z_j^l\right) \tag{5}$$

The decoder stage of FMN-UNet serves to synthesize high-resolution segmentation maps from the extracted contextual feature representations. In other words, $\hat{m}_n$ is converted to a lung segmentation mask in the original image dimensions (HxWxC). The decoder stage comprises four up-sampling blocks, with each block composed of one transposed convolution layer, one concatenate layer, and two convolution layers. The concatenate layer combines each up-sampling layer with the corresponding layer in the encoder stage. The skip connections between corresponding encoder-decoder stages allow for the transmission of low-level features from the encoder to the decoder, enabling the capture of fine details and local information in the segmentation maps. Equation 6 described the overall derivation of the segmented lung mask $\hat{y}_n$ by up-sampling the feature map $\hat{m}_n^{P\text{x}P\text{x}C_m}$.

$$\hat{y}_n = Conv\left(Conv\left(Concat\left(TransConv\left(\hat{m}_n\right)\right)\right)\right) \tag{6}$$

where *Concat* denotes concatenate layer and *TransConv* represents transposed convolution layer.

Since CXR lung segmentation is a binary segmentation task, each pixel in the segmentation mask at the output of the decoder stage should be in the range of 0, 1. To achieve this, the tangent hyperbolic function (tanh) is used at the decoder stage. As the tanh function produces values in the range of [-1, 1], negative values are thresholded to 0 to obtain the final segmentation mask. In order for each pixel in the segmentation mask to converge the ground truth value, we utilize Binary Cross Entropy loss (BCE) as depicted in Equation 7. The BCE loss, which penalizes binary values on a pixel-by-pixel basis, has been reported in the literature to offer a detailed capture of task-relevant features for lung segmentation.

$$\mathrm{BCE}\left(y_i, \hat{y}_i\right) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right] \tag{7}$$

3. Experimental design

3.1. Datasets

To comprehensively evaluate the lung segmentation performance of FMN-UNet, three publicly available CXR datasets are utilized: the Japanese Society of Radiological Technology (JSRT) dataset [47], the Montgomery County X-ray (MC) dataset [48], and The Shenzhen (SZ) dataset [48]. These datasets offer diverse and representative samples for assessing the effectiveness and generalization of FMN-UNet.

The JSRT dataset comprises a total of 247 posterior-anterior CXR images with dimensions of 2048x2048 pixels. Among these, 90 images depict healthy lungs without any abnormalities, while the remaining images contain various lung nodules. The MC dataset, created by the Department of Health and Human Services in Montgomery County, Maryland, USA, consists of 138 frontal CXR images with dimensions of either 4020x4892 or 4892x4020 pixels. Among these, 80 cases are healthy, while the remaining 58 cases show tuberculosis. The SZ dataset, collected by Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China, consists of 662 frontal CXR images with an average size of 3000x3000 pixels. Within this dataset, 326 cases are healthy, while the remaining 336 cases exhibit tuberculosis. To ensure consistency, all images were downsampled to 256x256 pixels using the nearest neighbor interpolation technique. The datasets are split into 80% for training, 10% for validation, and 10% for testing, while evaluation is conducted using the 10-fold cross-validation.

3.2. Architectural details and model implementation

The FMN-UNet takes input images of size 256x256 pixels with 3 channels representing RGB in CXR images. It generates lung segmentation masks of size 256x256x1 as the output. The encoder stage of FMN-UNet consists of down-sampling blocks, each comprising two convolution layers with 3x3 filters, one batch normalization layer, one dropout layer with a dropout factor of 0.3, and one max pooling layer with 2x2 filters. The ReLU activation function is applied to the convolution layers. The decoder stage of FMN-UNet includes up-sampling blocks, which consist of one transposed convolution layer, one concatenate layer, and two convolution layers, all with 3x3 filters and ReLU activation. The transposed convolution layer uses a 3x3 convolution kernel. The tanh function is used as output activation at the decoder output. For the FMN stage, the token patch size is set to 8x8 pixels, and the embedding dimensions are set to 768. Three basic focal layers are used in FMN, with a depth configuration of [2,3,2]. These layers have a focal level of 2 and a focal window size of 3. The classification layer and softmax activation are removed from the FMN output, allowing it to generate embedding vectors.

FMN-UNet was implemented using TensorFlow and executed on two NVidia RTX 3090 GPUs. The model was trained using the Adam optimization algorithm with a batch size of 24 and a learning rate of 5×10^{-4} for a total of 90 epochs. The base filter number for UNet was set to 48, and it automatically adjusted as per the original UNet multipliers in each block. The datasets are split into 80% for training, 10% for validation, and 10% for testing. A 10-fold validation technique is employed for evaluation.

3.3. Performance evaluation

The performance of FMN-UNet and other competing lung segmentation techniques is assessed using three common metrics, Dice Similarity Coefficient (DSC), Jaccard Index (JI), and Accuracy metric (ACC). The ACC evaluates the correctness of each pixel, while DSC and JI measure the spatial overlap between the binary ground truth and the predicted segmentation masks. DSC, JI, and ACC are calculated as in Equations 8, 9, and 10, respectively.

$$\text{DSC}\left(y_n, \hat{y}_n\right) = \frac{2\left|y_n \cap \hat{y}_n\right|}{\left|y_n\right| + \left|\hat{y}_n\right|} \tag{8}$$

$$\text{JI}\left(y_n, \hat{y}_n\right) = \frac{\left|y_n \cap \hat{y}_n\right|}{\left|y_n\right| + \left|\hat{y}_n\right| - \left|y_n \cap \hat{y}_n\right|} \tag{9}$$

$$\text{ACC}\left(y_n, \hat{y}_n\right) = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP represents true positive (correctly predicted white pixel), FP represents false positive (incorrectly predicted white pixel), TN represents true negative (correctly predicted black pixel), and FN represents false negative (incorrectly predicted black pixel).

3.4. Competing methods

To provide a comprehensive evaluation of FMN-UNet and emphasize FMN's unique contribution to the UNet architecture, we employed both UNet-based competing methods and other segmentation methods. The network architecture and training procedures for all methods were adapted from their respective original works. We strictly adhered to the hyperparameters specified in the original studies for each competing method. If there were any missing hyperparameters or incomplete experimental results in the competing methods, we conducted experiments using the relevant datasets to address these gaps.

For UNet-based competing segmentation models, we first use a medical image analysis guided UNet method [49]. Dense-UNet [50] enhances lung segmentation performance by incorporating dense connectivity between various layers. The improved UNet (i-UNet) [51] method uses pretrained Efficientnet-b4 as the encoder and the LeakyReLU activation function in the decoder to efficiently extract lung region and avoid gradient instability. The modified UNet (M-UNet) method [5] divides the image into smaller patches and performs the segmentation task using the information from two branches: a classifier CNN and a patch-based modified UNet. This approach leverages the advantages of both branches to achieve accurate segmentation results. TransUNet [42] architecture integrates the ViT between the encoder and decoder stages of the UNet architecture.

The A-LugSeg [9] technique consists of two subnetworks, an RCNN-based segmentation subnetwork and a refinement subnetwork for coarse segmentation. The AC-RegNet [52] provides an anatomically constrained neural networks-based contribution to CXR lung segmentation performance using anatomical priors into deep learning-based image registration. RecBSeg [2] is an automatic lung segmentation technique consisting of four steps: image acquisition, linear CNN-based initial segmentation, reconstruction using residual blocks in CNN, and final segmentation.

4. Experimental results

4.1. Ablation study

This section provides a detailed analysis of the effects of various components of the FMN-UNet architecture on lung segmentation performance. All ablation experiments have been performed using the JSRT dataset. Firstly, ablation experiments originating from UNet architecture are presented in Table 1. Table 1 (a) presents a comparison between the performance metrics of UNet, which was chosen as the segmentation backbone, and FCN, a popular structure that does not include skip connections. UNet outperforms FCN, achieving a 0.8% higher DSC, 1.5% higher JI, and 0.6% higher ACC with a moderate computational increase. In Section (b) of Table 1, the impact of UNet depth on the performance of FMN-UNet is examined. The number of trainable weights and memory consumption increases with increasing depth. Yet, the benefits from elevated degrees of depth might not influence performance beyond a certain level. Our results suggest that a depth of 5 offers a favorable compromise between complexity and performance, so FMN-UNet(5) has been selected as the main segmentation model in this study.

Table 1. Ablation experiments on UNet backbone. (a) Performance comparison between UNet and FCN architectures. (b) Impact of UNet depth on FMN-UNet.

		DSC	JI	ACC	Parameters
a	FCN	96.4±0.2	93.2±0.4	97.8±0.1	17.5 M
	UNet	97.2±0.3	94.7±0.6	98.4±0.2	19.5 M
b	FMN-UNet(4)	97.3±0.3	94.9±0.6	98.4±0.2	230.2 M
	FMN-UNet(5)	98.0±0.3	95.5±0.5	98.8±0.1	244.4 M
	FMN-UNet(6)	97.9±0.2	95.5±0.4	98.7±0.1	271.1 M

In Figure 2, the segmentation results obtained for the FCN, UNet, and FMN-UNet methods are shown using randomly selected test samples from the test dataset. In Figure 2, (a) and (d) show the lung segmentation masks, (b) and (e) depict the comparison of the boundaries of the lung segmentation masks with the ground truth, and (c) and (f) demonstrate the correctly and incorrectly identified regions by the competing methods and FMN-UNet on the original images. It can be observed that the segmentation mask generated by FCN is relatively smooth and fails to capture sharp transitions. UNet outperforms FCN in accurately identifying lung regions. However, among all the methods, FMN-UNet stands out as the most effective in capturing fine details and preserving intricate features. Upon examining the contours of each method's masks, it can be observed that the FMN-UNet's segmentation closely matches the original CXR image.

The impact of different variables in the FMN stage of the FMN-UNet architecture on performance is analyzed in Table 2. In Section (a) of Table 2, the depth and number of focal modulation layers are examined. While there is a noticeable improvement in performance with increasing depth, it also leads to a significant increase in computational complexity due to the growing number of parameters. Therefore, a focal modulation depth of (2,3,2) was selected to strike a balance. Furthermore, as the number of parameters to be trained increases, it becomes crucial to augment the training dataset size to mitigate the risk of overfitting. Section (b)

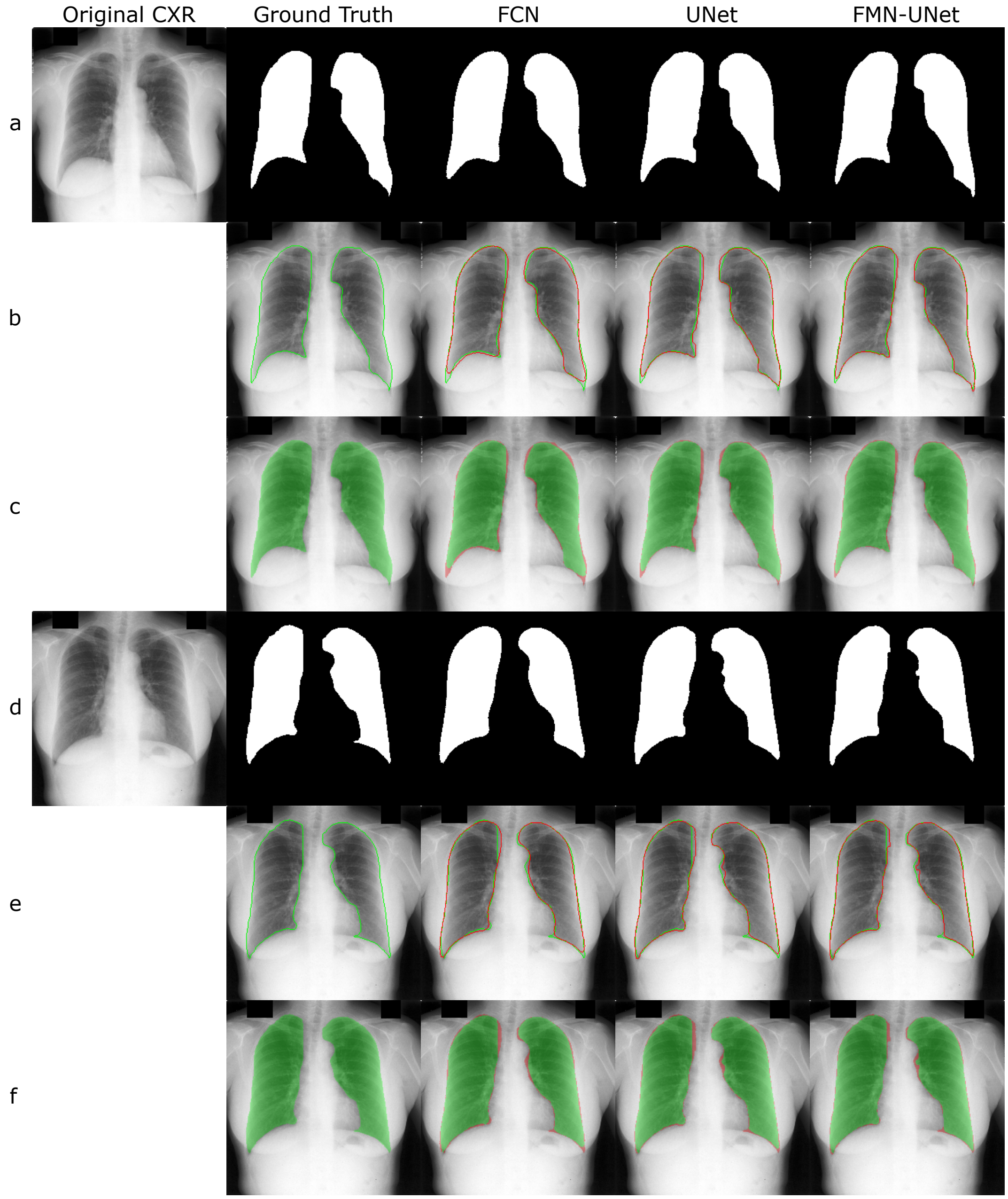

Figure 2. The segmentation performance of FCN, UNet, and FMN-UNet is evaluated. (a) and (d) show the lung segmentation masks, (b) and (e) depict the comparison of the boundaries of the lung segmentation masks with the ground truth, and (c) and (f) demonstrate the correctly and incorrectly identified regions. In (b) and (e), the green lines represent the ground truth information, while the red lines represent the segmentation mask produced by the relevant method. In (c) and (f), the green regions represent correctly identified, while the red regions represent incorrectly identified regions. Among the methods, FMN-UNet stands out as the most effective in capturing fine details and preserving intricate features.

of Table 2 investigates the impact of the embedding dimension parameter on performance. Choosing a higher embedding dimension, such as 1024, results in a considerable increase in trainable parameters, potentially leading to overfitting. Hence, an embedding dimension of 768 was preferred to maintain optimal performance without encountering overfitting issues.

Table 2. Ablation experiments on FMN stage. (a) Performance evaluation based on depth and number of focus modulation layers. (b) Performance evaluation based on embedding dimension parameter.

		DSC	JI	ACC	Parameters
a	FMN(2,2)	97.8±0.2	95.4±0.4	98.7±0.1	75.2 M
	FMN(2,3,2)	98.0±0.3	95.5±0.5	98.8±0.1	244.4 M
	FMN(4,6,4)	98.1±0.3	95.5±0.5	98.8±0.1	323.3 M
b	FMN(256)	97.7±0.3	95.4±0.6	98.6±0.1	59.6 M
	FMN(768)	98.0±0.3	95.5±0.5	98.8±0.1	244.4 M
	FMN(1024)	97.5±0.2	94.9±0.4	98.5±0.2	644.5 M

4.2. Comparative analysis

This section presents a comparative analysis of FMN-UNet against other UNet-based methods and other competing segmentation methods. To achieve this, the section is divided into two parts: competing segmentation methods and UNet-based segmentation methods. Firstly, in Section (a) of Table 3, three competing segmentation techniques, A-LugSeg, AC-RegNet, and RecBSeg, are compared with FMN-UNet for the CXR lung segmentation task. Compared to competing segmentation techniques, FMN-UNet demonstrates superior performance against competing segmentation methods by an average of 2.8% on the DSC metric, an average of 1.5% on the JI metric, and an average of 1.4% on the ACC metric on the JSRT dataset. In the MC dataset, it demonstrates superior performance against competing segmentation methods by an average of 2.9% on the DSC metric and an average of 1.4% on the ACC metric. In the JI metric, it lags behind other techniques by 0.6% on average. In the SZ dataset, it demonstrates superior performance against competing segmentation methods by an average of 3.8% on the DSC metric, an average of 4.0% on the JI metric, and an average of 1.0% on the ACC metric. According to the average of all datasets, FMN-UNet demonstrates superior performance against competing segmentation methods by an average of 3.1% on the DSC metric, an average of 1.6%, and an average of 1.5% on the ACC metric.

In Section (b) of Table 3, five UNet-based segmentation techniques, namely UNet, Dense-UNet, i-UNet, M-UNet, and TransUNet, are compared with FMN-UNet for CXR lung segmentation task. Compared to UNet-based segmentation techniques, FMN-UNet demonstrates superior performance against competing segmentation methods by an average of 0.7% on the DSC metric, an average of 0.4% on the JI metric, and an average of 0.6% on the ACC metric on the JSRT dataset. In the MC dataset, it demonstrates superior performance against competing segmentation methods by an average of 0.8% on the DSC metric, an average of 0.7% on the JI metric, and an average of 0.6% on the ACC metric. In the SZ dataset, it demonstrates superior performance against competing segmentation methods by an average of 0.2% on the DSC metric, an average of 0.2% on the JI metric, and an average of 0.2% on the ACC metric. According to the average of all datasets, FMN-UNet demonstrates superior performance against competing segmentation methods by an average of 0.6% on the DSC metric, an average of 0.4%, and an average of 0.5% on the ACC metric.

Table 3. Evaluation of AUC performance in relation to competing methods. (a) Competing segmentation techniques, (b) UNet-based techniques.

		JSRT			MC			SZ		
		DSC	JI	ACC	DSC	JI	ACC	DSC	JI	ACC
a	A-LugSeg [9]	97.3	95.8	97.2	97.1	95.5	97.0	-	-	-
	AC-RegNet [52]	94.3	-	-	95.3	-	-	93.1	-	-
	RecBSeg [2]	94.1	92.2	97.6	93.6	98.1	97.0	90.6	87.8	97.1
b	UNet [49]	97.2	94.7	98.4	97.9	95.9	99.0	95.4	91.6	97.9
	Dense-UNet [50]	97.6	95.3	-	97.9	95.9	-	-	-	-
	i-UNet [51]	97.9	95.8	98.5	97.7	95.5	98.9	-	-	-
	M-UNet [5]	96.5	95.0	97.4	96.1	94.3	97.3	-	-	-
	TransUNet [42]	97.4	94.7	98.4	98.1	96.0	99.1	95.4	91.7	97.9
	FMN-UNet	98.0	95.5	98.8	98.3	96.2	99.2	95.6	91.8	98.1

5. Discussion and conclusion

This study proposes an end-to-end CAD approach named FMN-UNet, which achieves high performance and relatively low computational complexity for lung segmentation in CXR images. To build FMN-UNet, we conducted an initial set of explorations to comparatively evaluate FCN and UNet backbones for encoding input images into a compact latent space. Corroborated by the result in Table 1, the UNet backbone was found superior to FCN for subsequent performance in the downstream segmentation tasks, so our proposed model leverages the UNet encoder in its initial stage. The traditional UNet architecture lacks sensitivity to long-range context and adaptation to sample-specific features. While attention-augmented CNN approaches can improve adaptation, they still show suboptimal sensitivity to a long-range context [36–38]. To boost sensitivity to long-range dependencies, vision transformer (ViT) models have been proposed for lung segmentation in CXR images [42–44]. Yet, canonical ViT models suffer from quadratic computational burdens with respect to image size, so direct integration of ViT and CNN modules can be challenging. To improve the capture of contextual representations while alleviating computational costs, FMN-UNet leverages local and global attentional mechanisms cascaded with convolutional encoding/decoding blocks. Note that FMN-UNet demonstrates superior performance against the TransUNet method, which directly inserts a ViT module in the midpoint of a UNet architecture, by an average of 0.3% on the DSC metric, an average of 0.4% on the JI metric, and an average of 0.2% on the ACC metric.

To comprehensively analyze the performance of FMN-UNet, a comparative analysis with both UNet-based segmentation techniques and other segmentation techniques is presented in Table 3. Since the proposed technique is implemented with 256x256 image dimensions, comparable techniques from the literature that include similar resolutions have been used in the comparisons. Overall, FMN-UNet demonstrates superior performance against baselines by an average of 1.8% on the DSC metric, an average of 1.0% on the JI metric, and an average of 1.0% on the ACC metric. Upon examining the visual results in Figure 2, it can be observed that FMN-UNet outperforms competing methods in capturing fine details.

Several avenues of development can further enhance the performance of FMN-UNet. The ability of the image feature maps or embedding vectors generated by the encoder stage to represent the original image significantly impacts the segmentation performance of the FMN-Unet. Therefore, robust deep learning models such as diffusion models can bring significant performance improvements in this area [53, 54]. Increasing image resolutions or utilizing feature dependencies at each stage to better model spatial context can improve performance, but it also significantly increases computational complexity. Managing this trade-off, especially when using high-resolution original images, will greatly enhance performance.

In future studies, our focus will be on generating more representative embedding vectors in the early network stages to address the mentioned limitations. Specifically, we will work on developing a deep encoder capable of producing more representative feature maps while managing the computational load for higher-resolution images and deeper architectures.

Acknowledgment

This work was supported by a TÜBİTAK 118C543 grant awarded to Ş.Öztürk. Ş.Ö.: Conceptualization, methodology, data curation, software, formal analysis, visualization, investigation, validation, writing – original draft. T.Ç.: Conceptualization, methodology, supervision, writing – original draft.

References

[1] Li X, Shen L, Xie X, Huang S, Xie Z et al. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. Artificial intelligence in medicine 2020; 103: 101744. https://doi.org/10.1016/j.artmed.2019.101744

[2] Souza JC, Diniz JOB, Ferreira JL, da Silva GLF, Silva AC et al. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. Computer methods and programs in biomedicine 2019; 177: 285-296. https://doi.org/10.1016/j.cmpb.2019.06.005

[3] Kvak D, Chromcová A, Biroš M, Hrubý R, Kvaková K et al. Chest x-ray abnormality detection by using artificial intelligence: A single-site retrospective study of deep learning model performance. BioMedInformatics 2023; 3 (1): 82-101. https://doi.org/10.3390/biomedinformatics3010006

[4] Torres-Mejía G, Smith RA, Carranza-Flores MDLL, Bogart A, Martínez-Matsushita L et al. Radiographers supporting radiologists in the interpretation of screening mammography: a viable strategy to meet the shortage in the number of radiologists. BMC cancer 2015; 15 (1): 1-12. https://doi.org/10.1186/s12885-015-1399-2

[5] Rahman MF, Zhuang Y, Tseng TLB, Pokojovy M, McCaffrey P et al. Improving lung region segmentation accuracy in chest X-ray images using a two-model deep learning ensemble approach. Journal of Visual Communication and Image Representation 2022; 85: 103521. https://doi.org/10.1016/j.jvcir.2022.103521

[6] Chowdary GJ, Kanhangad V. A Dual-Branch Network for Diagnosis of Thorax Diseases From Chest X-Rays. IEEE Journal of Biomedical and Health Informatics 2022; 26 (12): 6081-6092. https://doi.org/10.1109/JBHI.2022.3215694

[7] Alshmrani GMM, Ni Q, Jiang R, Pervaiz H, Elshennawy NM. A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. Alexandria Engineering Journal 2023; 64: 923-935. https://doi.org/10.1016/j.aej.2022.10.053

[8] Singh A, Lall B, Panigrahi BK, Agrawal A, Agrawal A et al. Deep LF-Net: Semantic lung segmentation from Indian chest radiographs including severely unhealthy images. Biomedical Signal Processing and Control 2021; 68: 102666. https://doi.org/10.1016/j.bspc.2021.102666

[9] Peng T, Gu Y, Ye Z, Cheng X, Wang J. A-LugSeg: Automatic and explainability-guided multi-site lung detection in chest X-ray images. Expert Systems with Applications 2022; 198: 116873. https://doi.org/10.1016/j.eswa.2022.116873

[10] Peng T, Xu TC, Wang Y, Li F. Deep belief network and closed polygonal line for lung segmentation in chest radiographs. The Computer Journal 2022; 65 (5): 1107-1128. https://doi.org/10.1093/comjnl/bxaa148

[11] Saad MN, Muda Z, Ashaari NS, Hamid HA. Image segmentation for lung region in chest X-ray images using edge detection and morphology. In 2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014); 46-51. https://doi.org/10.1109/ICCSCE.2014.7072687

[12] Xu T, Mandal M, Long R, Cheng I, Basu A. An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. Computerized Medical Imaging and Graphics 2012; 36 (6): 452-463. https://doi.org/10.1016/j.compmedimag.2012.04.005

[13] Liu C, Zhao R, Pang M. Lung segmentation based on random forest and multi-scale edge detection. IET Image Processing 2019; 13 (10): 1745-1754. https://doi.org/10.1049/iet-ipr.2019.0130

[14] Ahmad WSHM, Zaki WMD, Ahmad Fauzi MF. Lung segmentation on standard and mobile chest radiographs using oriented Gaussian derivatives filter. Biomedical engineering online 2015; 14: 1-26. https://doi.org/10.1186/s12938-015-0014-8

[15] Sangamithraa PB, Govindaraju S. Lung tumour detection and classification using EK-Mean clustering. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) 2016; 2201-2206. https://doi.org/10.1109/WiSPNET.2016.7566533

[16] Leader JK, Zheng B, Rogers RM, Sciurba FC, Perez A et al. Automated lung segmentation in X-ray computed tomography: development and evaluation of a heuristic threshold-based scheme1. Academic radiology 2003; 10 (11): 1224-1236. https://doi.org/10.1016/S1076-6332(03)00380-5

[17] Annangi P, Thiruvenkadam S, Raja A, Xu H, Sun X et al. A region based active contour method for x-ray lung segmentation using prior shape and low level features. In 2010 IEEE international symposium on biomedical imaging: from nano to macro 2010; 892-895. https://doi.org/10.1109/ISBI.2010.5490130

[18] Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE transactions on medical imaging 2013; 33 (2): 577-590. https://doi.org/10.1109/TMI.2013.2290491

[19] Öztürk Ş, Çelik E, Çukur T. Content-based medical image retrieval with opponent class adaptive margin loss. Information Sciences 2023; 637: 118938. https://doi.org/10.1016/j.ins.2023.118938

[20] Öztürk Ş, Çukur T. Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets. IEEE Journal of Biomedical and Health Informatics 2022; 26 (9): 4679-4690. https://doi.org/10.1109/JBHI.2022.3187215

[21] Saidy L, Lee CC. Chest X-ray image segmentation using encoder-decoder convolutional network. In 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) 2018; 1-2. https://doi.org/10.1109/ICCE-China.2018.8448537

[22] Mittal A, Hooda R, Sofat S. LF-SegNet: A fully convolutional encoder–decoder network for segmenting lung fields from chest radiographs. Wireless Personal Communications 2018; 101: 511-529. https://doi.org/10.1007/s11277-018-5702-9

[23] Rashid R, Akram MU, Hassan T. Fully convolutional neural network for lungs segmentation from chest X-rays. In Image Analysis and Recognition: 15th International Conference, ICIAR 2018; 71-80. https://doi.org/10.1007/978-3-319-93000-8_9

[24] Hooda R, Mittal A, Sofat S. An efficient variant of fully-convolutional network for segmenting lung fields from chest radiographs. Wireless Personal Communications 2018; 101: 1559-1579. https://doi.org/10.1007/s11277-018-5777-3

[25] Nishio M, Fujimoto K, Togashi K. Lung segmentation on chest X-ray images in patients with severe abnormal findings using deep learning. International Journal of Imaging Systems and Technology 2021; 31 (2): 1002-1008. https://doi.org/10.1002/ima.22528

[26] Rajaraman S, Folio LR, Dimperio J, Alderson PO, Antani SK. Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations. Diagnostics 2021; 11 (4): 616. https://doi.org/10.3390/diagnostics11040616

[27] Abedalla A, Abdullah M, Al-Ayyoub M, Benkhelifa E. Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. PeerJ Computer Science 2021; 7: e607. https://doi.org/10.7717/peerj-cs.607

[28] Munawar F, Azmat S, Iqbal T, Grönlund C, Ali H. Segmentation of lungs in chest X-ray image using generative adversarial networks. Ieee Access 2020; 8: 153535-153545. https://doi.org/10.1109/ACCESS.2020.3017915

[29] Chen C, Dou Q, Chen H, Heng PA. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018; 143-151. https://doi.org/10.1007/978-3-030-00919-9_17

[30] Eslami M, Tabarestani S, Albarqouni S, Adeli E, Navab N et al. Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. IEEE transactions on medical imaging 2020; 39 (7): 2553-2565. https://doi.org/10.1109/TMI.2020.2974159

[31] Chen Y, Zhang H, Wang Y, Liu L, Wu QJ et al. TAE-Seg: Generalized Lung Segmentation via Tilewise AutoEncoder Enhanced Network. IEEE Transactions on Instrumentation and Measurement 2022; 71: 1-13. https://doi.org/10.1109/TIM.2022.3217870

[32] Xie Y, Wu Z, Han X, Wang H, Wu Y et al. Computer-aided system for the detection of multicategory pulmonary tuberculosis in radiographs. Journal of Healthcare Engineering 2020. https://doi.org/10.1155/2020/9205082

[33] Gaggion N, Mansilla L, Mosquera C, Milone DH, Ferrante E. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. IEEE Transactions on Medical Imaging 2023; 42 (2): 546 - 556. https://doi.org/10.1109/TMI.2022.3224660

[34] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017; 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[35] De Silva MS, Narayanan BN, Hardie RC. A Patient-Specific Algorithm for Lung Segmentation in Chest Radiographs. AI 2022; 3 (4): 931-947. https://doi.org/10.3390/ai3040055

[36] Arora R, Saini I, Sood N. Multi-label segmentation and detection of COVID-19 abnormalities from chest radiographs using deep learning 2021. Optik; 246: 167780. https://doi.org/10.1016/j.ijleo.2021.167780

[37] Arvind S, Tembhurne JV, Diwan T, Sahare P. Improvised light weight deep CNN based U-Net for the semantic segmentation of lungs from chest X-rays. Results in Engineering 2023; 17: 100929. https://doi.org/10.1016/j.rineng.2023.100929

[38] Ghali R, Akhloufi MA. ARSeg: An Attention RegSeg Architecture for CXR Lung Segmentation. In 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science; 291-296. https://doi.org/10.1109/IRI54793.2022.00068

[39] Zhang L, Liu A, Xiao J, Taylor P. Dual encoder fusion u-net (defu-net) for cross-manufacturer chest x-ray segmentation. In 2020 25th International Conference on Pattern Recognition, 9333-9339. https://doi.org/10.1109/ICPR48806.2021.9412718

[40] Cao F, Zhao H. Automatic lung segmentation algorithm on chest X-ray images based on fusion variational auto-encoder and three-terminal attention mechanism. Symmetry 2021; 13 (5): 814. https://doi.org/10.3390/sym13050814

[41] Dalmaz O, Yurt M, Çukur T. ResViT: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging 2022; 41 (10): 2598-2614. https://doi.org/10.1109/TMI.2022.3167808

[42] Chen J, Lu Y, Yu Q, Luo X, Adeli E et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint 2021; arXiv:2102.04306.

[43] Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint 2021; arXiv:2107.05274.

[44] Huang X, Chen J, Chen M, Chen L, Wan Y. TDD-UNet: Transformer with double decoder UNet for COVID-19 lesions segmentation. Computers in Biology and Medicine 2022; 151: 106306. https://doi.org/10.1016/j.compbiomed.2022.106306

[45] Yang J, Li C, Dai X, Gao J. Focal modulation networks. Advances in Neural Information Processing Systems 2022; 35: 4203-4217.

[46] Yang J, Li C, Zhang P, Dai X, Xiao B et al. Focal self-attention for local-global interactions in vision transformers. arXiv preprint 2021; arXiv:2107.00641.

[47] Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T et al. 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. American Journal of Roentgenology 2000; 174 (1): 71-74. https://doi.org/10.2214/ajr.174.1.1740071

[48] Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery 2014; 4 (6): 475. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[49] Öztürk Ş. Image inpainting based compact hash code learning using modified U-Net. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2020; 1-5. https://doi.org/10.1109/ISMSIT50672.2020.9255239

[50] Yahyatabar M, Jouvet P, Cheriet F. DenseUnet a light model for lung fields segmentation in Chest X-Ray images. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society; 1242-1245. https://doi.org/10.1109/EMBC44109.2020.9176033

[51] Liu W, Luo J, Yang Y, Wang W, Deng J et al. Automatic lung segmentation in chest X-ray images using improved U-Net. Scientific Reports 2022; 12 (1): 8649. https://doi.org/10.1038/s41598-022-12743-y

[52] Mansilla L, Milone DH, Ferrante E. Learning deformable registration of medical images with anatomical constraints. Neural Networks 2020; 124: 269-279. https://doi.org/10.1016/j.neunet.2020.01.023

[53] Özbey M, Dar SU, Bedel HA, Dalmaz O, Özturk Ş et al. Unsupervised medical image translation with adversarial diffusion models. arXiv preprint 2022; arXiv:2207.08208.

[54] Dar SU, Öztürk Ş, Korkmaz Y, Elmas G, Özbey M et al. Adaptive diffusion priors for accelerated mri reconstruction. arXiv preprint 2022; arXiv:2207.05876.