ORIGINAL ARTICLE **DOI: https://doi.org/10.18502/fem.v9i3.20020**

Detecting COVID-19-infected regions in lung CT scan through a novel dual-path Swin Transformer-based network

Zeinab Momeni pour[1], Ali Asghar Beheshti Shirazi[1]*

1. Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran.

***Corresponding author:** Seyed Ali asghar Beheshti Shirazi; **Email:** abeheshti@iust.ac.ir

Published online: 2025-09-14

Abstract: **Objective:** Deep learning-based automatic segmentation provides significant advantages over traditional manual segmentation methods in medical imaging. Current approaches for segmenting regions of Coronavirus disease 2019 (COVID-19) infections mainly utilize convolutional neural networks (CNNs), which are limited by their restricted receptive fields (RFs) and consequently struggle to establish global context connections. This limitation negatively impacts their performance in accurately detecting complex details and boundary patterns within medical images.

Methods: This study introduces a novel dual-path Swin Transformer-based network to address these limitations and enhance segmentation accuracy. Our proposed model extracts more informative 3D input patches to capture long-range dependencies and represents both large and small-scale features through a dual-branch encoder. Furthermore, it integrates features from the two paths via the new transformer interactive fusion (TIF) module. The architecture also incorporates an inductive bias by including a residual convolution (Res-conv) block within the encoder.

Results: The proposed network has been evaluated using a 5-fold cross-validation technique, alongside data augmentation, on the publicly available COVID-19-CT-Seg and MosMed datasets. The model achieved Dice coefficients of 0.872 and 0.713 for the COVID-19-CT-Seg and MosMed datasets, respectively, demonstrating its effectiveness relative to prior methodologies.

Conclusion: The significant improvements in segmentation accuracy, demonstrated by the achieved Dice coefficients on the COVID-19-CT-Seg and MosMed datasets, highlight the potential of our approach to enhance automated segmentation in medical imaging.

Keywords: COVID-19; Long-Range Information; Vision Transformer; Segmentation

Cite this article as: Momeni pour Z, Beheshti Shirazi AA. Detecting COVID-19-infected regions in lung CT scan through a novel dual-path Swin Transformer-based network. Front Emerg Med. 2025;9(3):e24.

1. Introduction

Coronavirus disease 2019 (COVID-19) is diagnosed using a combination of laboratory methods, clinical examinations, and imaging techniques, with physicians typically analyzing imaging results manually (1). Common imaging findings associated with COVID-19 include peripheral and bilateral ground-glass opacities, consolidative pulmonary opacities, localized vascular enlargement, and bronchiectasis (1-3). The manual analysis of these findings is time-consuming and may be linked to a shortage of human resources. Conversely, when it comes to diagnostic purposes, specialists may have significantly differing views regarding the same subject. Additionally, a specialist's diagnostic accuracy may vary based on several factors, including distraction and boredom, among others. Currently, artificial intelligence (AI) algorithms have become important clinical tools, improving and speeding up the diagnostic process. Experiences and insights acquired during the coronavirus epidemic have demonstrated their value in integrating AI algorithms for managing the disease and mitigating its impact.

Since building a large-scale, high-resolution, and precisely annotated dataset is a time-consuming and costly process that also relies on medical expertise, the majority of studies have evaluated their models using two widely used public datasets, namely COVID-19-CT-Seg (4) and MosMed (5), which include 20 and 50 images, respectively. A predominant number of these studies have developed U-Net-based CNNs for the segmentation of COVID-19-infected areas (6-8).

Furthermore, several recent studies have attempted to improve segmentation accuracy by adding modules into their networks. Singh et al. (9), for example, introduced Lung-INFseg, a new segmentation model that makes use of slice-based input and receptive field-aware modules. This approach had a Dice coefficient of 0.803 on the COVID-19-CT-Seg dataset. Zheng et al. (10) introduced the 3D CU-Net architecture. It leveraged data augmentation techniques, a pyramid fusion module, and the Tversky loss function, re-

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

sulting in Dice scores of 0.778 and 0.668 on the COVID-19-CT-Seg and MosMed datasets, respectively.

Owais et al. (11) proposed a convolutional network named meta-domain adaptive segmentation network (MDA-SN), which is based on MobileNetV2. The model was trained separately on the COVID-19-CT-Seg and MosMed datasets and used grouped and multi-scale dilated convolution layers, a residual attention mechanism, and an adaptive data normalization module. In this work, cross-dataset evaluations were conducted to assess generalization capability, leading to an average Dice coefficient of 0.759 for the two datasets.

Given that an increase in the number of convolutional layers in a CNN model can result in the loss of critical details, studies (12,13) have shown that transformers present a viable alternative for CNNs in the structure of models. Geng et al. (12) developed an encoder-decoder model called STC-Net, in which convolutional blocks were used for extracting local features, and the ReSwin Transformer blocks were applied to capture long-range information. This model was introduced to improve the accuracy of local feature extraction and the global context, while maintaining a balance between computational efficiency and performance accuracy. In the study (13), a hierarchical agent transformer network (HATNet) was introduced by Tian et al. (13). This model is based on agent transformer blocks. These blocks are designed to extract features with high precision and use a non-local attention mechanism to model global features with linear complexity. To prevent the degradation of features during the processes of extraction and fusion, the modules, namely diversity restoration and full-scale bidirectional feature pyramid network, were used in sequence. Lastly, in studies (12,13) where STCNet and HATNet models were designed, Dice scores of 0.799 and 0.841 were reported on the small COVID-19-CT-Seg dataset, respectively.

Recently, transformer-based architectures, primarily designed for sequence-to-sequence modeling in natural language processing, have sparked considerable interest within the computer vision (CV) community. The multi-head self-attention (MSA) mechanism inherent in these models facilitates the effective establishment of global relationships among the sequence tokens and increases the model's capacity to capture long-range dependencies, particularly for pixel-based tasks in computer vision (14,15). Accordingly, we designed a 3D U-shaped architecture that can extract more optimal local and global features by leveraging the strengths of CNNs and the Swin Transformer blocks.

2. Methods

2.1. Study design and setting

Multi-scale feature representation is critical for optimizing vision transformers in medical image segmentation (14,16-19). This paper describes a novel Swin Transformer-based U-shaped encoder-decoder architecture. This architecture partitions images into non-overlapping patches at two resolutions: large and small. A dual-scale encoder processes patches to extract complementary features at various semantic levels. To facilitate the effective integration of multi-scale representations, we propose a transformer interactive f usion (TIF) module that allows for cross-scale feature interaction. The proposed framework significantly improves the conventional U-shaped architecture by leveraging the Swin Transformer's hierarchical modeling capabilities and the strengths of multi-scale processing, offering enhanced performance for medical image segmentation tasks.

Notably, the proposed model operates in three dimensions and is evaluated through the application of data augmentation techniques alongside a 5-fold cross-validation methodology on the COVID-19-CT-Seg (4) and MosMed (5) public datasets.

2.2. Dataset of COVID-19-CT-Seg

The COVID-19-CT-Seg dataset comprises 20 lung CT scan images, accompanied by three benchmark conditions. It includes a subset of patients diagnosed with COVID-19, where binary pixel masks delineating the infection regions are provided. The CT scans in this dataset possess dimensions of 512×512 and 630×630, comprising multiple slices (4,6). This dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike International (CC BY-NC-SA) license (20). Also, this dataset is a result of a collaborative effort between the Coronacases Initiative and Radiopaedia's open data initiative (4).

2.3. Dataset of MosMed

The MosMed dataset encompasses a total of 1,110 CT scan images from subjects, of which 42% are male, 56% are female, and the gender of the remaining 2% remains unidentified (21). Importantly, a limited subset of this dataset, comprising 50 out of 1,110 individuals, has been subjected to expert annotation (5,7). In the annotation process for each image, ground glass opacities and consolidation regions were identified as positive pixels on the corresponding mask. The masks obtained have been supplied in NIfTI format and converted into Gzip archives (22). Also, this dataset is governed by the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License (5,22).

2.4. Data augmentation techniques

The following volumetric data augmentation techniques were deployed to boost the robustness and generalizability of our model:

Random cropping: To ensure informative training samples, sub-volumes were randomly selected from the original volume, each including at least a portion of the infected area.

Rotation: Volumes were randomly rotated around each axis within ±15 degrees. This helped the model become invariant to minor orientation changes in lung position.

Shifting: Random translations were applied along each axis

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

with a shift limit of ±5% of the volume size.

Scaling: 3D volumes were randomly resized by a factor of 0.6 to 0.8, making the model more robust to changes in the size of infected areas.

Intensity adjustment: Brightness adjustment involved scaling the voxel intensities across the 3D volume by a constant factor, which is typically between [0.9 and 1.1].

2.5. Swin Transformer-based dual multiscale architecture

Figure 1 illustrates the overall structure of the proposed network. This architecture is grounded in a two-path encoder. The input is processed through two parallel patch embedding layers. The first branch, with a patch size of 4×4×4, produces an output of shape H/4× W/4×D/4×48. The second branch, with a patch size of 2×2×2, generates a feature map of shape H/2× W/2×D/2×48. Each encoder branch passes through a Res-conv Block to extract local features, followed by a Swin Transformer to capture global dependencies. At the beginning of each stage, the combination of the Res-conv block with the Swin Transformer block plays a vital role in image processing and the hierarchical analysis of features.

The shortcut connection within the Res-conv block facilitates the network's ability to learn identity mappings alongside other transformations. Moreover, the shortcut connection addresses critical challenges, including the vanishing gradient problem, complications associated with the training process, and the necessity for extensive datasets (23).

The output generated from the Swin Transformer is subsequently passed to the merge layer at each stage. This layer connects neighboring patches to preserve spatial information and fine-grained details of the images. Furthermore, this layer reduces the dimensionality of the combined features by half (14). As such, it decreases the number of patches available for processing by the following Swin Transformer block, thereby improving the network's capacity to learn long-range dependencies more efficiently.

Within the architecture presented, the TIF module is applied at each stage to fuse multi-scale features and enable efficient interactions between them. The features extracted at each encoder stage are integrated into the CNN-based decoder through skip connections, as illustrated in figure 1. The decoder part of this architecture resembles the decoder component of the SwinUNETR (24). In the bottleneck section, the resolution of the feature map is increased via a deconvolution layer, which is subsequently linked to the feature map from the preceding stage. This process is similarly applied to the other layers. Ultimately, the final output is subjected to a convolution layer with a kernel size of 1×1×1, resulting in the generation of a segmented image.

2.6. Swin Transformer block

This block comprises two sub-blocks (Figure 2). The block's architecture includes a normalization layer, a window-based MSA mechanism, and a multi-layer perceptron (MLP). In this context, the notation W-MSA refers to conventional window-based MSA modules, while SW-MSA denotes shifted window-based MSA modules (24). The default window size is set to 7. In each sub-block, the output from the window-based module is transferred to the normalization layer. This layer stabilizes the output within an optimal range, contributes to a more stable training process (25). Subsequently, an MLP is applied to this output. The residual connection within the Swin Transformer facilitates the model's learning of richer feature representations by appending the processed features to the original input. According to figure 2, the initial sub-block output (l) in the Swin Transformer block is obtained as follows:

$$\hat{Z}^{l}=\text{W-MSA}(\text{LayerNorm}(Z^{l-1}))+Z^{l-1} \quad (1)$$

$$Z^{l}=\text{MLP}(\text{LayerNorm}(\hat{Z}^{l}))+\hat{Z}^{l}, \quad (2)$$

The W-MSA module employs multiple parallel self-attention (SA) mechanisms, each addressing distinct aspects of the input data. The W-MSA module generates attention maps by integrating intuitions from various dimensions, thereby facilitating interpretability regarding the model's focus during the segmentation task. In the initial sub-block, SA is applied to the non-overlapping windows of the patches, with the SA scores calculated among the patches within each window. Consequently, the computational scope of SA is confined to local windows rather than encompassing all image patches.

The second sub-block introduces SA calculations within shifted partitions (26). The second sub-block output (l+1) is derived as follows (Figure 2):

$$\hat{Z}^{l+1}=\text{SW-MSA}(\text{LayerNorm}(Z^{l}))+Z^{l} \quad (3)$$

$$Z^{l+1}=\text{MLP}(\text{LayerNorm}(\hat{Z}^{l+1}))+\hat{Z}^{l+1}, \quad (4)$$

The second sub-block enhances the model's capability to comprehend long-range relationships and contextual information within the images. Specifically, SW-MSA computes attention across adjacent windows by cyclically shifting the windows, thereby enabling the model to capture broader contextual data. This methodology employs an efficient batch processing technique by cyclically shifting windows to the upper left (14). This approach creates batch windows that contain multiple smaller non-adjacent sub-windows in the feature map. Despite being spread out, the total number of batch windows remains constant, similar to the standard window partitioning method, where each batch window has a fixed size. This consistency ensures efficient image processing. In essence, the method ensures that the processing mechanism remains regular while allowing for more flexible interactions between regions. In addition, both Window-based Multi-head Self-Attention (W-MSA) and Shifted Window-based Multi-head Self-Attention (SW-MSA) incorporate relative position bias. This enables the model to process the features of each token, while capturing their relative spatial positions within a window. Rather than relying on fixed absolute coordinates, the model learns token relationships, enhancing its ability to capture spatial structure and dependencies.

Figure 3 presents the attention map generated by the model

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

during its decision-making process. This map highlights regions in the input image that receive the highest focus during prediction, thereby improving the interpretation of model behavior.

2.7. Transformer interactive fusion module (TIF)

The TIF module is applied at each stage to fuse multi-scale features. Figure 4 illustrates the structure of this module.

This module initially receives two features from two branches. To clarify, for the outputs of the two branches at the same phase i (where i=1,...,5), we denote the outputs of the prime branch as $F^i=[f^i{}_1,f^i{}_2,\ldots,f^i{}_{h\times w\times d}]\in \mathbb{R}^{C\times(h\times w\times d)}$ and the outputs of the complementary branch as $G^i=[g^i{}_1,g^i{}_2,\ldots,g^i{}_{(h/2\times w/2\times d/2)}]\in\mathbb{R}^{C\times(h/2\times w/2\times d/2)}$. Subsequently, the transformation output G^i is obtained through the following operation:

$$\hat{g}^i=\text{Flatten (Avgpool}(G^i)) \quad (5)$$

where $\hat{g}^i$ represents the global abstracted data derived from G^i, facilitating its interaction with F^i at the pixel level. The concatenation of F^i and $\hat{g}^i$ results in a sequence of 1+h×w×d tokens, which are subsequently input into the transformer layer to compute the SA representation:

$$\hat{F}^i=\text{Transformer}([\hat{g}^i,f^i{}_1,f^i{}_2,\ldots,f^i{}_{h\times w\times d}])=[\hat{f}^i{}_0,\hat{f}^i{}_1,\hat{f}^i{}_2,\ldots,\hat{f}_{(h\times w\times d)}]\in\mathbb{R}^{C\times(1+h\times w\times d)}$$
$$F^i{}_{out}=[\hat{f}^i{}_1,\hat{f}^i{}_2,\ldots,\hat{f}^i{}_{h\times w\times d}]\in\mathbb{R}^{C\times(h\times w\times d)} \quad (6)$$

$Where F^i{}_{out}$ represents the outcome of the small-scale branch within the TIF module. This approach establishes connections between each token in F^i and the entirety of G^i, thereby enabling small-scale features to access high-level contextual information from the larger-scale branches. Finally, the output from each transformer is processed by the convolution layer following a series of reshaping, upsampling operations, and concatenation of the feature matrices (Figure 4).

2.8. Objectives

The primary objective of this study was to assess the performance and effectiveness of current methodologies used for segmenting areas impacted by COVID-19. In particular, we aimed to analyze various segmentation techniques, comparing their performance, reliability, and computational efficiency.

2.9. Statistical analysis

We implemented the proposed model in Python and trained using an NVIDIA 1080 Ti graphics card. The proposed model was evaluated using the COVID-19-CT-Seg and the MosMed datasets. The evaluation process included 5-fold cross-validation and data augmentation techniques, with the model subjected to training for 500 epochs utilizing the DiceCELoss loss function, and a batch size of 1. The AdamW optimizer was utilized to optimize the parameters of the proposed model, with an initial learning rate of $1e^{-4}$. A cosine decay learning rate scheduler with linear warm-up was employed to adjust the learning rate during training, where the warm-up phase spanned the first 50 epochs.

To assess segmentation performance, we employed common metrics such as the Dice coefficient, sensitivity, and specificity, which collectively measure the effectiveness of our model in detecting COVID-19 lesions, indicating the proximity of predicted outputs to actual labels.

Specifically, the Dice coefficient reflects the degree of overlap between the model's segmented region and the target mask, with values ranging from zero (indicating no overlap) to one (indicating complete overlap) (27).

3. Results

In this work, we first evaluated the performance of UNETR and SwinUNETR models on the COVID-19-CT-Seg and MosMed datasets for the infection segmentation task. On the COVID-19-CT-Seg dataset, UNETR achieved a Dice coefficient of 0.852, while SwinUNETR demonstrated superior segmentation accuracy with a Dice coefficient of 0.866. Similarly, the Dice values obtained on the MosMed dataset were 0.69 for UNETR and 0.706 for SwinUNETR. Due to the superior performance of SwinUNETR compared to UNETR, we chose the SwinUNETR architecture as the basis for our model.

We subsequently optimized the SwinUNETR and developed a dual-path multi-scale network aimed at detecting COVID-19 lesions. In addition to segmenting COVID-19 infectious regions in the two datasets, we also segmented lung regions within the COVID-19-CT-Seg dataset. Tables 1 and 2 present the evaluation results of our approach.

4. Discussion

In situations characterized by disease outbreaks and a shortage of human resources, deploying deep learning-based automatic algorithms, known for their rapid processing and high generalizability, can improve diagnostic efficiency significantly. As with other diseases, deep learning algorithms are currently employed to study COVID-19. Most existing studies use U-Net-based CNNs for COVID-19 lesion segmentation, with these networks extracting features via convolutional layers. For example, in studies (6-8) that used CNN models, Dice scores of 0.761, 0.673, and 0.704 were reported on a dataset of 20 CT images.

A significant limitation of the U-Net and similar architectures is their insufficient ability to learn long-range dependencies. These networks primarily depend on local RFs, which highlight nearby spatial information. As a result, the features provided to the decoder of these networks are often deficient in spatial resolution. Furthermore, many of these networks do not effectively distinguish between background and foreground pixels due to the application of uniform filters across the input data. As a result, such models may struggle to optimally segment lesions of varying scales. In summary, CNN-based models demonstrate constraints in their capacity to extract global features, learn long-range dependencies, and

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

preserve high-resolution features. The limitations of CNNs hinder their ability to diagnose border patterns of COVID-19-related lesions accurately and to differentiate these from other respiratory diseases, despite the critical importance of boundary pattern detection in medical imaging.

Recent studies (12,13) have shown that transformer-based architectures can significantly improve feature extraction performance and reduce the amount of challenges in the segmentation task. Since the model's ability to model spatial dependencies between slices is critical for CT volumetric segmentation, approach (13) performed better than approach (12). In this study, leveraging this insight, we introduce our innovative strategies within a 3D framework to achieve higher diagnostic accuracy than recent approaches.

In the proposed model, Swin Transformer blocks are used for feature extraction, producing attention maps that improve the model's focus on pertinent regions of the image. The SA mechanisms within the Swin Transformer blocks enable the adaptation of receptive field sizes across various regions of the image. This flexibility significantly improves our model's ability to identify COVID-19 lesions of various shapes and sizes. Additionally, as the proposed model encompasses two paths, it can facilitate large-scale and small-scale information extraction, allowing for the effective capture of features within image patches and pixel-level information. We designed a transformer block-based TIF module to combine features, striking a balance between pixel-level characteristics and overall contextual information. Furthermore, the proposed model integrates an inductive bias by implementing the Res-conv block prior to the Swin Transformer block. The inductive bias refers to the assumptions that guide the algorithm's predictions on unseen data, allowing it to prioritize specific solutions regardless of the data it has previously encountered (28). Given the importance of local details in medical imaging, we focused on the decoder component of the model based on convolutional layers. The proposed model has been evaluated on two publicly datasets, namely COVID-19-CT-Seg (4) and MosMed (5). On the COVID-19-CT-Seg dataset, the proposed model increased the Dice score to 0.872 from 0.841, the highest value in earlier research by Tian et al. (13).

Therefore, our model performs more effectively than recent research due to the global modeling capability of Swin Transformer blocks and the use of multi-resolution and dual-path strategies.

Table 3 depicts the trade-offs between the complexity of our models (in terms of the number of parameters and FLOPs) and their inference efficiency on the COVID-19-CT-Seg dataset. This table illustrates that the dual-path Swin Transformer-based model, despite its fewer parameters than UNETR, requires a high number of FLOPs and exhibits the longest inference time. Nevertheless, this model outperforms UNETR and SwinUNETR in segmentation, despite the higher computational cost.

The designed model was once again evaluated on a different dataset known as MosMed. In this instance, the Dice value for this dataset increased by 4.5% relative to a prior work conducted by Zheng et al. (10). Thus, the results obtained exhibit the proposed model's ability to discern patterns and boundaries as well as accurately identify target regions (positive pixels) over earlier approaches.

In the task of COVID-19 segmentation using CT images, one of the main challenges is the severe class imbalance infected regions are typically much smaller than the healthy areas or background. As seen in Figure 5, infected regions (positive pixels) are far outnumbered by non-infected regions (negative pixels). It reflects a class imbalance between positive and negative pixels, with the predominance of negative samples during training resulting in comparatively higher accuracy in detecting negative pixels than positive ones. Consequently, as observed in Tables 1 and 2, the specificity values are higher than the sensitivity values in identifying infected regions.

Known as a combination of Dice Loss and Cross-Entropy Loss functions, the DiceCELoss loss function helped our model better segment small and large regions simultaneously. Dice Loss enables the model to appropriately identify smaller classes, e.g., the infected areas, by emphasizing the overlap between predictions and ground truth. In contrast, Cross-Entropy Loss enables the model to accurately predict dominant classes, e.g., the background, by minimizing differences between predicted and actual class probabilities. In other words, Dice Loss is sensitive to small classes, whereas Cross-Entropy Loss is more suitable for large classes. Combining the two acts similarly to class weighting helped balance the contribution of each class during training. In the presented work, regions of interest (ROIs) were randomly sampled from the input 3D images, with a focus on areas affected by infection. This sampling strategy was aimed at addressing the issue of class imbalance by broadening the representation of minority classes (i.e., infected areas) during training. By ensuring that the model encompasses more examples of these underrepresented regions, it becomes equipped further to learn their features and enhance segmentation performance. Following this targeted patch sampling, we employed data augmentation techniques on the selected patches. This subject not only helped avoid overfitting but also improved the model's generalizability. The integrated approach of focused sampling and augmentation leads to more accurate and robust detection of infection regions in 3D medical images.

In addition to segmenting COVID-19 infectious regions in the two datasets, we also segmented lung regions within the COVID-19-CT-Seg dataset, achieving a Dice score of 0.977. As indicated in table 1, the Dice score associated with the lung mask exceeds that of the infection mask for several reasons:

1. The boundaries of the lungs exhibit relatively stable descriptive features. The morphology, positioning, and anatomical structure of the lungs remain consistent, whereas COVID-19 lesions display significant variability in shape, size, and localization (29).

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

2. The delineation of lung boundaries in CT scan images is particularly distinct owing to the tissue-air interface. The pronounced contrast between air-filled lung tissue and surrounding anatomical structures plays a critical role in enhancing segmentation outcomes (30). In contrast, the margins of lesions may intermingle with healthy lung tissue or adjacent structures, complicating the segmentation process.

3. Lung tissue generally exhibits a relatively uniform intensity profile, whereas lesions manifest diverse intensity patterns. This disparity creates challenges in distinguishing between healthy lung regions and areas affected by COVID-19 (29).

The main challenge lies in accurately segmenting the COVID-19 infection regions to ensure the precise detection of positive pixel classifications. Consequently, a higher occurrence of false negatives is observed in the infection segmentation compared to lung segmentation.

The designed network can serve as a versatile architecture and deliver a unified solution for diverse objectives, while traditional CNNs, typically require specialized architectures tailored to specific tasks.

5. Limitations

Although the proposed approach demonstrates promising performance, there are several limitations that warrant attention in future research. The reliance on extensive, COVID-19-related annotated 3D medical imaging datasets poses a considerable challenge. The process of collecting and manually labeling such data is inherently labor-intensive and costly, which often restricts the diversity and volume of training data that is available. Major concern is the risk of overfitting, given the small sample size and the complex, variable nature of COVID-19-related imaging patterns. To mitigate overfitting, more practical and actionable strategies can be used in future research. Beyond conventional techniques (e.g., rotation, scaling, and intensity variations), the integration of other methods can be considered for data augmentation, including:

• Elastic deformation, which simulates natural anatomical variability by applying nonlinear spatial distortions to the tissue structures;

• CutMix and MixUp, which generate hybrid samples by combining patches from different images, thereby increasing structural diversity and regularizing the model;

• Transformers-based augmentation, which leverages attention mechanisms to apply more context-aware and structurally informed transformations to 3D volumes.

For synthetic data generation, investigating the use of Conditional Generative Adversarial Networks (cGANs) and Denoising Diffusion Probabilistic Models (DDPMs) can also be considered an aim of future studies. However, a multi-step validation framework is proposed to ensure the clinical plausibility and utility of the generated samples:

• Employing quantitative similarity metrics such as the Fréchet Inception Distance (FID) and the Structural Similarity Index Measure (SSIM) to evaluate the fidelity of synthetic images;

• Conducting expert reviews by trained radiologists to evaluate the pathological realism of generated infection patterns.

Ultimately, the performance of the proposed framework can be enhanced through the targeted methodological advancements. The enhancement will lead to the improved clinical utility, scalability, and generalization of the framework across broader medical imaging tasks.

6. Conclusion

Given the critical importance of boundary patterns and detailed local features in medical imaging, the proposed network effectively harnesses high-resolution features through its dual-path encoder. This design fosters a balance between local and global features, aided by the TIF module. In this approach, data augmentation techniques and Res-conv blocks implemented in the encoder contribute to optimizing the training process, thereby enhancing model performance even with limited datasets. By analyzing the results obtained and comparing them with findings from previous studies, we anticipate that our model will yield substantial advancements in the segmentation of medical images.

7. Declarations

7.1. Acknowledgement

The authors have no acknowledgments to disclose.

7.2. Authors' contribution

ABS and ZM contributed in the conceptualization and methodology of the study, with ABS overseeing the entire process. ZM took the lead in writing the original draft, while both authors collaborated in the writing, review, and editing phases. Furthermore, ABS handled the project administration.

7.3. Conflict of interest

The authors declare that there are no conflicts of interest associated with this article. The research was carried out in an objective manner, and no financial or personal relationships have impacted the findings or interpretations presented.

7.4. Funding

This research project was carried out without any external funding or financial assistance.

7.5. Data availability

The authors affirm that the data from the study are accessible and will be provided upon request.

7.6. Using artificial intelligence chatbots

The content of this article was not produced with the assistance of any artificial intelligence chatbot.

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

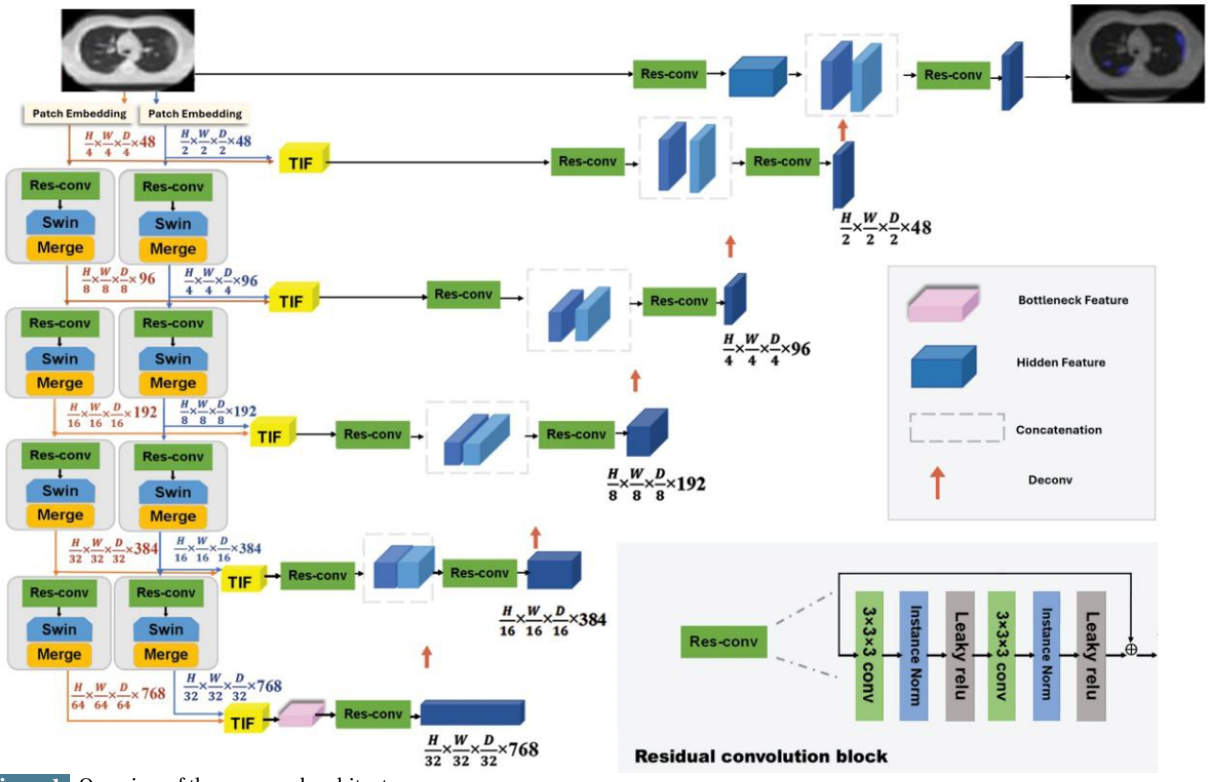
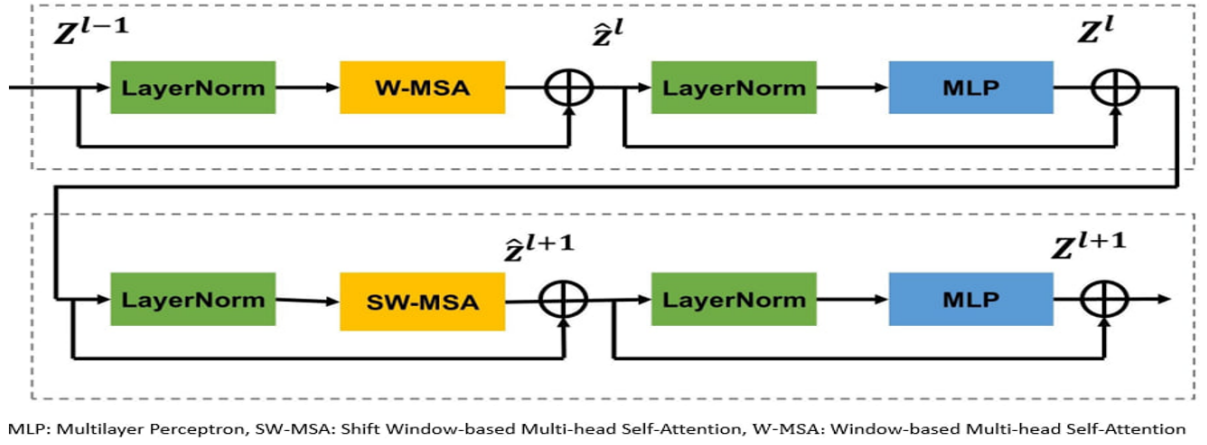


Figure 1 Overview of the proposed architecture



MLP: Multilayer Perceptron, SW-MSA: Shift Window-based Multi-head Self-Attention, W-MSA: Window-based Multi-head Self-Attention

Figure 2 The structure of the Swin Transformer block

Table 1 Performance results of the proposed model on the COVID-19-CT-Seg dataset

Metric	Fold1	Fold2	Fold3	Fold4	Fold5	AVG
Infection mask						
Dice	0.871	0. 882	0.864	0.869	0.875	0.872
Sensitivity	0.880	0.874	0.909	0.873	0.902	0.887
Specificity	0.938	0.973	0.982	0.982	0.971	0.969
Lung mask						
Dice	0.972	0.979	0.972	0.983	0.982	0.977
Sensitivity	0.985	0.987	0.985	0.991	0.990	0.988
Specificity	0.994	0.998	0.997	0.999	0.999	0.997

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

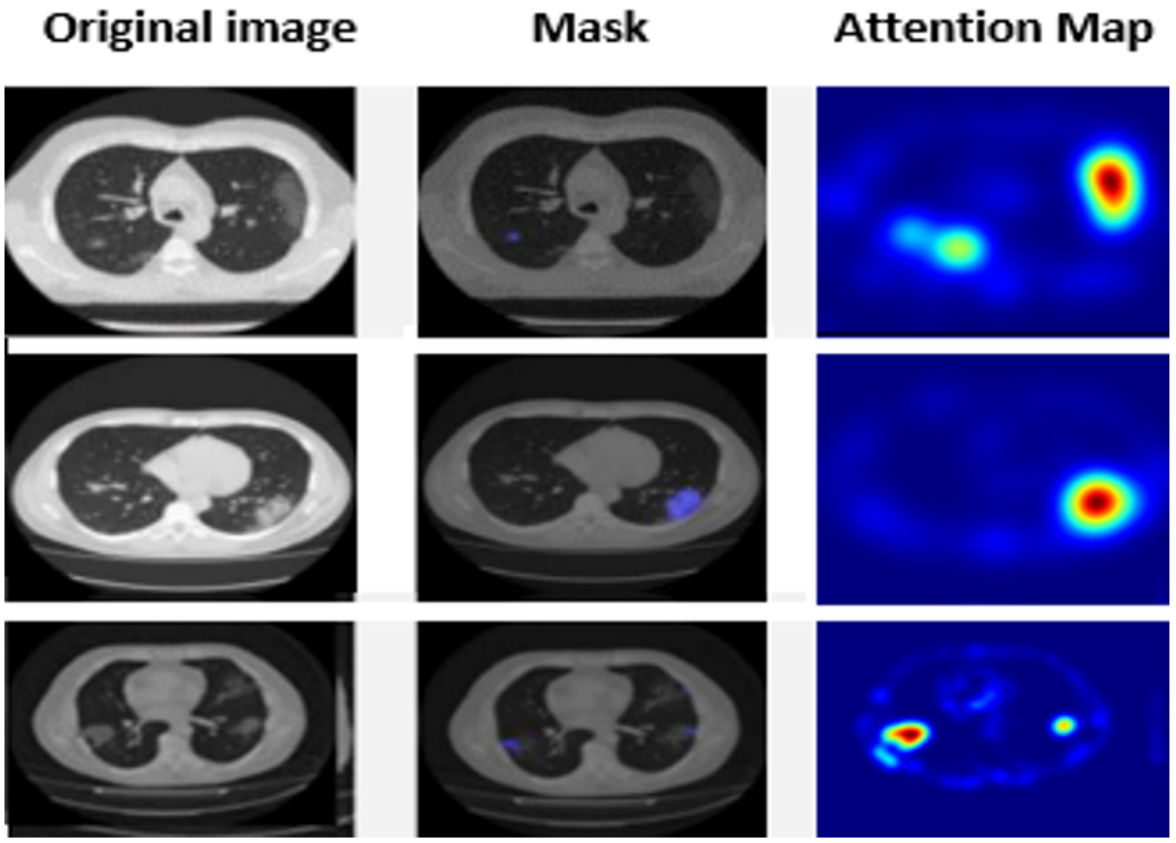


Figure 3 Attention map visualization

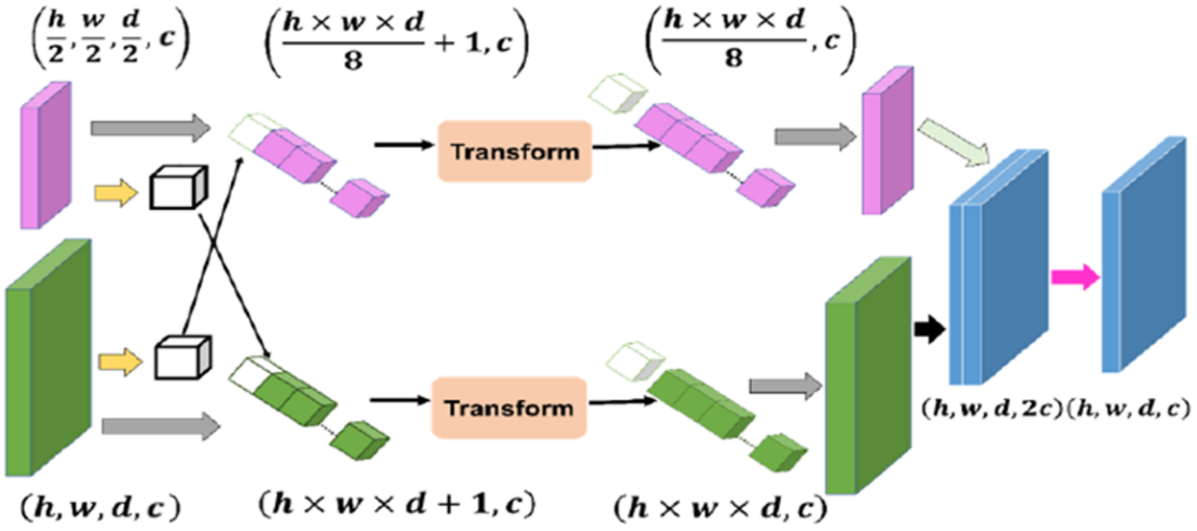


Figure 4 Depicts overview of the TIF module

Table 2 Performance results of the proposed model on MosMed dataset

			Infection mask			
Metric	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Fold5**	**AVG**
Dice	0.732	0.682	0.672	0.705	0.772	0.713
Sensitivity	0.785	0.724	0.716	0.730	0.841	0.759
Specificity	0.996	0.996	0.999	0.995	0.999	0.997

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

Table 3 Comparison of the number of parameters, FLOPs, and averaged inference time of our models in the COVID-19-CT-Seg experiments

Model	#Params (M)	FLOPs (G)	Inference Time (s)
UNETR	94.28	42.11	12.08
SwinUNETR	33.7	67.403	14.23
Dual-path Swin Transformer-based model	61.02	124.23	20.45

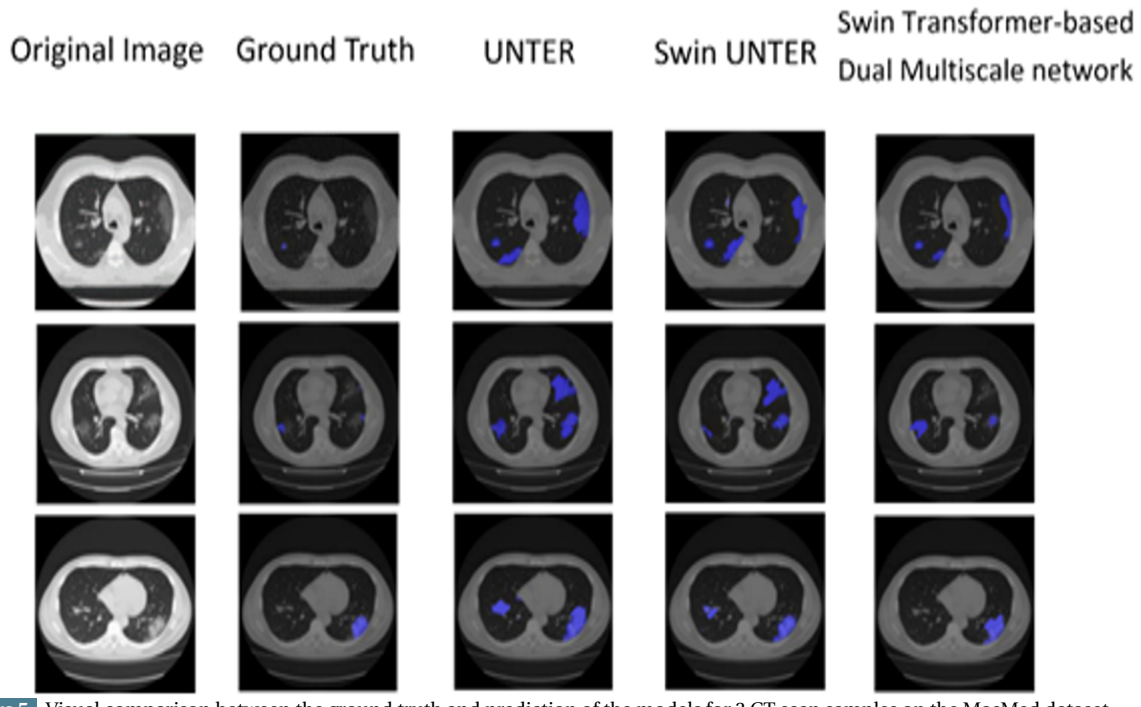


Figure 5 Visual comparison between the ground truth and prediction of the models for 3 CT scan samples on the MosMed dataset

References

1. Filchakova O, Dossym D, Ilyas A, Kuanysheva T, Abdizhamil A, Bukasov R. Review of COVID-19 testing and diagnostic methods. Talanta. 2022;244:123409.
2. Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology. 2020;295(3):685-91.
3. Zhao W, Zhong Z, Xie X, Yu Q, Liu J. Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study. AJR Am J Roentgenol. 2020;214(5):1072-7.
4. Ma J, Ge Y, Wang Y, et al. COVID-19 CT Lung and Infection Segmentation Dataset. Zenodo. Published April 20, 2020.
5. Morozov SP, Andreychenko AE, Blokhin IA, Gelezhe PB, Gonchar AP, Nikolaev AE, et al. MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic. Digital Diagnostics. 2020;1(1):49-59.
6. Müller D, Rey IS, Kramer F. Automated chest ct image segmentation of covid-19 lung infection based on 3d u-net. arXiv preprint arXiv:200704774. 2020.
7. Ma J, Wang Y, An X, Ge C, Yu Z, Chen J, et al. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. Medical physics. 2021;48(3):1197-210.
8. Wang Y, Zhang Y, Liu Y, Tian J, Zhong C, Shi Z, et al. Does non-COVID-19 lung lesion help? investigating transferability in COVID-19 CT image segmentation. Comput Methods Programs Biomed. 2021;202:106004.
9. Kumar Singh V, Abdel-Nasser M, Pandey N, Puig D. Lung-infseg: Segmenting covid-19 infected regions in lung ct images based on a receptive-field-aware deep learning framework. Diagnostics. 2021;11(2):158.
10. Zheng R, Zheng Y, Dong-Ye C. Improved 3D U-Net for COVID-19 chest CT image segmentation. Scientific Programming. 2021;2021(1):9999368.
11. Owais M, Hassan T, Afzal N, Khan SH, Velayudhan D, Ganapathi II, et al. Meta-domain adaptive framework for efficient diagnostic assessment of lung infection using CT radiographs. 2024.
12. Geng P, Tan Z, Wang Y, Jia W, Zhang Y, Yan H. STCNet: Alternating CNN and improved transformer network for COVID-19 CT image segmentation. Biomedical Signal Processing and Control. 2024;93:106205.
13. Tian Y, Mao Q, Wang W, Zhang Y. Hierarchical agent transformer network for COVID-19 infection segmentation. Biomedical Physics Engineering Express.

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

2025;11(2):025055.
14. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Transactions on Instrumentation and Measurement. 2022;71:1-15.
15. Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution Swin transformer for automatic medical image segmentation. Sensors. 2023;23(7):3420.
16. Cai Z, Fan Q, Feris RS, Vasconcelos N, editors. A unified multi-scale deep convolutional neural network for fast object detection. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14; 2016: Springer.
17. Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L, editors. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020.
18. Nah S, Hyun Kim T, Mu Lee K, editors. Deep multi-scale convolutional neural network for dynamic scene deblurring. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
19. Chen C-F, Fan Q, Mallinar N, Sercu T, Feris R. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. arXiv preprint arXiv:180703848. 2018.
20. Zhang Q, Ren X, Wei B. Segmentation of infected region in CT images of COVID-19 patients based on QC-HC U-net. Scientific Reports. 2021;11(1):22854.
21. Ibrahim MR, Youssef SM, Fathalla KM. Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment. J Ambient Intell Humaniz Comput. 2023;14(5):5665-88.
22. Morozov SP, Andreychenko AE, Pavlov N, Vladzymyrskyy A, Ledikhova N, Gombolevskiy V, et al. Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:200506465. 2020.
23. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
24. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, et al., editors. Self-supervised pre-training of swin transformers for 3d medical image analysis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022.
25. Ba JL. Layer normalization. arXiv preprint arXiv:160706450. 2016.
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision; 2021.
27. Li C-F, Xu Y-D, Ding X-H, Zhao J-J, Du R-Q, Wu L-Z, et al. MultiR-net: a novel joint learning network for COVID-19 segmentation and classification. Computers in Biology and Medicine. 2022;144:105340.
28. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:180601261. 2018.
29. Oda M, Hayashi Y, Otake Y, Hashimoto M, Akashi T, Mori K, editors. Lung infection and normal region segmentation from CT volumes of COVID-19 cases. Medical Imaging 2021: Computer-Aided Diagnosis; 2021: SPIE.
30. Sun S, Bauer C, Beichel R, editors. Robust active shape model based lung segmentation in CT scans. Fourth International Workshop on Pulmonary Image Analysis; 2011.

Copyright © 2025 Tehran University of Medical Sciences
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org /licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.