Archives of Academic Emergency Medicine. 2025; 13(1): e21

OPEN ACCESS

ORIGINAL RESEARCH

Identifying COVID-19-Infected Segments in Lung CT Scan Through Two Innovative Artificial Intelligence-Based Transformer Models

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

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

Received: September 2024; Accepted: November 2024; Published online: 16 December 2024

Abstract: **Introduction:** Automatic systems based on Artificial intelligence (AI) algorithms have made significant advancements across various domains, most notably in the field of medicine. This study introduces a novel approach for identifying COVID-19-infected regions in lung computed tomography (CT) scan through the development of two innovative models. **Methods:** In this study we used the Squeeze and Excitation based UNet TRansformers (SE-UNETR) and the Squeeze and Excitation based High-Quality Resolution Swin Transformer Network (SE-HQRSTNet), to develop two three-dimensional segmentation networks for identifying COVID-19-infected regions in lung CT scan. The SE-UNETR model is structured as a 3D UNet architecture with an encoder component built on Vision Transformers (ViTs). This model processes 3D patches directly as input and learns sequential representations of the volumetric data. The encoder connects to the decoder using skip connections, ultimately producing the final semantic segmentation output. Conversely, the SE-HQRSTNet model incorporates High-Resolution Networks (HRNet), Swin Transformer modules, and Squeeze and Excitation (SE) blocks. This architecture is designed to generate features at multiple resolutions, utilizing Multi-Resolution Feature Fusion (MRFF) blocks to effectively integrate semantic features across various scales. The proposed networks were evaluated using a 5-fold cross-validation methodology, along with data augmentation techniques, applied to the COVID-19-CT-Seg and MosMed datasets. **Results:** Our experimental results demonstrate that the Dice value for the infection masks within the COVID-19-CT-Seg dataset improved by 3.81% and 4.84% with the SE-UNETR and SE-HQRSTNet models, respectively, compared to previously reported work. Furthermore, the Dice value for the MosMed dataset increased from 66.8% to 69.35% and 70.89% for the SE-UNETR and SE-HQRSTNet models, respectively. **Conclusion:** These improvements indicate that the proposed models exhibit superior efficiency and performance relative to existing methodologies.

Keywords: COVID-19; Segmentation; Self-Attention; Squeeze and Excitation; Swin Transformer; Vision Transformer

Cite this article as: Momeni Pour Z, Beheshti Shirazi AA. Identifying COVID-19-Infected Segments in Lung CT Scan Through Two Innovative Artificial Intelligence-Based Transformer Models. Arch Acad Emerg Med. 2025; 13(1): e21. https://doi.org/10.22037/aaemj.v13i1.2515.

1. Introduction

The emergence of coronavirus disease or COVID-19 marked a pivotal moment in the history of global health systems, presenting numerous challenges across medical sciences, public health, and societal frameworks. The rapid and accurate identification of COVID-19-infected regions in computed tomography (CT) scan images is crucial for patient diagnosis and treatment planning. However, manual segmentation of these infected areas is both labor-intensive and prone to variability, making it difficult to ensure consistent results. To address this challenge, artificial intelligence (AI)-based models have been increasingly utilized for automating segmentation tasks (1). Despite advancements in AI-driven medical imaging, precise segmentation of COVID-19-infected lung regions remains challenging due to the diverse manifestations of the disease in CT scans. Over time and throughout the progression of the COVID-19 pandemic, significant insights have been gained. Leveraging these insights, in conjunction with robust AI algorithms, facilitates effective crisis management and mitigation of future consequences when faced with comparable situations. The variability in the size, shape, and location of COVID-19 lesions poses notable challenges for medical professionals during the diagnostic process; thus, AI algorithms present valuable solutions to these difficulties. A substantial body of research has employed encoder-decoder networks for the automated delineation of COVID-19 lesions (2). These networks predominantly utilize convolutional layers in both the encoder and decoder components, with feature extraction being primarily performed by these layers. Notably, many studies have proposed architectures based on the U-Net model (3-5). Paluru et al. suggested that their segmentation approach is characterized by significantly fewer parameters than the U-Net architecture, rendering it advantageous for deployment on resource-constrained platforms such as mobile devices (6). In their investigation, the Anam-Net architecture was utilized to seg-

***Corresponding Author:** Seyed Ali Asghar Beheshti Shirazi; Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran.Email: abeheshti@iust.ac.ir, Phone: 0098-21-73225620, Fax: 0098-21-7322 5777, ORCID: https://orcid.org/0000-0001-9603-5117.

ment both normal and abnormal regions within lung CT images. This network features a symmetric encoder-decoder structure and employs an adapted label weighting technique to enhance performance during the training phase. Due to its lightweight design, which resulted in reduced training and testing times, this approach is deemed suitable for practical applications, potentially facilitating local training within clinical settings. The authors reported a Dice coefficient of 0.972 for the normal class and 0.755 for the abnormal class utilizing the COVID-19 Chest CT public dataset (6). Aswathy et al. proposed an innovative strategy aimed at enhancing segmentation accuracy by implementing two successive 3D U-Net networks for the detection of infected areas (7). Their methodology involved initially segmenting lung regions, followed by the application of post-processing techniques, which allowed the identified regions to be input into the subsequent network. This patch-based approach achieved Dice values of 0.9246 and 0.82 in the first and second stages, respectively (7).

In summary, these studies have endeavored to improve segmentation accuracy by considering a variety of factors and employing convolutional networks for the segmentation of COVID-19-infected regions in chest CT scan images. However, the capacity of these networks to identify long-range relationships among pixels and extract contextual and global features is inherently limited. This inability is due to factors such as feature extraction via convolutional layers, reliance on shared weight assignments within localized areas, uniform filter application, recovery of high-resolution representations from low-resolution ones, and the utilization of local receptive fields (RFs) (8, 9). Additionally, these networks have prioritized nearby spatial information and treated all pixels equally, which undermines their ability to effectively detect lesions that exhibit varied shapes and scales. Recent studies have indicated that transformer-based models may outperform convolutional networks and address the limitations of these architectures, particularly in scenarios involving limited datasets. The advantages of transformer models have gained increasing recognition among researchers across diverse domains within computer vision (CV). Consequently, there is a growing interest in utilizing these models for tasks encompassing image classification, object recognition, representation learning, and semantic segmentation(10-16) . Given that the annotation of 3D images can be both time-consuming and costly for experts, transformer models may serve as valuable tools within clinical applications. This study aimed to introduce a novel approach for identifying COVID-19-infected regions in lung CT scan through the development of two innovative transformer models.

2. Methods

2.1. Study design and setting

In this study we used the Squeeze and Excitation based UNet TRansformers (SE-UNETR) and the Squeeze and Excitation based High-Quality Resolution Swin Transformer Network (SE-HQRSTNet), to develop two three-dimensional segmentation networks for identifying COVID-19-infected regions in lung CT scan.

The performance of our models was evaluated using two publicly available datasets: COVID-19-CT-Seg (17) and MosMed (18). The COVID-19-CT-Seg dataset consists of 20 three-dimensional chest CT scan images, accompanied by three benchmarks. These images were sourced from the Coronacases Initiative and Radiopedia, with resolutions of 512×512 and 630×630 across multiple slices, respectively (3, 17). The MosMed dataset was obtained from various hospitals in Moscow, Russia, and comprises 50 images that have been annotated by expert radiologists (18). To enhance the performance of the proposed models, we employed data augmentation (DA) techniques alongside a 5-fold cross-validation methodology. The DA techniques included scaling, random cropping, shifting, rotation, and adjustment of voxel intensity values.

It is notable that the COVID-19-CT-Seg dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike International (CC BY-NC-SA) certificate (19). This dataset is a collaborative effort between Coronacases Initiative and Radiopaedia's open data disclosure (17). Additionally, the MosMed dataset we used consists of anonymized human lung CT scans, including those with and without COVID-19-related findings. This dataset is governed by a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License (18, 20). Both datasets are available publicly and we declare that our study adheres to the strict guidelines of handling de-identified data and emphasize that our research's purpose is to contribute to the scientific understanding of COVID-19.

2.2. SE-UNETR architecture

The SE-UNETR model incorporates a stack of transformer layers consisting of three sublayers: normalization layer, multi-head self-attention (MSA), and multilayer perceptron (MLP), alongside a Squeeze and Excitation (SE) mechanism designed to suppress irrelevant features (Figure 1). The transformer block focuses on capturing features across various locations and resolutions through multiple parallel self-attention (SA) mechanisms, thereby extracting salient features from critical regions of the image (11). Within this model, the SE block initially captures global information by generating a squeezed channel descriptor through the application of an average pooling operation on the feature maps (21).

This output is subsequently processed using two fully connected (FC) layers to create channel-wise scaling factors (21). The adaptive weights produced during this process enhance the prominence of the most relevant features by emphasizing significant channels (22). This recalibration technique improves the discriminative capacity of the model for the detection of COVID-19 lesions.

This network is designed following a U-shaped architecture in which the encoder component is connected to the decoder through skip connections (Figure 1). The encoder and decoder serve distinct yet complementary roles, the encoder is specifically responsible for learning contextual information, whereas the decoder focuses on extracting local information. For this network, we generated a one-dimensional sequence from the input volume with dimensions H × W × D × C, where H, W, and D represent the spatial resolution, and C represents the number of input channels. The input volume was segmented into flattened, uniform, non-overlapping patches. If each patch has dimensions P× P× P, then the length of the sequence is given by N = (H × W × D)/P^3. For this analysis, a patch resolution of 16 × 16 × 16 is employed.

Initially, a linear layer and an embedding layer were applied to the sequence of patches. Subsequently, the sequence undergoes normalization before passing through three linear layers, each with distinct weights, resulting in the generation of three matrices corresponding to the query (q), key (k), and value (v). The MSA sublayer was configured with 12 heads. The SA is determined using the following equation (8): SA = v [Softmax ($qk^T/\sqrt{K_h}$)]; where v denotes the values within the input sequence, and K_h is a scaling factor computed as K/n (8).

We selected a sequence representation (Z_3, Z_6, Z_9, Z_{12}) from the transformer and changed their resolution by passing through several blocks consisting of 2 × 2 × 2 deconvolutional, 3 × 3 × 3 convolutional, batch normalization, and ReLU activation function. The resulting output was then processed through an SE block, facilitating the integration of the encoder's output with the upsampled feature maps from the decoder. To enhance information representation and retention, a 2 × 2 × 2 deconvolutional layer is applied to the output of the last transformer layer. The resultant feature map was concatenated with Z_9, and this concatenated result was subsequently fed into successive convolutional layers, another SE block, and a deconvolutional layer. This process was iteratively repeated for the remaining layers. The integration of the transformer block, the SE mechanism, and skip connections collectively contribute to the extraction of optimal features.

Ultimately, the final output is processed through a 1 × 1 × 1 convolutional layer to generate a mask (Figure 1).

2.3. SE-HQRSTNet architecture

The structure of the SE-HQRSTNet is illustrated in Figure 2. We considered the input image to have dimensions D × H × W × C. Where D, H, and W represent the spatial dimensions, and C indicates the number of input channels. The input image was segmented into smaller patches using a 3D convolutional block. The resulting patch embedding vectors were then fed into the first Swin Transformer of the initial stage. The present model is organized into four stages, with each stage generating features at different resolutions (Figure 2). Each stage comprises parallel Swin Transformers, SE mechanisms, and patch merging blocks. The Swin Transformer blocks at each stage process and refines feature maps of varying resolutions. Part a of Figure 3 provides a visual representation of a Swin Transformer block, which incorporates two attention mechanisms: Windowed MSA (W-MSA) and Shifted Windowed MSA (SW-MSA) (23).

In this architecture, unlike the previous architecture, attention calculations were performed within small shifted windows that encompass multiple tokens. Our model effectively captures complex non-linear relationships between image pixels and contextual information through the utilization of the Swin Transformer block in conjunction with the SE mechanism. The number of attention heads within the Swin Transformer blocks varies according to the resolutions: 3 heads for D/4 × H/4 × W/4, 6 heads for D/8 × H/8 × W/8, 12 heads for D/16 × H/16 × W/16, and 24 heads for D/32 × H/32 × W/32. Notably, the output resolution of each Swin Transformer block corresponds to its input resolution. An SE block was integrated after each Swin Transformer block, except for the fourth Swin Transformer block within the fourth stage. As depicted in Figure 2 , a patch merging block follows the SE block, performing down-sampling on the SE output and thereby reducing its spatial dimensions.

The patch merging block of the first stage produces a feature map with a resolution of D/8 × H/8 × W/8, which is subsequently transmitted to the second stage. In the second stage, this process yields two feature maps with resolutions of D/8 × H/8 × W/8 and D/16 × H/16 × W/16.

Similarly, in the third stage, three feature maps were generated with resolutions of D/8 × H/8 × W/8, D/16 × H/16 × W/16, and D/32 × H/32 × W/32. The Multi-Resolution Feature Fusion (MRFF) block was implemented following each stage, except for the first stage (Figure 2). Part b of Figure 3 illustrates the functionality of the MRFF block after the fourth stage. This block receives four feature maps with resolutions of D/4 × H/4 × W/4, D/8 × H/8 × W/8, D/16 × H/16 × W/16, and D/32 × H/32 × W/32. The multi-resolution feature fusion (MRFF) block converts the resolution of the input feature maps into the other three resolutions by employing patch merging and patch expanding techniques. For instance, a feature map with a resolution of D/4 × H/4 × W/4 traverses three patch merging layers to produce features with resolutions of D/8 × H/8 × W/8, D/16 × H/16 × W/16, and D/32 × H/32 × W/32.

Each patch merging layer reduces the feature resolution to half of its original dimension, thereby generating various resolutions. Subsequently, the features obtained from the MRFF are aggregated at corresponding resolutions to retain comprehensive information. The resultant outputs are then processed through residual blocks to generate the final output (Figure 3). This design ensures that the MRFF block effectively extracts and combines suitable features. The outputs from the third MRFF block are subsequently connected after being converted to the resolution of D/4 × H/4 × W/4. The processed output then undergoes a sequence of trans-

formations, which includes being passed through a residual block, followed by patch expanding and convolutional operations, ultimately yielding the predicted output. This architecture obviates the necessity of recovering high-resolution representations from low-resolution ones, thereby preserving the integrity of the highest-resolution features throughout the processing stages.

2.4. Objectives

The primary objective of this study was to evaluate the performance and efficacy of contemporary methodologies in segmenting regions affected by COVID-19.

2.5. Statistical analysis

We designed two three-dimensional (3D) semantic segmentation models. These models consider the interdependencies between different channels in the feature set by the SE blocks. We utilized PyTorch 1.8.1 as our primary deep learning framework with CUDA 10.2 and cuDNN 8.1.1 to optimize the Convolutional Neural Network (CNN) models. Both models were rigorously evaluated using a 5-fold cross-validation approach alongside DA techniques, employing two publicly available datasets: COVID-19-CT-Seg and MosMed.

For the evaluation of the models, we calculated their screening performance characteristics (sensitivity, specificity, and Dice value) based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. Dice value was calculated as 2TP/2TP+FP+FN.

We employed the DiceCELoss function as our loss metric, which combines the Dice loss and cross-entropy loss functions. This hybrid approach capitalized on the advantages inherent in both loss functions.

We also utilized grid search to explore a range of batch sizes, evaluating model performance for each batch size. The process began with larger batch sizes and progressively reduced the batch size to ensure the model fit within the available memory constraints. The models were executed with a batch size of 1, utilizing the AdamW optimization algorithm with an initial learning rate set to 10^{-4}.

3. Results

Our models segmented lung regions and COVID-19-infected areas within the COVID-19-CT-Seg dataset. Tables 1 and 2 show the screening performance characteristics of the proposed models on the COVID-19-CT-Seg dataset. The values of the two tables demonstrate that lung segmentation faced fewer challenges than the segmentation of infection regions. Table 3 presents the performance of the proposed models on the MosMed dataset.

Figure 4 shows a visual comparison between the ground truth and prediction of the models before and after applying SE blocks on the COVID-19-CT-Seg dataset for 3 CT scan samples of COVID-19 patients.

Table 4 illustrates the enhanced performance of the proposed methodologies before and after applying the SE block compared to recent studies.

4. Discussion

In prior research, a significant challenge associated with CNN models is their performance relative to the overall image scale. In these studies, filters are applied locally to each pixel within the image, resulting in feature maps that predominantly reflect the relationships among nearby pixels. Consequently, while CNNs are adept at extracting local features, they often fall short of capturing optimal global features and the long-range relationships among pixels (24). Efforts to enhance the network's capacity to extract global features such as increasing filter sizes and the number of layers tend to elevate model complexity, computational demands, error rates, and execution times. Segmentation accuracy may be influenced by multiple factors, including the specific annotation guidelines, variability among annotators, and the quality of the images used. We designed two new models to solve the challenges of a CNN model. In this study, the first designed model is a U-shaped architecture that includes a series of transformers and SE mechanisms for effectively capturing contextual information and fine details. This patch-based network utilizes the SA mechanism within the transformer block to assess the interrelationships between each patch and all other patches. Given that CT medical images exhibit varying spatial distributions, and lesion sizes in COVID-19 patients differ, this network can adapt to various lesion sizes through the transformer block. The decoder component of the model is structured around a CNN framework and integrated with the SE mechanism, enabling the model to capture local information robustly. In this architecture, skip connections are employed to link the features extracted by the encoder to those processed by the decoder. As the number of patches increases within this model, the computational burden and complexity increase, potentially impairing the model's ability to produce precise deep predictions. This limitation constitutes a notable drawback for this architecture. Our second model can overcome the limitations of the first model.

The second model is also patch-based, which computes the SA in small, shifted windows containing several 3D tokens. This model concentrates on salient features through the window-based module. Subsequently, the SE block effectively suppresses less relevant features and amplifies the learning of distinctive attributes. Moreover, in this network, MRFF blocks help to create new features with more detailed and richer semantic details by exchanging information at different resolutions.

We used the proposed models to segment lung regions and COVID-19-infected areas. The experiment results showed superior segmentation of lung areas compared to COVID-19-infected regions. This superiority can be attributed to the relatively consistent and stable anatomical structure of the lungs, as well as their intensity patterns in CT scan im-

ages. In contrast, COVID-19 lesions exhibit considerable variability in shape, size, position, and intensity patterns. Furthermore, the infected regions often present indistinct boundaries, whose edges may merge with adjacent structures, thereby additional complexities are created to the segmentation of these lesions.

Table 4 illustrates the SE-UNETR and SE-HQRSTNet models improving the Dice coefficient for the COVID-19-CTSeg dataset from 82% (the highest reported value in prior literature) to 85.81% and 86.84%, respectively. Furthermore, these models increased the Dice coefficient for the MosMed dataset from 66.8% to 69.35% and 70.89%, respectively. Consequently, our proposed approaches have demonstrated superior accuracy compared to traditional CNNs by considering the information from all pixels within an image. Additionally, the results of our evaluations indicated that the SE-HQRSTNet model outperforms the SE-UNETR model across both datasets. As a result, SE-HQRSTNet has superior convergence capabilities and can effectively capture and comprehend the nuances of image details and content.

We wish to emphasize that even minor enhancements can have substantial real-world implications. In the context of COVID-19, timely and accurate diagnosis is essential for effective patient management and public health responses. An increase in accuracy can lead to fewer false negatives, ensuring that more cases are promptly identified and treated. This is critical for controlling transmission and improving patient outcomes. Thus, the proposed segmentation models can serve as standardized and targeted tools that contribute to the evaluation and interpretation of imaging findings.

5. Limitations

It is notable that our study also has some potential limitations related to model overfitting and increased complexity. Relying solely on fine-tuning may lead to overfitting, particularly when the training dataset is limited or when the model architecture lacks the robustness to generalize across diverse inputs. However, the incorporation of the SE module enhances the model's robustness by enabling it to better discern important features through channel-wise attention. Furthermore, while the added complexity of the SE module can enhance the model's ability to learn from intricate patterns in the data, this complexity is essential for distinctive features in CT images of COVID-19 infections.

6. Conclusions

In the current study, we proposed two segmentation models, which outperform recent methodologies across two publicly available datasets.

Notably, among the two models, SE-HQRSTNet exhibited a greater capacity for learning fine-grained features while effectively capturing contextual information. This approach preserved high-resolution feature maps and facilitated the continuous exchange of information across different resolutions through the MRFF block. Furthermore, this strategy enabled the model to learn global dependencies and identify non-linear relationships among feature channels by employing a window-based module in conjunction with the SE mechanism. Given that CT medical imaging produces high-resolution images, the resolution of feature maps is critical for achieving accurate segmentation outcomes.

7. Declarations

7.1. Acknowledgments

The authors have no acknowledgments to report.

7.2. Authors contributions

Ali Asghar Beheshti Shirazi and Zeinab Momeni Pour contributed to the conceptualization and methodology of the study, with Ali Asghar Beheshti Shirazi providing supervision throughout the process. Zeinab Momeni Pour was responsible for writing the original draft, while both authors were involved in the writing, review, and editing stages. Additionally, Ali Asghar Beheshti Shirazi managed the project administration.

7.3. Conflict of interest

The authors declare that there is no conflict of interest related to this article. All aspects of the research were conducted in an impartial manner, and no financial or personal relationships have influenced the results or interpretations presented.

7.4. Funding and supports

This research project was conducted without any external funding or financial support.

7.5. Data availability

The authors declare that data from the study are available and will be provided if anyone needs them.

7.6. Using artificial intelligence chatbots

The content of this article was not generated using any artificial intelligence chatbot.

References

1. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. Phys Med Biol. 2020;65(20):20TR01.
2. Ter-Sarkisov A. Detection and segmentation of lesion areas in chest CT scans for the prediction of COVID-19. MedRxiv. 2020:2020.10. 23.20218461.
3. Müller D, Rey IS, Kramer F. Automated chest ct image segmentation of covid-19 lung infection based on 3d u-net. arXiv preprint arXiv:2007.04774. 2020.
4. 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.
5. 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.
6. Paluru N, Dayal A, Jenssen HB, Sakinis T, Cenkeramaddi LR, Prakash J, et al. Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Trans Neural Netw Learn Syst. 2021;32(3):932-46.
7. Aswathy A, SS VC. Cascaded 3D UNet architecture for segmenting the COVID-19 infection from lung CT volume. Scientific Reports. 2022;12.
8. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al., editors. Unetr: Transformers for 3d medical image segmentation. Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2022.
9. Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution Swin transformer for automatic medical image segmentation. Sensors. 2023;23(7):3420.
10. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.
13. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H, editors. Training data-efficient image transformers & distillation through attention. International conference on machine learning; 2021.
14. Beal J, Kim E, Tzeng E, Park DH, Zhai A, Kislyuk D. Toward transformer-based object detection. arXiv preprint arXiv:2012.09958. 2020.
15. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077-90.
16. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al., editors. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision; 2021.
17. COVID-19-CT-Seg dataset. https://zenodo.org/record/3757476#.
18. 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.
19. 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.
20. Morozov SP, Andreychenko AE, Pavlov NA, Vladzymyrskyy AV, Ledikhova NV, Gombolevskiy VA, et al. Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465. 2020.
21. Hu J, Shen L, Sun G, editors. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
22. Lee S, Lee M. MetaSwin: a unified meta vision transformer model for medical image segmentation. PeerJ Comput Sci. 2024;10:e1762.
23. 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.
24. Kundu S, Sundaresan S, editors. Attentionlite: Towards efficient self-attention models for vision. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021.
25. 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.
26. Zheng R, Zheng Y, Dong-Ye C. Improved 3D U-Net for COVID-19 chest CT image segmentation. Sci Program. 2021;2021(1):9999368.

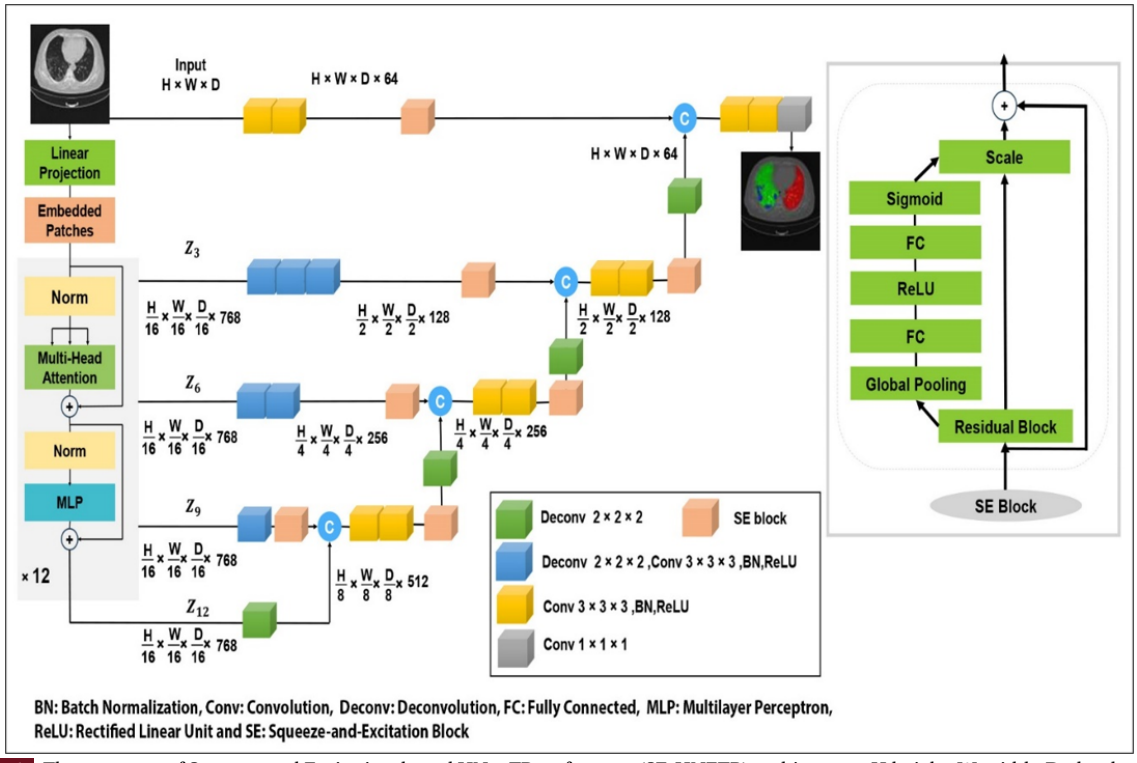

Figure 1: The structure of Squeeze and Excitation-based UNet TRansformers (SE-UNETR) architecture. H:height; W: width; D: depth.

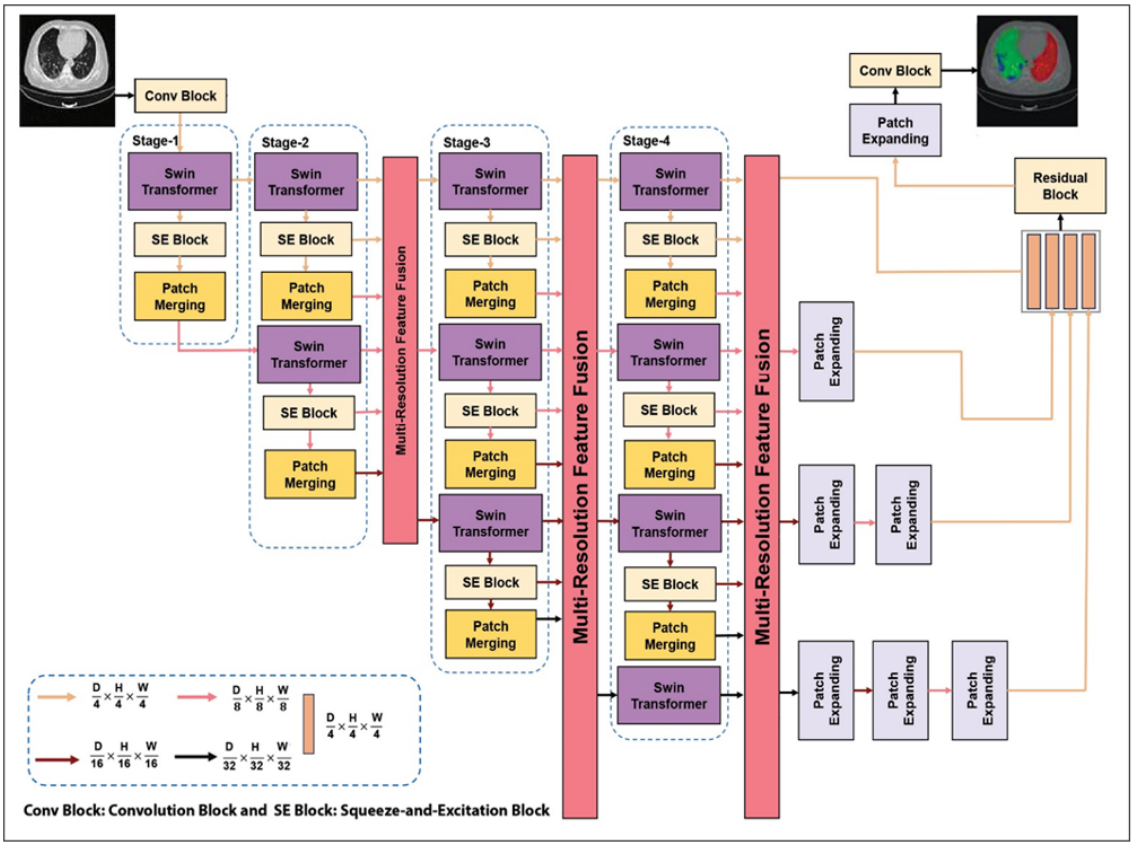

Figure 2: The structure of the Squeeze and Excitation-based High-Quality Resolution Swin Transformer Network (SE-HQRSTNet) architecture. H:height; W: width; D: depth.

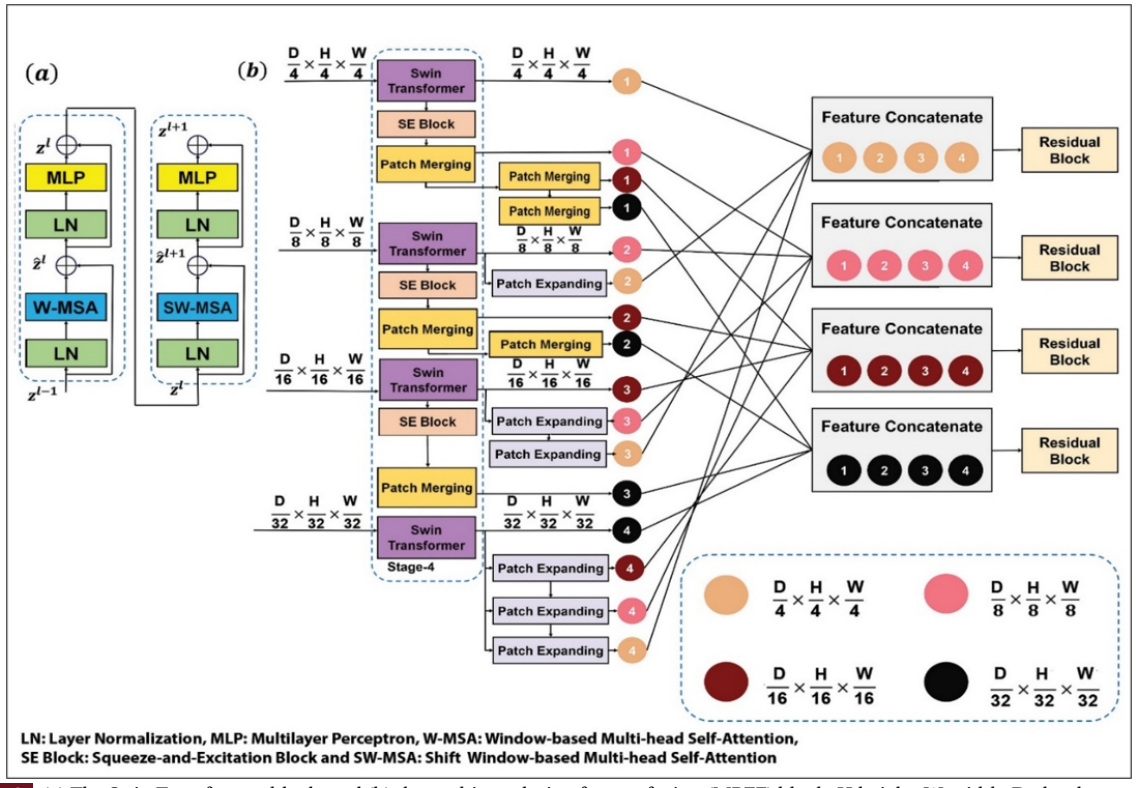

Figure 3: (a) The Swin Transformer block, and (b) the multi-resolution feature fusion (MRFF) block. H:height; W: width; D: depth.

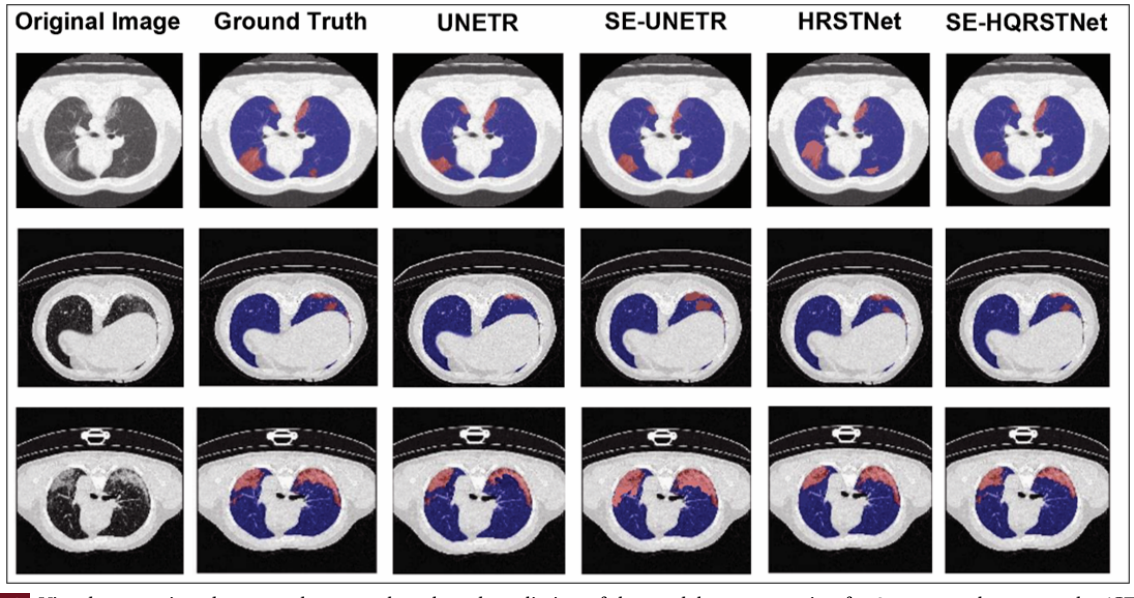

Figure 4: Visual comparison between the ground truth and prediction of the models segmentation for 3 computed tomography (CT) scan samples of COVID-19 patients. SE: Squeeze-and-Excitation, UNETR: UNEt TRansformers, HRSTNet: High-Resolution Swin Transformer Network, HQRSTNet: High-Quality Resolution Swin Transformer Network.

Table 1: Screening performance characteristics of the Squeeze and Excitation-based UNet TRansformers (SE-UNETR) model on COVID-19-CT-Seg dataset

SE-UNETR	Specificity	Sensitivity	Dice
Infection mask			
Fold1	0.9248 (± 0.0150)	0.8689 (± 0.0120)	0.8592 (± 0.0115)
Fold2	0.9587 (± 0.0112)	0.8596 (± 0.0127)	0.8577 (± 0.0122)
Fold3	0.9780 (± 0.0103)	0.8980 (± 0.0131)	0.8683 (± 0.0118)
Fold4	0.9643 (± 0.0126)	0.8593 (± 0.0125)	0.8481 (± 0.0117)
Fold5	0.9679 (± 0.0118)	0.8829 (± 0.0130)	0.8576 (± 0.0120)
Average	0.9587 (± 0.0122)	0.8737 (± 0.0127)	0.8581 (± 0.0118)
Lung mask			
Fold1	0.9940 (± 0.0025)	0.9775 (± 0.0050)	0.9565 (± 0.0060)
Fold2	0.9992 (± 0.0018)	0.9783 (± 0.0049)	0.9672 (± 0.0058)
Fold3	0.9951 (± 0.0023)	0.9790 (± 0.0051)	0.9590 (± 0.0059)
Fold4	0.9983 (± 0.0021)	0.9879 (± 0.0044)	0.9648 (± 0.0057)
Fold5	0.9994 (± 0.0015)	0.9981 (± 0.0040)	0.9686 (± 0.0054)
Average	0.9972 (± 0.0020)	0.9841 (± 0.0047)	0.9632 (± 0.0057)

All measures are presented with 95% confidence interval.

Table 2: Screening performance characteristics of the Squeeze and Excitation-based High-Quality Resolution Swin Transformer Network (SE-HQRSTNet) model on COVID-19-CT-Seg dataset

SE-HQRSTNet	Specificity	Sensitivity	Dice
Infection mask			
Fold1	0.9386 (± 0.0105)	0.8789 (± 0.0123)	0.8669 (± 0.0111)
Fold2	0.9684 (± 0.0098)	0.8690 (± 0.0134)	0.8794 (± 0.0125)
Fold3	0.9792 (± 0.0083)	0.9057 (± 0.0115)	0.8698 (± 0.0104)
Fold4	0.9797 (± 0.0092)	0.8694 (± 0.0136)	0.8599 (± 0.0118)
Fold5	0.9689 (± 0.0101)	0.8975 (± 0.0129)	0.8660 (± 0.0109)
Average	0.9669 (± 0.0096)	0.8841 (± 0.0127)	0.8684 (± 0.0113)
Lung mask			
Fold1	0.9958 (± 0.0030)	0.9798 (± 0.0082)	0.9690 (± 0.0075)
Fold2	0.9984 (± 0.0025)	0.9893 (± 0.0076)	0.9789 (± 0.0068)
Fold3	0.9975 (± 0.0029)	0.9888 (± 0.0071)	0.9664 (± 0.0073)
Fold4	0.9990 (± 0.0021)	0.9881 (± 0.0065)	0.9795 (± 0.0077)
Fold5	0.9997 (± 0.0017)	0.9979 (± 0.0061)	0.9791 (± 0.0069)
Average	0.9980 (± 0.0024)	0.9887 (± 0.0071)	0.9745 (± 0.0072)

All measures are presented with 95% confidence interval.

Table 3: Screening performance characteristics of the proposed models on MosMed dataset

SE-UNETR	Specificity	Sensitivity	Dice
Fold1	0.9868 (± 0.0025)	0.7378 (± 0.0152)	0.7098 (± 0.0145)
Fold2	0.9889 (± 0.0021)	0.7058 (± 0.0183)	0.6698 (± 0.0168)
Fold3	0.9862 (± 0.0023)	0.6999 (± 0.0149)	0.6686 (± 0.0123)
Fold4	0.9859 (± 0.0018)	0.7186 (± 0.0173)	0.6799 (± 0.0159)
Fold5	0.9985 (± 0.0012)	0.7936 (± 0.0187)	0.7397 (± 0.0174)
Average	0.9892 (± 0.0020)	0.7311 (± 0.0172)	0.6935 (± 0.0154)
SE-HQRSTNet	**Specificity**	**Sensitivity**	**Dice**
Fold1	0.9958 (± 0.0015)	0.7698 (± 0.0134)	0.7287 (± 0.0125)
Fold2	0.9955 (± 0.0016)	0.7194 (± 0.0178)	0.6799 (± 0.0156)
Fold3	0.9981 (± 0.0011)	0.7055 (± 0.0165)	0.6680 (± 0.0142)
Fold4	0.9963 (± 0.0013)	0.7285 (± 0.0159)	0.6998 (± 0.0143)
Fold5	0.9929 (± 0.0020)	0.8276 (± 0.0184)	0.7684 (± 0.0175)
Average	0.9957 (± 0.0015)	0.7501 (± 0.0164)	0.7089 (± 0.0148)

SE: Squeeze-and-Excitation, UNETR: UNEt TRansformers, HQRSTNet: High-Quality Resolution Swin Transformer Network.

Table 4: Comparison of results of our models with previous studies

Author	Dataset	Splitting Type	Method: Dice
Müller et al. (3)	COVID-19-CT-Seg	5-Fold	3D U-Net:0.761
Ma et al. (4)	COVID-19-CT-Seg	5-Fold	nnU-Net:0.673
Wang et al. (5)	COVID-19-CT-Seg	5-Fold	3D U-Net: 0.704
Singh et al. (25)	COVID-19-CT-Seg	Train:70% Validation:10% Test:20%	LungINFseg:0.8034
Aswathy et al. (7)	COVID-19-CT-Seg	Train:60% Validation:20% Test:20%	Cascaded 3D U-Net:0.820
Our method	COVID-19-CT-Seg	5-Fold	UNETR:0.8519 SE-UNETR:0.8581 HRSTNet: 0.8663 SE-HQRSTNet: 0.8684
Zheng et al. (26)	MosMed	5-Fold	3D CU-Net:0.668
Our method	MosMed	5-Fold	UNETR:0.6901 SE-UNETR:0.6935 HRSTNet: 0.7072 SE-HQRSTNet: 0.7089

SE: Squeeze-and-Excitation ,UNETR: UNEt TRansformers , HQRSTNet: High-Quality Resolution Swin Transformer Network, HRSTNet: High-Resolution Swin Transformer Network.