

A Comparative Study on Image Segmentation Models in COVID-19 Diagnosis



Sheng Xu, Shuwen Chen, and Mike Chen

Abstract With Omicron sweeping the world, it has brought huge pressure on the healthcare system, and quick diagnosis of pneumonia caused by COVID-19 using chest computed tomography (CT) scans plays a key role in saving lives. Image processing techniques have been widely used to analyze CT scans and other medical images, which also facilitate the diagnosis and treatment of Coronavirus disease. This article introduces several image segmentation models that are used to facilitate the diagnosis of COVID-19 and applies several recently proposed deep learning (DL) models to the task, including UNet Transformers (UNETR++) and Dual Attention-guided Efficient Transformer (DAE-Former). We compare the performance of these models and provide a thorough analysis of different methods. The experimental results show that for the task of segmenting COVID-19 lesion areas, both Transformer-based models obtained better performance in terms of the value of mIoU compared with U-Net, a CNN-based model, while U-Net obtained higher accuracy. The DAE-Former model has superior anti-noise ability to UNETR++, whereas UNETR++ is more robust in terms of domain transfer.

Keywords Deep learning · Medical image processing · Image segmentation · COVID-19 · UNETR++ · DAE-former

Sheng Xu, Shuwen Chen—These authors contributed equally to this work. Sheng Xu and Shuwen Chen are regarded as co-first authors.

S. Xu · S. Chen · M. Chen (✉)
School of Physics and Information Engineering, Jiangsu Second Normal University,
Nanjing 211200, China
e-mail: mc277509@ohio.edu

S. Chen · M. Chen
Jiangsu Province Engineering Research Center of Basic Education Big Data Application,
Nanjing 211200, China

1 Introduction

According to the latest data provided by the World Health Organization (WHO), as of 18th December 2022, over 649 million confirmed cases and over 6.6 million deaths have been reported globally since the outbreak of COVID-19 [1]. All viruses, including SARS-CoV-2, the virus that causes COVID-19, change over time. Since the beginning of the coronavirus pandemic, a number of prominent variants have been seen, including Alpha, Beta, Delta, and Omicron. According to recent studies on coronavirus variants, the Omicron variant has higher transmission rate and brings a significant risk of neutralizing antibody escape from convalescent patients [2, 3], catalyzing the fourth wave of the pandemic in many countries worldwide.

The early identification of COVID-19 cases plays a crucial role to provide fast and efficient treatment to the effected patients. Reverse transcription polymerase chain reaction (RT-PCR) is widely used to detect infectious agents, for example, SARS-CoV-2 [4]. In spite of its simplicity, specificity and sensitivity, RT-PCR still suffers from the long processing time (4–8 h) as well as false-negative and false-positive results [5]. Therefore, the use of chest radiographs and chest CT scans in the clinical diagnosis of COVID-19 becomes a necessary addition, especially in severe cases of pneumonia, as shown in Fig. 1.

With the increase in confirmed and suspected COVID-19 cases, manually analyzing such a large amount of CT scans in a limited time may incur errors. Automatic analysis of CT scans using image processing techniques provides an efficient alternative approach [7]. We present a review of machine learning approaches related to the task of image segmentation, which is a crucial step in the automatic diagnosis of COVID-19 [8]. We also apply several recently proposed deep learning

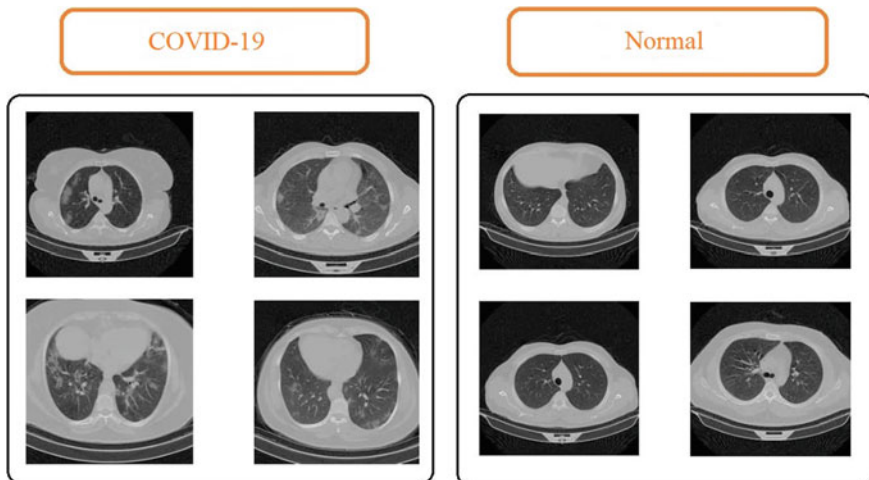


Fig. 1 Examples of normal and COVID-19 infected chest CT scans [6]

models to the task and compare the performance of different models. The experimental results show that for the task of segmenting COVID-19 lesion areas, both Transformer-based models (DAE-Former and UNETR++) obtained better performance in terms of the value of mIoU compared with U-Net, a CNN-based model, while U-Net obtained higher accuracy. The DAE-Former model has superior anti-noise ability to UNETR++, whereas UNETR++ is more robust in terms of domain transfer.

The rest of the paper is organized as follows: first, we introduce several image segmentation techniques that are widely used in medical image processing and the diagnosis of COVID-19; second, we apply two recently proposed deep learning models to the task, then compare and analyze the experimental results; finally, we end with concluding remarks.

2 Machine Learning Approaches Used in the Task of Image Segmentation

Image segmentation is the process of partitioning a digital image into multiple image segments, which is an important step in the automatic diagnosis of COVID-19. It locates regions of interest (ROI) in chest X-ray or CT images, for example, lungs, lobes, bronchopulmonary segments and infected regions or lesions. Image segmentation techniques can facilitate the diagnosis of COVID-19 by locating infected regions in X-ray and CT images, thus helping radiologists to make clinical decisions [9]. This section summarizes related works in image segmentation and its applications in the diagnosis of COVID-19.

2.1 Evaluation Metrics for Image Segmentation

In general, image segmentation evaluation methods can be divided into two main categories: empirical evaluation and analytical evaluation methods. The empirical methods analyze a segmentation method with respect to its findings and outputs, whereas the analytical methods assess the method based on its complexity, functionality, and utilities. In this paper, we focus on the empirical methods. In addition to accuracy, there are four basic quantitative metrics, which are true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Some other metrics are also widely used in image segmentation. For example, true positive rate (TPR), also known as sensitivity or recall, measures the fraction of pixels that are true positives among the total class-relevant pixels.

$$\text{Recall} = \text{Sensitivity} = \text{TPR} = \frac{TP}{TP + FN} \quad (1)$$

False positive rate (FPR), also known as specificity, measures the ability of predicting negative examples.

$$Specificity = FPR = \frac{TN}{TN + FP} \quad (2)$$

Precision is the proportion of positive samples that are predicted to be positive.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

F1-score is the harmonic average of precision and recall, with the maximum value being 1 and the minimum value being 0.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

ROC-AUC refers to the area under the ROC curve, which is a curve formed by connecting the points of FPR and TPR, where the horizontal axis is FPR, and the vertical axis is TPR.

Similarly, PR-AUC refers to the area under the PR curve, which is a curve formed by the points of precision and recall, with recall on the horizontal axis and precision on the vertical axis.

Dice Similarity Coefficient (DSC) is used to measure the similarity between predicted labels and true labels; the larger the value, the higher the segmentation accuracy is

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (5)$$

Hausdorff distance (HD) indicates the maximum distance between the boundary of the predicted segmentation region and the boundary of the manually labeled region; the smaller the value, the better the segmentation quality is. Let X and Y be two nonempty sets, then define their Hausdorff distance as $d_H(X, Y)$.

$$d_H(X, Y) = \max\{d_{XY}, d_{YX}\} = \max\left\{\max_{x \in X} \min_{y \in Y} d(x, y), \max_{y \in Y} \min_{x \in X} d(x, y)\right\} \quad (6)$$

In order to eliminate unreasonably large distances caused by outliers and obtain a more robust estimate of the maximum error, 95% Hausdorff distance (HD95) is used instead, also known as 95th percentile of the ordered distance measures.

Mean Intersection over Union (mIoU): A standard measure of image segmentation calculates the average of the ratio of the intersection of the segmentation results and the true labels to the union of the two, across all classes. Let k represent the number of categories, and $(k + 1)$ means that the background class is added.

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + FP + TP} \quad (7)$$

2.2 Image Segmentation Methods

Various techniques have been applied to the task of image segmentation, ranging from the inversion technique, active contour models, and the threshold method to the edge flow method and the clustering technique. The broad success of deep learning models has prompted the development of new image segmentation approaches leveraging DL models, among which CNN-based and Transformer-based models are widely used and proved to be effective.

CNN-based segmentation methods: With the success of the U-Net architecture [10], variants of the U-Net model have been proposed to solve the image segmentation task, which are also used to detect lesion regions of lungs in medical images of COVID-19 patients. Table 1 shows the image segmentation performance of different CNN-based models. Zheng et al. [11] identified the lung region using a pre-trained U-Net, and the resulting segmentation mask is fed into a deep neural network to diagnose COVID-19 automatically. Gaál et al. [12] proposed the attention U-Net model to accomplish the task of image segmentation. Naqvi et al. [13] modified the U-Net architecture by adding a new block that applied morphological operations and binarization to the segmentation output from U-Net, which performed well in automatic lung segmentation. Qi et al. [14] applied logistic regression (LR) and random forest (RF) to features extracted from pneumonia lesions, and predicted hospital stay in patients with pneumonia associated with COVID-19. Wang et al. [15] introduced COPL-Net, a noise-robust framework, to learn from noisy labels for segmentation of pneumonia lesions. Mishra et al. [16] developed a fully convolutional neural network (FCNN) to solve segmentation of ultrasound (US) images. Traditional methods of tuning hyperparameters in these models are time-consuming and unstable. To address this, [17] used a Particle Swarm Optimization algorithm to build a PSO-guided Self-Tuning Convolution Neural Network (PSTCNN), allowing the model to automatically tune hyperparameters. This approach reduces human involvement and improves the selection of hyperparameter combinations to achieve a stable solution closer to the global optimum. Experimental results show that the PSTCNN performs exceptionally well, with high accuracy, sensitivity, specificity, precision, F1-score, Matthews Correlation Coefficient, and Fowlkes-Mallows Index. Compared to traditional methods, this optimization algorithm for hyperparameter tuning is faster and more effective. It is worth pointing out that CNN-based methods have been also widely explored in the diagnosis of other diseases, for example, lung cancer [18] and visual impairment [19].

Table 1 Image segmentation performance of different CNN-based models

Refs.	Technique	Image types	Performance
Zheng et al. [11]	U-Net	CT	ROC-AUC = 95.9%, PR-AUC = 97.6%
Gaál et al. [12]	Attention U-Net	X-ray	DSC = 97.5%
Naqvi et al. [13]	U-Net-based	X-ray	DSC = 98.3%
Qi et al. [14]	LR, RF	CT	Sensitivity = 100%, specificity = 89%
Wang et al. [15]	COPLE-Net	CT	DSC = 80.29%, HD95 = 18.72 mm
Mishra et al. [16]	FCNN	US	DSC = 79%, F1 = 83%, mIoU = 83%

In addition to the use in 2D images, CNN-based models have been explored in 3D image segmentation. Milletari et al. [20] designed a volumetric, fully convolutional neural network to accomplish 3D image segmentation. Çiçek et al. [21] proposed a network for volumetric segmentation which learned from sparsely annotated volumetric images by replacing 2D operations with 3D counterparts. Isensee et al. [22] introduced a segmentation framework called nnU-Net, which automatically configures itself, from preprocessing, network architecture, training to post-processing for any new task. The approaches above have been shown to perform well in the study of CT image segmentation to detect pneumonia caused by COVID-19 [23]. However, incorporating shape representations into CNN-based architectures, requires modeling long-range dependencies, which remains a challenge and affects the system performance [24].

Transformer-based segmentation methods: Vision Transformers (ViTs) have shown superior performance over CNN-based models by incorporating self-attention [25, 26]. Although ViT is applicable to image classification, it is challenging to directly adapt it to pixel-level dense predictions, for example, object detection and segmentation, mainly due to the high computational and memory costs even for common input image sizes. Therefore, the Pyramid Vision Transformer (PVT) was proposed, as shown in Fig. 2 [27]. ViT contains one Transformer encoder and was designed for image classification specifically; while PVT has several Transformer encoders resulting in multiple feature maps, which can be applied to different downstream tasks, for example, image classification, objection detection as well as image segmentation. Liu et al. [28] introduced shifted window based self-attention, which brought greater efficiency by limiting self-attention computation to non-overlapping local windows.

In addition to further improvement of CNN and Transformer based models [20, 22, 28], efforts were made to combine the two into hybrid models, for example, UNETR [29] and not-another Transformer nnFormer [30], and better performance has been achieved.

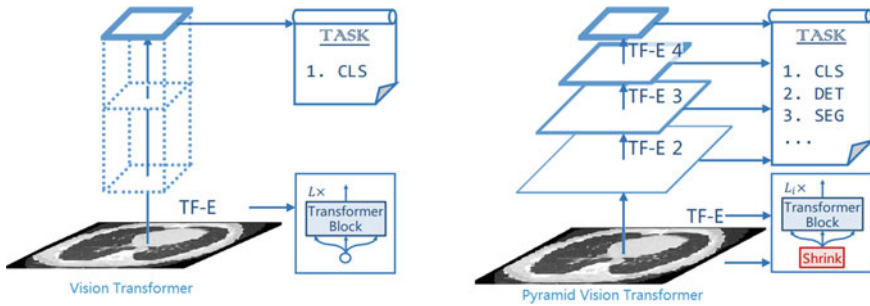


Fig. 2 Architectures of ViT and PVT. TF-E stands for transformer encoder, CLS stands for classification, DET stands for detection, and SEG stands segmentation

3 Automatic Diagnosis of COVID-19 Using DAE-Former and UNETR++

As described in the previous section, Transformer-based models as well as hybrids of Transformer and CNN have become state-of-the-art techniques for image segmentation. Therefore, we experiment with two recently proposed models for image segmentation to accomplish automatic diagnosis of COVID-19, which are DAE-Former [31] and UNETR++ [32]. DAE-Former is a Transformer-based model and UNETR++ is a hybrid model of CNN and Transformer. We introduce the two models and COVID-19 datasets as well as our experiments in detail in the following subsections.

3.1 Models

DAE-Former: A convolution-free U-Net-like hierarchical Transformer, incorporates the patch embedding module [33], efficient attention [34], transpose attention [35], efficient dual attention as well as skip connection cross attention (SCCA). As shown in Fig. 3, the encoder takes the tokenized output from the patch embedding module as input, and consists of 3 stacked encoder blocks, each of which is composed of two consecutive dual Transformer layers and a patch merging layer. The decoder expands the tokens by a factor of 2 in each block. SCCA is then used to fuse the output of each patch expanding layer with the features filtered by the skip connection from the corresponding encoder layer. The SCCA module cross-attends the features from the encoder and decoder layers, instead of simple concatenation, using efficient attention. The combined features go through two consecutive dual Transformer layers. Finally, a linear projection layer generates the resulting segmentation image.

Efficient attention drastically reduces the computational complexity, compared with the standard self-attention mechanism, with equally high representational power. Therefore, efficient attention is used to capture the spatial context of the input feature. Transpose attention scales quadratically with the embedding dimension, which is

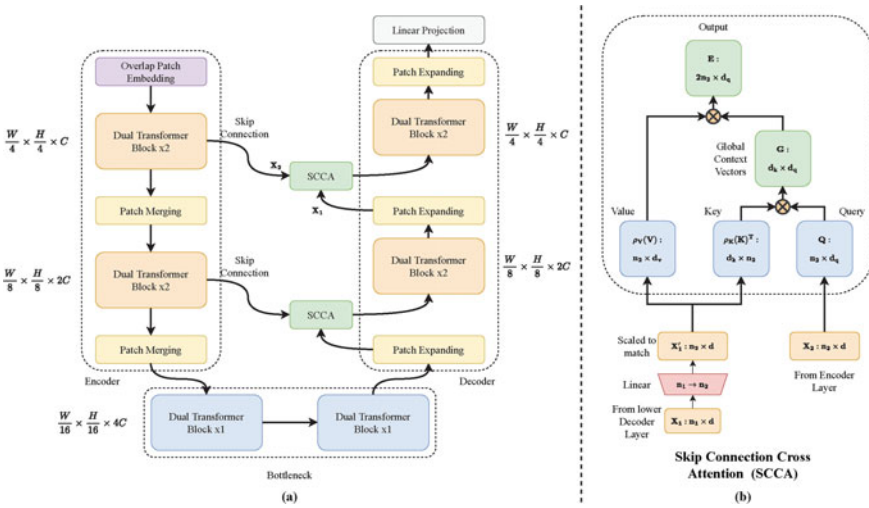


Fig. 3 The architecture of DAE-Former [31]

usually more computationally efficient than the standard self-attention mechanism, especially for larger images. The transpose attention is then used to capture the channel context of the input feature. As shown in Fig. 4, the dual Transformer block combines the transpose attention and efficient attention mechanism to capture the spatial and channel context simultaneously.

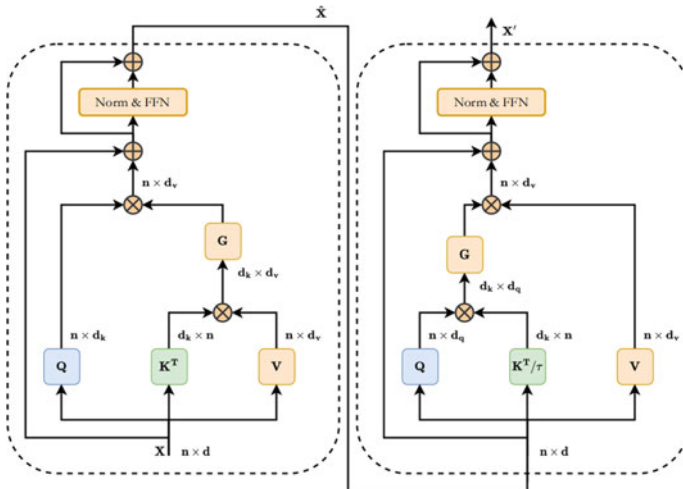


Fig. 4 The structure of dual transformer. It consists of an effective attention block followed by a Norm and FFN, and a channel attention block followed by a Norm and FFN to perform spatial and channel attention [31]

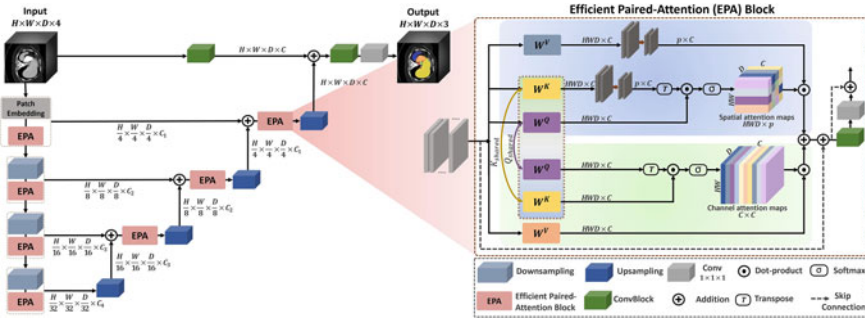


Fig. 5 Hierarchical encoder-decoder structure of the UNETR++ model [32]

UNETR++: A hybrid model for 3D medical image segmentation, achieves both high segmentation quality and efficiency in terms of parameters and computational cost. As shown in Fig. 4, the encoder takes the 3D patches as input, which are fed into an efficient paired-attention (EPA) block; the rest of the encoder are 3 stacked blocks, composed of downsampling layers and EPA. Skip connections are used between the encoder outputs and the decoder. The decoder consists of 4 stages, each of which contains an upsampling layer followed by the EPA block, except the last stage. Convolutional blocks are applied to the output of the last stage to generate the final segmentation masks.

The core component of the model is the EPA block, which incorporates both spatial and channel attention. Compared with the standard self-attention, the spatial attention module is able to reduce the complexity from quadratic to linear. The channel attention module is aimed to learn the relations between channel feature maps. As shown in Fig. 5, the weights of query and key mapping functions are shared between the two attention modules, which enables better feature representation and reduces the overall parameters as well. The spatial attention and channel attention outputs are used and transformed by convolution blocks.

3.2 COVID-19 Datasets

Table 2 lists the datasets used in the following experiments. Morozov et al. [36] constructed a dataset which contains CT scans annotated with COVID-19 related findings or without such findings. Ma et al. [37] built three benchmark datasets for lung and infection segmentation of CT scans of COVID-19 patients. COVID-19 [38] is a COVID-19 CT segmentation dataset, each image of which was segmented by experts using 3 labels: ground-glass, consolidation and pleural effusion. Soares et al. [39] annotated a large dataset of real patients CT scans, each of which was labeled as positive or negative for COVID-19.

Table 2 COVID-19 datasets

Type	Dataset	Description
CT-scans	Morozov et al. [36]	Chest CT scans with or without COVID-19 related findings
	Ma et al. [37]	COVID-19 CT lung and infection segmentation dataset
	COVID-19 [38]	COVID-19 CT segmentation dataset
	Soares et al. [39]	CT scans annotated as positive or negative for COVID-19

We combined the four datasets by using slice-level annotations to obtain segmentation labels and converted all images to 8-bits to have consistent depth. Finally, in order to reduce noise, we remove images that do not contain key information about lungs and deleted images that lack clear category labels or patient information. As a result, the dataset consists of 2729 pairs of images and segmentation masks of COVID-19 lesion areas.

3.3 Experiments

As mentioned in previous sections, CNN-based and Transformer-based models have been proven to be effective in image segmentation. We experiment with one CNN-based (U-Net) and two Transformer-based models (DAE-Former and UNETR++) to segment COVID-19 lesion areas. **Even though DAE-Former and UNETR++ have achieved state-of-the-art performance in image segmentation, no experiments have been done to show their performance in the segmentation of COVID-19 lesion areas.** U-Net is used as a baseline model.

Experiment setup: The models were implemented in Pytorch and trained using the AdaBelief optimization algorithm. The workstation used for the experiments has an i7 CPU @ 2.3 GHz, 16 GB RAM and NVIDIA GeForce GTX 3060. The dataset was split as 60% for training, 20% for validation and 20% for testing.

Quantitative and qualitative analysis: We adopt standard evaluation metrics for image segmentation, and report results in terms of mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), accuracy (AC) and precision (P) (Table 3).

Table 3 Experimental results of segmentation of COVID-19 lesion areas

Technique results	
U-Net	mIoU: 83.36%, DSC: 91.48%, AC: 99.44%, P: 91.44%
UNETR++	mIoU: 85.16%, DSC: 89.95%, AC: 99.31%, P: 91.02%
DAE-Former	mIoU: 86.45%, DSC: 90.93%, AC: 98.34%, P: 90.16%

The results show that although U-Net has higher accuracy, its mIoU is the lowest among the three models, which indicates that U-Net obtains smaller segmented lesion area. As shown in Fig. 6, U-Net discarded part of the lesion area, which resulted in the segmented lesion area being smaller than the other two models.

Subsequently, we changed the noise rate and noise type, and the segmentation performance of the three models all deteriorated. As shown in Fig. 7, the mIoU value of UNETR++ dropped more than that of the other two models. It indicates that its anti-noise ability is inferior to the other models.

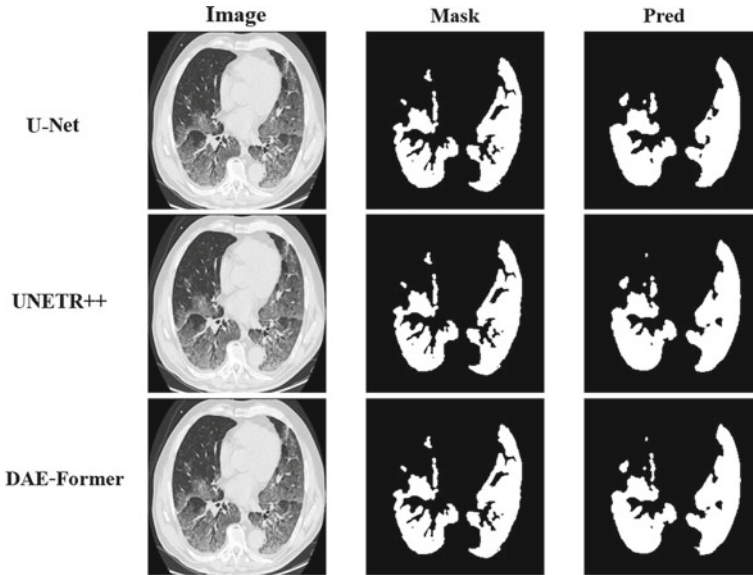


Fig. 6 Comparison of segmentation results from different models

Fig. 7 Comparison of the mIoU value of different models with or without noise

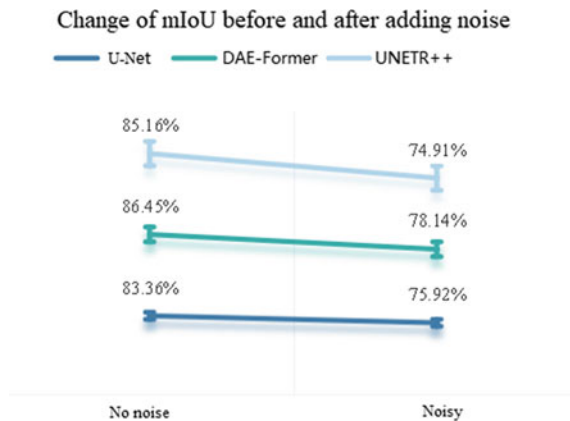


Table 4 Comparison of segmentation performance of different models on Synapse

Technique	#Params [M]	DSC \uparrow (%)	HD \downarrow
nnFormer [30]	150.5	86.57	10.63
TransUNet [40]	105.28	77.48	31.69
Swin-Unet [33]	27.17	79.13	21.55
UNETR++ [32]	42.96	87.22	7.53
DAE-Former [31]	48.1	82.43	17.46

In addition to the dataset presented in this paper, the common Synapse dataset, consisting of multiple organ segmentation datasets, was widely used for the evaluation of image segmentation. In Table 4, the results of several models are added for comparison, and the results show that UNETR++ and DAE-Former achieve better segmentation performance with a low parameter number.

UNETR++ obtains a DSC value of 87.22% on the Synapse dataset, while reaching 89.95% of DSC on the COVID-19 dataset, which may be caused by the fact that the Synapse dataset is more diverse in terms of types of organs. DAE-Former behaves similarly and obtains the DSC value of 82.43% and 90.93% on the Synapse and COVID-19 dataset respectively. This comparison also shows that UNETR++ is more robust in terms of domain transfer.

4 Conclusion

We review recent deep learning models for the task of image segmentation, most of which performed well in automatic diagnosis of COVID-19, i.e., accurate identification of lung infections. We also apply two state-of-the-art Transformer-based models (DAE-Former and UNETR++) to segment COVID-19 lesion areas. The experimental results show that for the task of segmenting COVID-19 lesion areas, both Transformer-based models obtained better performance in terms of the value of mIoU compared with U-Net, a CNN-based model, while U-Net obtained higher accuracy. The DAE-Former model has superior anti-noise ability to UNETR++, whereas UNETR++ is more robust in terms of domain transfer. In future, we would experiment with generative adversarial network (GAN) as well as geometric data enhancement technologies to enrich and preprocess datasets, in order to generate data with high quality and quantity, which may further improve the performance of the system.

References

1. Feng, C., et al.: Age and sex differences among mildly symptomatic and asymptomatic patients with omicron infection in 2022 in Shanghai, China. *J Multi. Healthc.* **15**, 1909–1919 (2022)
2. Silva, S.J.R.D., et al.: Recent insights into SARS-CoV-2 omicron variant. *Rev. Med. Virol.* e2373
3. Tian, D., et al.: The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. *J. Med. Virol.* **94**(6), 2376–2383 (2022)
4. He, J.-L., et al.: Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. *Respir. Med.* **168**, 105980 (2020)
5. Tahamtan, A., Ardebili, A.: Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev. Mol. Diagn.* **20**(5), 453–454 (2020)
6. Rahimzadeh, M., Attar, A., Sakhaei, S.M.: A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset. *Biomed. Signal Process. Control* **68**, 102588 (2021)
7. Bullock, J., et al.: Mapping the landscape of artificial intelligence applications against COVID-19. *J. Artif. Intell. Res.* **69**, 807–845 (2020)
8. Shi, F., et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **14**, 4–15 (2021)
9. Vineth Ligi, S., et al.: Radiological analysis of COVID-19 using computational intelligence: a broad gauge study. *J. Healthc. Eng.* **2022**, 5998042 (2022)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional Networks for Biomedical Image Segmentation. Springer, Berlin
11. Zheng, C., et al.: Deep learning-based detection for COVID-19 from chest CT using weak label. *MedRxiv* (2020)
12. Gaál, G., Maga, B., Lukács, A.: Attention u-net based adversarial architectures for chest x-ray lung segmentation. *arXiv preprint arXiv:2003.10304* (2020)
13. Naqvi, S.A.J., et al.: Improved lung segmentation based on U-Net architecture and morphological operations. *arXiv preprint arXiv:2210.10545* (2022)
14. Qi, X., et al.: Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. *MedRxiv* (2020)
15. Wang, G., et al.: A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Trans. Med. Imaging* **39**(8), 2653–2663 (2020)
16. Mishra, D., et al.: Ultrasound image segmentation: a deeply supervised network with attention to boundaries. *IEEE Trans. Biomed. Eng.* **66**(6), 1637–1648 (2018)
17. Wang, W., et al.: PSTCNN: explainable COVID-19 diagnosis using PSO-guided self-tuning CNN. *Biocell* **47**(2), 373–384 (2023)
18. Ren, Z.: LCDAE: data augmented ensemble framework for lung cancer classification. *Technol. Cancer Res. Treat.* **21** (2022). <https://doi.org/10.1177/15330338221124372>
19. Wang, J.: Artificial intelligence for visually impaired. *Displays* **77**(Article ID: 102391) (2023)
20. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *IEEE* (2016)
21. Çiçek, Ö., et al.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Springer, Berlin (2016)
22. Isensee, F., et al.: NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021)
23. Tilborghs, S., et al.: Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients. *arXiv preprint arXiv:2007.15546* (2020)
24. Feyjie, A.R., et al.: Semi-supervised few-shot learning for medical image segmentation. *arXiv preprint arXiv:2003.08462* (2020)

25. Raghu, M., et al.: Do vision transformers see like convolutional neural networks? *Adv. Neural. Inf. Process. Syst.* **34**, 12116–12128 (2021)
26. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)* (2020)
27. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 568–578 (2021)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. <https://arxiv.org/abs/2103.14030v2> (2021)
29. Hatamizadeh, A., Yang, D., Roth, H., Xu, D.: UNETR: Transformers for 3d medical image segmentation. <https://arxiv.org/abs/2103.10504> (2021)
30. Zhou, H.-Y., et al.: nnformer: interleaved transformer for volumetric segmentation. *arXiv preprint [arXiv:2109.03201](https://arxiv.org/abs/2109.03201)* (2021)
31. Azad, R., et al.: DAE-Former: dual attention-guided efficient transformer for medical image segmentation. *arXiv preprint [arXiv:2212.13504](https://arxiv.org/abs/2212.13504)* (2022)
32. Shaker, A., et al.: UNETR++: delving into efficient and accurate 3D medical image segmentation. *arXiv preprint [arXiv:2212.04497](https://arxiv.org/abs/2212.04497)* (2022)
33. Cao, H., et al.: Swin-UNET: Unet-like pure transformer for medical image segmentation. *arXiv preprint [arXiv:2105.05537](https://arxiv.org/abs/2105.05537)* (2021)
34. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: attention with linear complexities. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp 3530–3538. <https://doi.org/10.1109/WACV48630.2021.00357> (2021)
35. Ali, A., et al.: Xcit: cross-covariance image transformers. *Adv. Neural. Inf. Process. Syst.* **34**, 20014–20027 (2021)
36. Morozov, S.P., et al.: Mosmeddata: chest ct scans with covid-19 related findings dataset. *arXiv preprint [arXiv:2005.06465](https://arxiv.org/abs/2005.06465)* (2020)
37. Ma, J., et al.: Toward data-efficient learning: a benchmark for COVID-19 CT lung and infection segmentation. *Med. Phys.* **48**(3), 1197–1210 (2021)
38. MEDSEG. COVID-19. 2020. Available from: <http://medicalsegmentation.com/covid19/>
39. Soares, E., et al., SARS-CoV-2 CT-scan dataset: a large dataset of real patients CT scans for SARS-CoV-2 identification, p. 2020-04. *MedRxiv* (2020)
40. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation (2021). *arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306)*