ELSEVIER

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Biomedical Signal Processing and Control 80 (2023) 104366

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Cov-TransNet: Dual branch fusion network with transformer for COVID-19 infection segmentation

Yanjun Peng, Tong Zhang *, Yanfei Guo

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, 266590, Shandong, China

ARTICLE INFO

Keywords:
COVID-19
Deep learning
Transformer
Semantic segmentation
Attention mechanism
CT images

ABSTRACT

Segmentation of COVID-19 infection is a challenging task due to the blurred boundaries and low contrast between the infected and the non-infected areas in COVID-19 CT images, especially for small infection regions. COV-TransNet is presented to achieve high-precision segmentation of COVID-19 infection regions in this paper. The proposed segmentation network is composed of the auxiliary branch and the backbone branch. The auxiliary branch network adopts transformer to provide global information, helping the convolution layers in backbone branch to learn specific local features better. A multi-scale feature attention module is introduced to capture contextual information and adaptively enhance feature representations. Specially, a high internal resolution is maintained during the attention calculation process. Moreover, feature activation module can effectively reduce the loss of valid information during sampling. The proposed network can take full advantage of different depth and multi-scale features to achieve high sensitivity for identifying lesions of varied sizes and locations. We experiment on several datasets of the COVID-19 lesion segmentation task, including COVID-19-CT-Seg, UESTC-COVID-19, MosMedData and COVID-19-MedSeg. Comprehensive results demonstrate that COV-TransNet outperforms the existing state-of-the-art segmentation methods and achieves better segmentation performance for multi-scale lesions.

1. Introduction

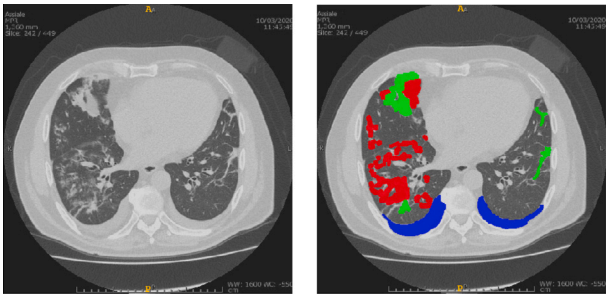
COVID19 is an infectious disease caused by the SARS-CoV-2 virus. Since the disease spread rapidly and caused a high number of fatalities, it has been declared a global pandemic by the World Health Organization. It is essential to detect the disease as early as possible in order to be able to treat it better. Chest CT, a traditional non-invasive imaging technique with high precision and speed, has been proven to have high sensitivity to diagnose COVID-19. Almost all patients with COVID-19 have characteristic CT features during the course of the disease [1–3]. Ground glass opacity (GGO), consolidation, and pleural effusion (PE) are the most common types of lesions. And the CT manifestations of them are shown in Fig. 1. Therefore, visual qualitative analysis based on chest CT images is expected to accurately assess the severity of COVID-19 and guide the clinical treatment combined with clinical information [4,5].

However, rapid detection, pathological tissue location and manual labeling are laborious duties for radiologists, extending the time required to analyze CT images. Therefore, the automatic annotation of COVID-19 infection areas in CT images is of great significance. Deep learning methods have fundamentally changed feature selection and extraction, and gradually become an effective technology to detect chest abnormalities and pathology [6–8].

Recently, deep learning algorithm has provided a unique solution for identifying COVID-19 in clinical chest CT images [9–11]. Wang et al. [12] designed a new priority attention strategy to extend residual learning and coupled two three-dimensional ResNets into a single model. Their network can effectively identify COVID-19 in chest CT volumes. Ardakani et al. [13] used ten well-known convolutional neural networks to identify COVID-19 pneumonia in clinical CT images. They concluded that ResNet-101 can be used as an auxiliary tool in the radiology department. Hu et al. [14] used deep supervised method to learn the multi-scale features of imbalanced clinical data. An adaptive auxiliary loss algorithm was designed based on the integration of effective sample numbers and a weighted regularization item. It performs well in imbalanced data of varying degrees. Zhou et al. [15] decomposed three-dimensional segmentation problem into three two-dimensional ones, and developed a novel data enhancement module to simulate the evolution of infection, which significantly solved the problem of insufficient training data. Fan et al. [16] designed a novel parallel decoder to combine high-level features and obtain a global map. Then, the reverse attention mechanism and the edge attention

* Corresponding author.
E-mail address: t_1998zhang@163.com (T. Zhang).

https://doi.org/10.1016/j.bspc.2022.104366
Received 23 May 2022; Received in revised form 6 September 2022; Accepted 30 October 2022
Available online 8 November 2022
1746-8094/© 2022 Elsevier Ltd. All rights reserved.



(a) Original CT image

(b) Labeled CT image

Fig. 1. A sample of CT image with different lesion type. (Red color for the GGO lesions, green color for the consolidation lesions and blue color for the PE lesions).

mechanism were used to strengthen the edge feature representation. Due to less effective labeled data, they used semi-supervised method to further improve the model learning ability and achieved higher performance. Mu et al. [17] presented a new multi-scale and multi-level feature aggregation method, and the multi-monitoring method was used to introduce more boundary information to improve segmentation performance.

Nevertheless, current work is difficult to achieve satisfactory results in multi-scale COVID-19 infection location and segmentation, due to the intricate distribution of lesions, the low contrast between infected regions and other organisms, and noise interference in CT images. Moreover, manual labeling is subjective. Therefore, automatic segmentation of COVID-19 lung infection regions is still hard work. Considering the above issues, our contributions are as follows:

1. We propose a dual branch fusion network named COV-TransNet to automatically segment the new coronavirus infection regions in CT images. The network fully utilizes the advantage of Convolutional Neural Network (CNN) and Transformer. Experimental results in different datasets suggest that the proposed model can achieve advanced performance.
2. We improve the convolution layer with feature activation module, so that more important features can be retained in the sampling process.
3. A novel multi-scale attention mechanism is introduced to make the network more capable of generating context feature representations and adaptively refining feature maps.
4. In order to evaluate the effectiveness of incorporating each module, we have conducted ablation experiments. ROC curves and PR curves also have been plotted for a better visual comparison.
5. A pre-processing method for removing non-human parts from CT images is proposed, which will narrow the focus of interest and facilitate model training.
6. As a demonstration of the robustness of our proposed model, we have participated in Kaggle's "COVID-19 CT Images Segmentation" challenge and achieved an F1 score of 0.69452.

2. Related works

2.1. Transformer in computer vision

The Transformer architecture first appeared in the field of Natural Language Processing (NLP) [18]. Recently, it has been applied to Computer Vision (CV). Dosovitskiy et al. [19] proposed a pure transformer model called Vision Transformer (ViT), which is proven to work well with multiple datasets. Zheng et al. [20] presented the SETR model for the segmentation task of urban street scenes. It is allowed to model the global context in each layer of the whole. Recently, many medical image segmentation tasks have also incorporated the transformer architecture. Chen et al. [21] proposed TransUNet, which offers the advantages of transformer and U-Net. The introduction of the transformer structure makes up for the limitation of U-Net in modeling long-range dependencies. Since the transformer structure introduced in this model is implemented based on ViT, its advantages still need to be highlighted on models trained on large datasets. Valanarasu et al.[22] proposed the MedT model, which introduced a gated axial attention model to extend the existing architecture. We also introduce the transformer structure into the proposed network to capture long-range dependence for the task of COVID-19 infected region segmentation from chest CT slices.

2.2. Attention mechanism

With the development of artificial intelligence, the attention mechanism has been widely used in a variety of tasks in the fields of CV and NLP. It makes the network focus on the target area, so that more detailed information about the target can be obtained, as well as other useless information can be suppressed. At present, the attention mechanism is being researched by more scholars. Hu et al. [23] proposed the SE block, which uses the interdependence between convolution feature channels to improve the representation quality of the model. Based on SE block, a variant form called sSE block was proposed to add some spatial information by Roy et al. [24]. The scSE block was also proposed in this work which can integrate more spatial and channel information into the feature map. Similarly, for the sake of considering the information of the two dimensions of space and channel, Woo et al. [25] presented CBAM. This Module can generate attention maps along channel and spatial, then multiplies the attention maps to the input feature map for adaptively refining feature. Considering the recent related research, scholars are more inclined to use the attention mechanism to fuse information in channel and spatial dimensions to refine features so that the model is more robust to noise.

2.3. Dilated convolution

Dilated convolution was originally proposed for the task of image segmentation. Common image segmentation algorithms typically use pooling and convolution layers to increase the receptive field while also reducing the resolution of the feature map. Then we need to frequently use sampling to resize the feature map, causing a loss of accuracy. To avoid this problem, Yu et al. [26] proposed the concept of dilated convolution, which can support the exponential growth of the receptive field without loss of resolution. Subsequently, the ASPP [27] based on dilated convolution appeared, which can robustly segment multi-scale targets. Due to the different scale targets in the task of COVID-19 infection region segmentation, we also introduce a paralleled multi-scale module in the proposed model to segment multi-scale targets better.

3. Methods

In this study, the end-to-end COV-TransNet is proposed, which takes advantage of UNet [28] and Transformer. It can make full use of the global information to extract features related to semantic segmentation and directly use them for segmentation. A multi-scale attention mechanism is designed to effectively expand the receptive field while selecting more powerful feature representation. In addition, a feature activation module has been suggested and introduced into the proposed architecture.

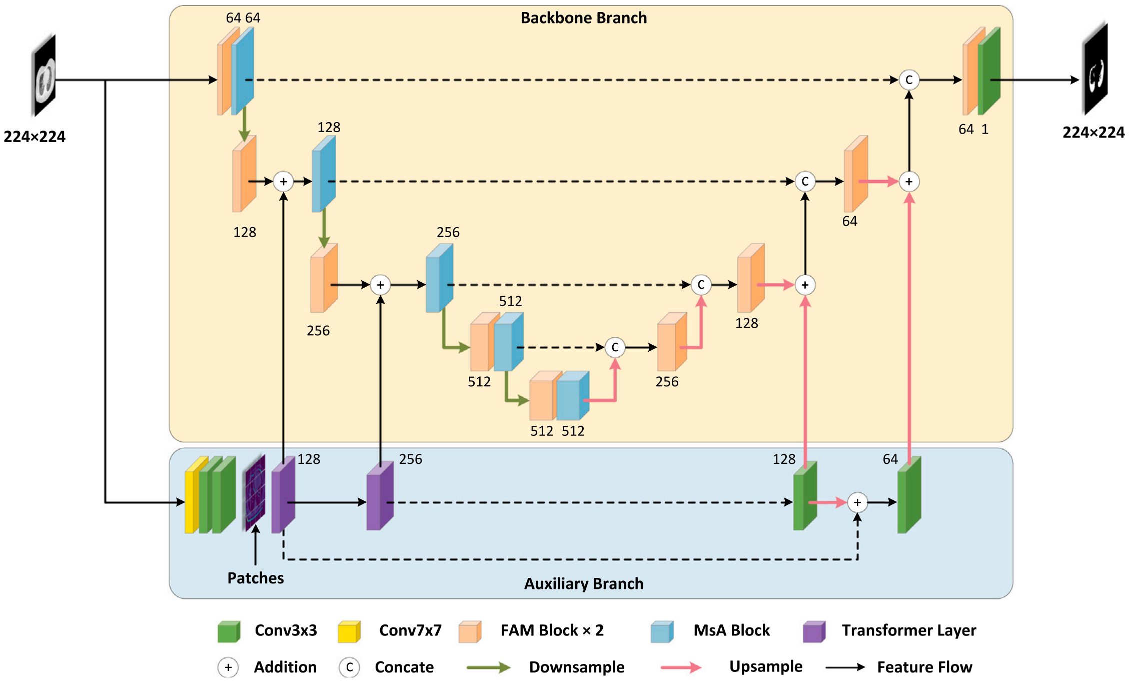


Fig. 2. The architecture of the proposed network with two branches. The network takes a CT slice as input and the visualized result is COVID-19 infection region.

3.1. Structure of the dual-branch fusion segmentation network

UNet provides a powerful method for medical image segmentation [28]. The encoder, decoder and skip connections form a U-shaped structure network. However, due to the intrinsic locality of convolution operations, UNet generally demonstrates limitations in explicitly modeling long-range dependency. Given that transformer has a global modeling capability, which relies on self-attention to capture long-range dependencies, we proposed COVTransNet with transformer.

The structure of the network is shown in Fig. 2. The backbone branch network inherits the basic mechanics of UNet. The encoder extracts high-level semantic features by stacking convolution layers, and gradually expands the receptive field. The decoder makes the feature map return to its original shape, at the same time, the corresponding semantic information of the encoding layer is added in the decoding process to reduce the semantic gap.

The auxiliary branch network is designed to serve as a global map provider benefitting from the transformer's natural advantages. The earlier generated global representation from the auxiliary branch is incorporated into the local feature extraction process of the backbone branch. In this way, the encoder of the backbone branch can pay more attention to long-range information, because of the global awareness, it also can capture better helpful local features and use them directly in semantic segmentation.

The backbone branch in Cov-TransNet uses convolution to extract features. To highlight the feature of COVID-19 infected areas as well as suppress the interference of irrelevant information, the network improves the continuous convolution of each layer by activating feature maps. A novel attention mechanism designed to enlarge the receptive field and refine the feature map adaptively. The output is a feature map with multi-scale feature information weighted by attention, guiding the network to capture different scales of infection lesions.

The auxiliary branch network uses a two-layer transformer as the encoder to obtain the feature map with extensive global feature information. The global feature information is incorporated into the backbone branch network without losing spatial information. More complete instance-related information will be retained during sampling owing to the guidance of global relationships. The auxiliary branch uses the conventional linear interpolation method for upsampling. The final output is based on the outputs of the two branches.

3.2. Feature Activate Module

The convolution can obtain different types of image features, making the image gradually abstracted into a concept representation with high-level semantic information. However, there may be a large amount of redundancy in feature maps, and there is a loss of spatial information during the sampling process, which may not be conducive to accurate segmentation. Therefore, the intrinsic features of targets can be further highlighted by enhancing the effective feature expression and suppressing the ineffective feature expression. In order to provide more useful semantic information, we have improved the convolution layer in the network within introducing the Feature Activation Module (FAM) which is shown in Fig. 3.

FAM activates the feature map after convolution so that more effective information on the feature map is enhanced, which means important features can be maintained. The loss of spatial information in the next down-sampling process can be reduced. Besides, the high-level semantic information spread to the low-level more effectively.

A 3 × 3 convolution is used for feature extraction. And then Sigmoid is conducted to obtain a feature vector which is used to activate the original feature map. Additionally, a 1 × 1 convolution layer is introduced to further selectively retain important feature representations and suppresses useless feature representations. Finally, the activated feature map is added to the original feature map. The skip connection in this block is used to alleviate the gradient vanishing problem caused by the sigmoid operation.

The FAM module for a certain feature map is expressed as:

$$FAM(X) = f_1(\sigma(f_1(f_3(X))) \times f_1(f_3(X))) + f_1(f_3(X)) \tag{1}$$

where X is the input feature, $FAM(X)$ is the activated feature map, f_n represents the convolution operation of kernel size n, σ represents the sigmoid activation. "×" and "+" represent the element multiplication and addition operations of the two tensors, respectively.

3.3. Multi-scale attention

The basic intention of the Multi-scale Attention (MsA) mechanism structure is to make the network learn important features and

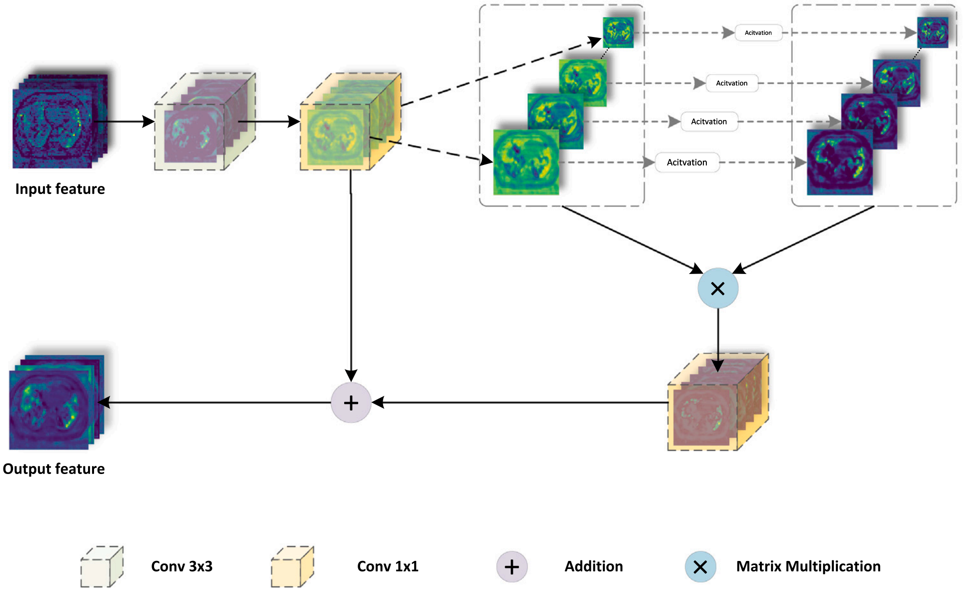


Fig. 3. The proposed Feature Activate Module, which is utilized to strengthen important features to retain them during sampling.

strengthen the multi-scale feature representation adaptively, so as to obtain a detailed feature map with more abundant contextual information. Inspired by Polarized Self-Attention [29] and Inception [30], we design a parallel combined structure of different size convolution kernels and different rate dilated convolutions. The result obtained in Ref. [29] demonstrates that it can maintain a relatively high internal resolution during the computation processing. As shown in Fig. 4, a dual-branch structure is introduced for channel and spatial attention calculations. A branch containing softmax is used to calculate channel or spatial importance information, while the other branch of reshape images is used to maintain high resolution. Finally, an additional branch is used to impose channel or spatial attention on a multi-scale integrated feature map.

The Channel Attention can be formulated as:

$$Output_{ch}(X) = \sigma(f_1(R(f_1(X))) \times \tau(R(f_1(X)))) \tag{2}$$

where X is the input feature, $Output_c h(X)$ is the output feature map in the channel dimension, f_n represents convolution operation of kernel size n, σ represents the sigmoid activation, R represents reshape operation, τ represents $SoftMax$ activation.

The Spatial Attention can be formulated as:

$$Output_{sp}(X) = \sigma(R(\tau(R(Pool(f_1(X)))) \times R(f_1(X)))) \tag{3}$$

where $Output_{sp}(X)$ is the output feature map in the spatial dimension, and $Pool$ represents the global pooling layer.

Finally, the Multi-scale Attention can be expressed as:

$$\begin{aligned} Output_{ms}(X) = & f_1(Cat(f_1(X), f_3^{r=3}(f_3(f_1(X))), \\ & f_5^{r=5}(f_5(f_1(X))))) + X \end{aligned} \tag{4}$$

$$\begin{aligned} Output(X) = & Output_{sp}(X) \times Output_{ms}(X) \\ & + Output_{ch}(X) \times Output_{ms}(X) \end{aligned} \tag{5}$$

where $Output_{ms}(X)$ is the output feature map of multi-scale, $Output(X)$ is the final output feature map, f_n^r represents dilated convolution operation of kernel size n and rate r, Cat represents concatenate operation.

3.4. Transformer branch

CNN has a powerful ability to capture the invariant law of local feature translation. Despite it has good performance, current local feature extraction still has problems. If we want to further improve the accuracy and positional precision of the segmentation, the network needs to encode the global positional presentation at an early stage, avoiding additional steps to help modify inaccurate local representations. Considering that a single convolution layer cannot capture long-distance information, the long-range feature capture capability of CNN is often obtained by stacking depth. However, the stacked convolution layer is still significantly weaker than Transformer in this respect. To alleviate the problem of the increased number of parameters and calculations caused by the transformer structure, we use only two transformer layers in the encoder and decoder. We adopt the Transformer structure in Fig. 5.

The axial attention mechanism proposed in Ref. [31] is applied in our Transformer structure, which improves the self-attention mechanism. Height-Axis represents axial attention layer with positive encodings along the height axis, Width-Axis represents axial attention layer with positive encoding along the width axis, and an axial attention layer broadcasts information along a specific axis. Multi-head attention mechanism was applied in both axial attention layers. The use of two axial attention layers on both the height and width axes not only captures the global information adequately, but also reduces computational effort and increases computational efficiency. The attention layer on the height axis and width axis can be defined as one-dimensional position sensitive self-attention respectively, which are defined as follows:

$$y_i^h j = \Sigma_{h=1}^{H} softmax(q_{ij}^T k_{hj} + q_{ij}^T r_{hj}^q + k_{hj}^T r_{hj}^k)(v_{hj} + r_{hj}^v) \tag{6}$$

$$y_{ij}^w = \Sigma_{w=1}^{W} softmax(q_{ij}^T k_{iw} + q_{ij}^T r_{iw}^q + k_{iw}^T r_{iw}^k)(v_{iw} + r_{iw}^v) \tag{7}$$

Where y_{ij}^w, y_{ij}^h respectively represents Height-Axis and Width-Axis self attention. The equations in Eqs. (6) and (7) follow the axial attention model with positive encodings proposed in work [31].

4. Experiment and results

4.1. Dataset

Currently, there are four datasets widely used by researchers in the COVID-19 infection region segmentation task, including COVID-19-CT-Seg [32], UESTC-COVID-19[33], MosMedData [34] and COVID-19-MedSeg [35]. CO-VID-19-CT-Seg consists of CT scans of 20 patients diagnosed with COVID-19. The left lung, right lung and infected areas

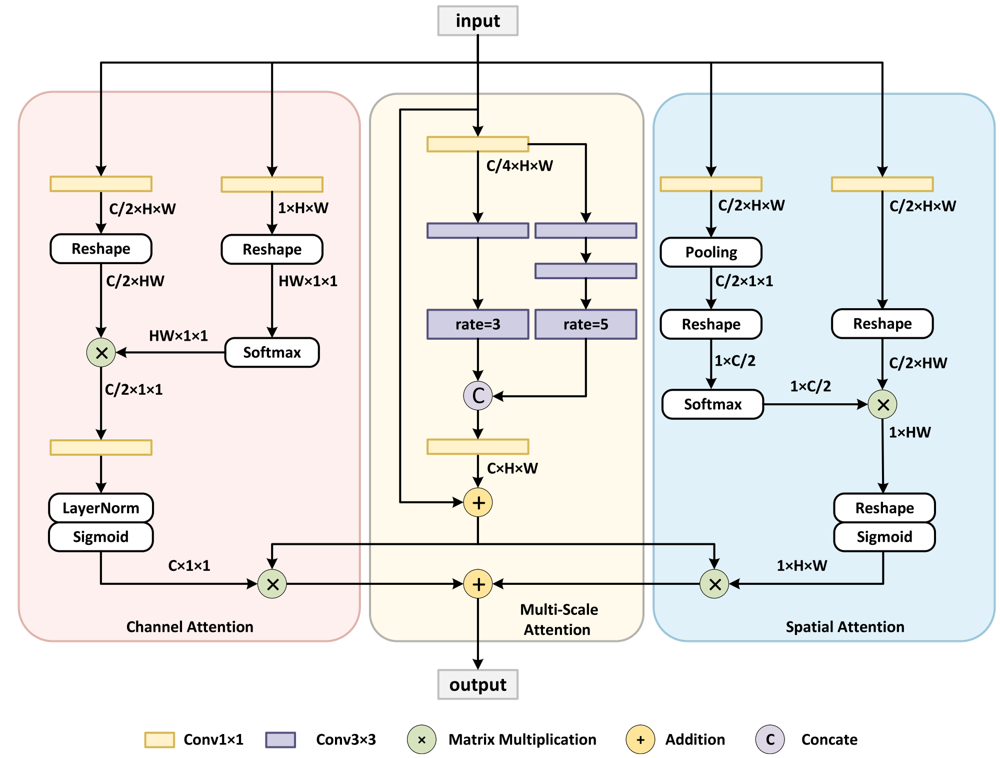


Fig. 4. Multi-scale attention calculates feature importance in channel and spatial dimensions to highlight the useful regions. Refined feature maps of multiple scales are fused together. The shape of tensors is shown, where C presents channel number, H presents the height and W presents the width.

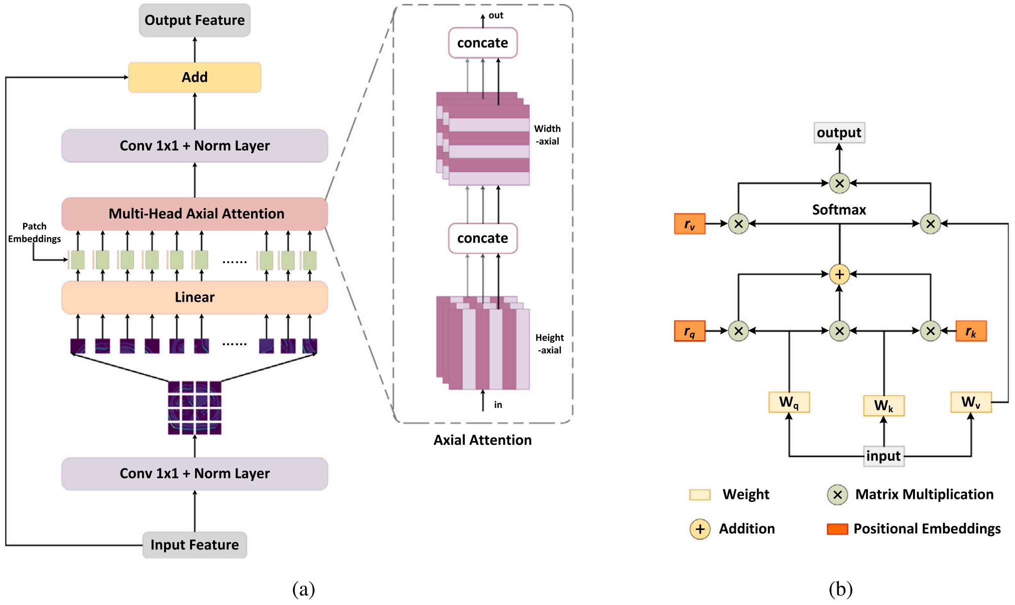


Fig. 5. (a) The main architecture of the transformer layer which is used in the auxiliary branch. (b) Axial Attention layer with positional embeddings applied along the width-axis.

were annotated at the pixel level by experienced radiologists. UESTC-COVID-19 consists of 120 labeled COVID-19 CT scans, of which 50 patients' CT volumes were annotated by experts and others were annotated by no-experts. We used the CT scans labeled by experts as part of our dataset. MosMedData was published by the Moscow Health Care Department in Russia and was distributed into 5 categories according to the degree of infection. But only the CT scans of 50 patients with mild symptoms in this dataset have been annotated, which were also included as part of our dataset. COVID-19-MedSeg is a collection of over 100 axial CT slices from more than 40 patients, which have been pre-processed with grayscale. The CT slices were annotated by radiologists with 3 labels: ground-glass (mask value =1), consolidation (=2) and pleural effusion (=3). The statistics of slices with infected areas in the dataset are shown in Fig. 6 directly.

We divided each dataset into a training set, a validation set and a test set. And then we chose to combine all the divided training and validation sets and mix them into a larger and more feature-rich training set, so that the model can fully learn more features of the infected regions. Considering that the data in COVID-19-MedSeg consists of over 100 CT slices from more than 40 patients, we cannot

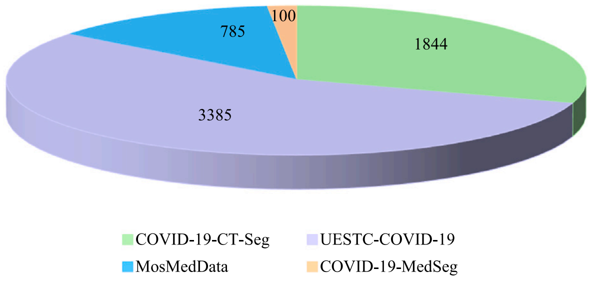


Fig. 6. The specific number of slices in the four datasets.

determine which CT images correspond to the same patient. In order to prevent the CTs of the same patient from being divided into the training set and the test set at the same time, which may lead to leakage of test data and affect the real experimental results, we chose to use the entire dataset as a part of the training dataset. The trained model was tested on the 3 divided test datasets, named COVID-19-CT-Seg-Test, UESTC-COVID-19-Test, and MosMedData-Test.

4.2. Dataset pre-processing

For the convenience of training model, we considered having a central labeling of the mask value of COVID-19 infections. The infected region was set as the foreground area (mask value =1) and the other region was set as background area. All CT slices except COVID-19-MedSeg have been normalized them into [0,1] using window width/level of 1500/-650. Since the CT slices in COVID-19-MedSeg have been grayscaled, we directly normalized to [0,1]. There are a lot of noisy information generated during the scanning process which can make an impact on the performance of the trained model. To avoid these effects, we decided to focus on the human part as the region of interest. For this purpose, we designed an algorithm described as follows. Firstly, the OSTU algorithm [36,37] is used to calculate a threshold(The calculated threshold is different for each CT axial slice). And we binarize the CT slice according to this threshold, setting the pixel values greater than this threshold to the foreground area and those less than or equal to this threshold to the background area. Then a binarized CT slice is obtained, which is shown like Fig. 7(b). However, some non-human parts will also be set to the foreground area. Since these parts occupy far fewer pixels than the human part and are non-connected with the human part, the objects can be removed easily. And then a human area mask with holes is obtained, as shown in Fig. 7(c). After filling these holes, we get a mask of the segmented body part, as shown in Fig. 7(d). It is easy to get the ROI by multiplying this mask image with the original CT image. The result of this algorithm is shown in Fig. 7(e).

4.3. Evaluation metrics

To evaluate the performance of the model in segmenting the COVID-19 infections of the lung, we measured the dice similarity coefficient (DSC), intersection over union (IoU), sensitivity, specificity and precision. DSC and IoU are commonly used to evaluate performance in medical image segmentation. Sensitivity and specificity are introduced to measure the ability of the model to distinguish positive and negative pixels. Precision reflects the ability of the model to predict accurately. These evaluation metrics are defined as follows:

$$DSC = \frac{2|G \cap P|}{|G| + |P|} \tag{8}$$

$$IoU = \frac{|G \cap P|}{|G| + |P| - |G \cap P|} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

where P and G represent the predicted and ground-truth masks, TP refers to the true positive, TN refers to the true negative, FP refers to the false positive, and FN refers to the false negative.

4.4. Experiment details

COV-TransNet was implemented based on the Pytorch framework. We used NVIDIA RTX 2080 Ti GPUs to accelerate our training process. Due to the limited computational resources, we unified the images to the size of 224 × 224. To further improve the generalization of the model, we performed random data augmentation on the training set by randomly cropping, flipping vertically or horizontally, and adding Gaussian noise. Data augmentation is implemented with the imgaug Python toolkit. In this work, we used the Adaptive Moment Estimation (Adam) optimizer [38] in the training process, with the initial learning rate set to 1e-4, the batch size set to 8 and the epoch set to 200. We updated the parameters of the model by the Tversky loss [39]. Tversky coefficient is a generalization of the DSC and IoU, and it can be defined as:

$$Tversky(P, G, \alpha, \beta) = \frac{|PG|}{|PG| + \alpha|P/G| + \beta|G/P|} \tag{13}$$

where P and G represent the predicted and ground-truth masks, α and β are parameters, which control the proportion of false negatives and false positives respectively. By adjusting α and β we can balance false positives and false negatives. In this work, we set α to 0.3 and β to 0.7. The generalized loss function is defined as follows:

$$L_{Tky}\alpha, \beta = \frac{\Sigma_{i=1}^{N} p_{0i} g_{0i}}{\Sigma_{i=1}^{N} p_{0i} g_{0i} + \alpha \Sigma_{i=1}^{N} p_{0i} g_{1i} + \beta \Sigma_{i=1}^{N} p_{1i} g_{0i}} \tag{14}$$

where p_{0i} is the probability that pixel i is the infected region and p_{1i} is the probability that pixel i is the normal region. Also, g_{0i} is 1 for an infected region and 0 for normal region and vice versa for g_{1i}. To enable better generalization of the model, we experimented with different combinations of parameters. Numerous experimental results show that setting β to 0.7 can achieve generalization and better performance on unbalanced data and effectively make the Network focus on reducing FNs and improving sensitivity.

4.5. Experiment results

4.5.1. Quantitative analysis

Several models were trained using the same approach as the comparison experiments, including UNet++[40], FD-UNet [41], TransUNet [21], UTNet [42] and some other networks [16,33,43–45] proposed for COVID-19 infections segmentation recently. To demonstrate that the proposed model is valid on different datasets, we test different models on three separated test datasets. The training hyperparameters of all compared models are the same as described in Section 4.4. The detailed experimental results are shown in Tables 1, 2, and 3. It can be seen clearly that our model outperforms the other compared models in evaluation metrics such as DSC, IoU and Sen. As can be seen from Table 1, COV-TransNet is about 4 percentage points higher than the highest results of other models used for segmentation of COVID-19 infections on DSC. Although UESTC-COVID-19 and MosMed consist of mild patients with fewer pixels marked as infected areas in slices, and the boundaries between infected areas and normal tissue are blurred, the evaluation of our model performs well in terms of quantitative analysis and evaluation metrics as shown in Tables 2 and 3. It is suggested that other models do not perform well on all test data. From these results, COV-TransNet not only has the most advanced results, but also has better generalization capabilities.

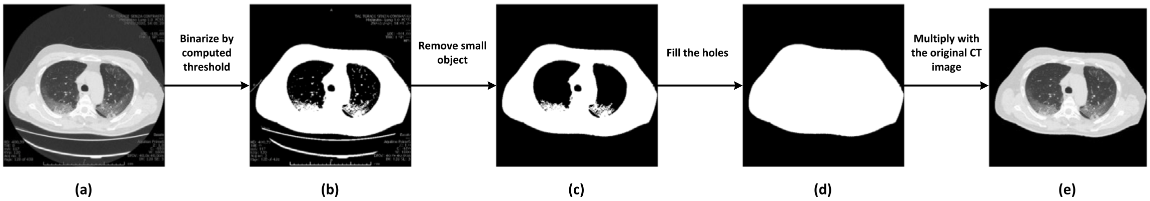
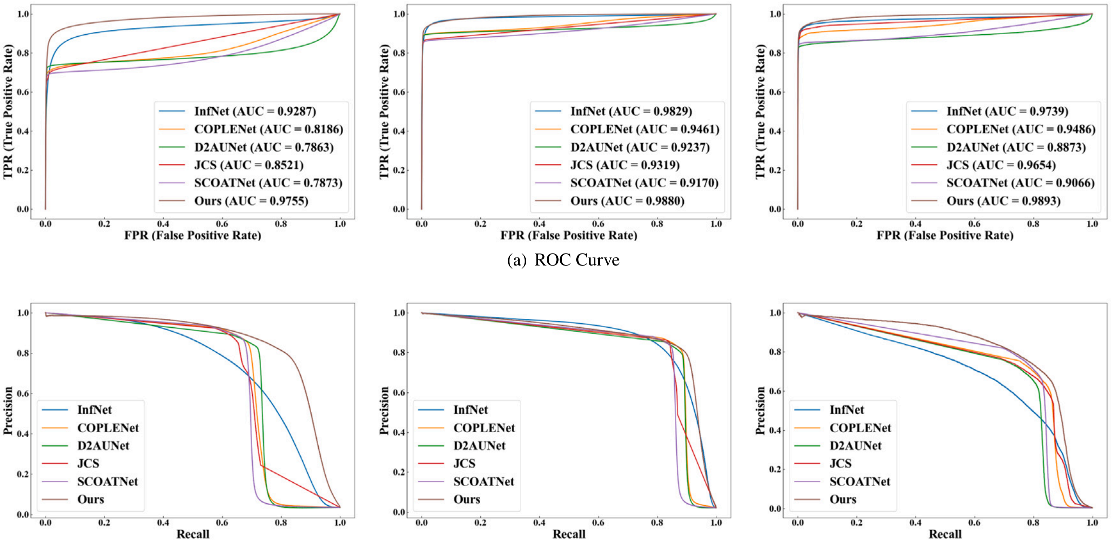

Binarize by computed threshold
Remove small object
Fill the holes
Multiply with the original CT image
(a)
(b)
(c)
(d)
(e)

Fig. 7. Visualization results during pre-processing, and (a) is the original image.



(a) ROC Curve

(b) PR Curve

Fig. 8. Performance comparisons of the proposed network with other models presented for the COVID-19 infection segmentation using COVID-19-CT-Seg-Test, UESTC-COVID-19-Test, MosMedData-Test.

Table 1
Comparison of the quantitative performance of 10 models in terms of different evaluation metrics in COVID-19-CT-Seg-Test. Sen is sensitivity, Spec is specificity and Prec is precision.

Model	DSC	IoU	Sen	Spec	Prec
UNet++[40]	0.758	0.613	0.671	0.997	0.876
FDUNet [41]	0.747	0.600	0.655	0.997	0.882
TransUNet [21]	0.721	0.568	0.617	0.997	0.883
UTNet [42]	0.749	0.603	0.654	0.997	0.885
InfNet [16]	0.737	0.583	0.630	0.997	0.892
CopleNet [33]	0.742	0.592	0.637	0.998	0.897
D2AUNet [43]	0.766	0.624	0.704	0.996	0.851
JCS [44]	0.746	0.600	0.675	0.995	0.838
SCOATNet [45]	0.737	0.586	0.626	0.998	0.903
Ours	0.803	0.673	0.772	0.995	0.842

Table 2
Comparison of the quantitative performance of 10 models in terms of different evaluation metrics in UESTC-COVID-19-Test. Sen is sensitivity, Spec is specificity and Prec is precision.

Model	DSC	IoU	Sen	Spec	Prec
UNet++[40]	0.846	0.738	0.845	0.997	0.852
FDUNet [41]	0.845	0.736	0.830	0.997	0.866
TransUnet [21]	0.822	0.702	0.837	0.996	0.814
UTNet [42]	0.845	0.737	0.854	0.997	0.841
InfNet [16]	0.845	0.743	0.840	0.997	0.860
CopleNet [33]	0.848	0.741	0.853	0.997	0.848
D2AUNet [43]	0.842	0.733	0.870	0.996	0.822
JCS [44]	0.845	0.738	0.840	0.997	0.856
SCOATNet [45]	0.840	0.729	0.833	0.997	0.853
Ours	0.855	0.751	0.874	0.996	0.842

Table 3
Comparison of the quantitative performance of 10 models in terms of different evaluation metrics in MosMedData-Test. Sen is sensitivity, Spec is specificity and Prec is precision.

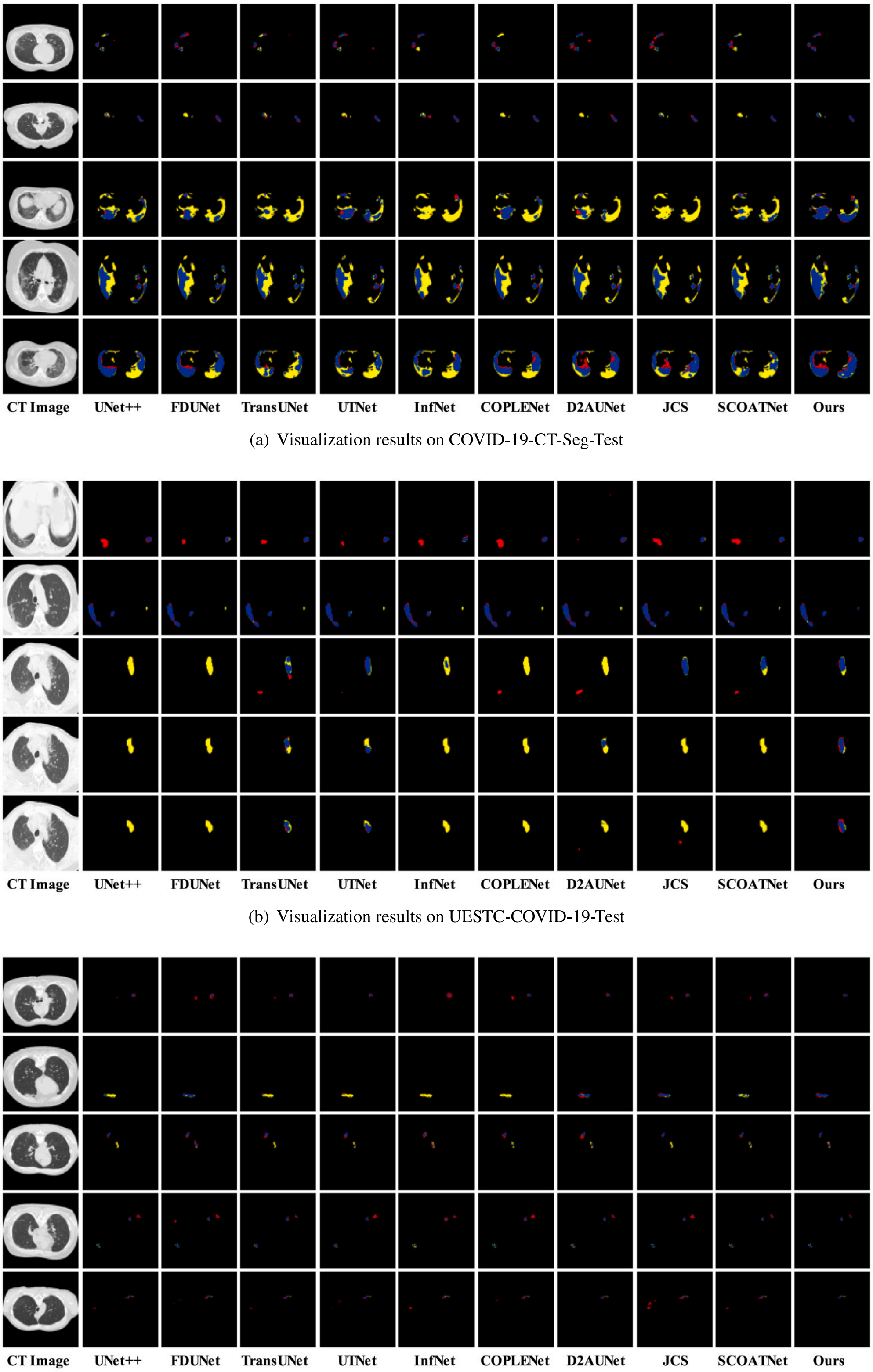
Model	DSC	IoU	Sen	Spec	Prec
UNet++[40]	0.704	0.552	0.745	0.999	0.671
FDUNet [41]	0.705	0.549	0.828	0.999	0.619
TransUnet [21]	0.701	0.549	0.735	0.999	0.675
UTNet [42]	0.726	0.572	0.816	0.999	0.661
InfNet [16]	0.704	0.541	0.735	0.999	0.670
CopleNet [33]	0.740	0.591	0.805	0.999	0.693
D2AUNet [43]	0.715	0.560	0.767	0.999	0.679
JCS [44]	0.717	0.566	0.793	0.999	0.663
SCOATNet [45]	0.741	0.596	0.763	0.999	0.731
Ours	0.751	0.607	0.800	0.999	0.717

In addition, we also plotted ROC curves and PR curves on test sets. As shown in Fig. 8, the curve of the proposed model consistently lies above most other models, the proposed model's AUCs achieve the highest results, which indicate that COV-TransNet has the smallest error in its predictions compared to the ground truth.

A further comparison between the tested models and ours in terms of parameters (Params) and floating point operations per second (FLOPs) is presented in Table 4. In order to be fair, they were compared under the same hardware and software settings. The results demonstrate that our model improves the segmentation performance without consuming too much time and space.

4.5.2. Qualitative analysis

For making the performance of the proposed model more intuitive, we visualize the prediction results of the model. Since we performed a two-class segmentation task, we labeled the GT and predicted results



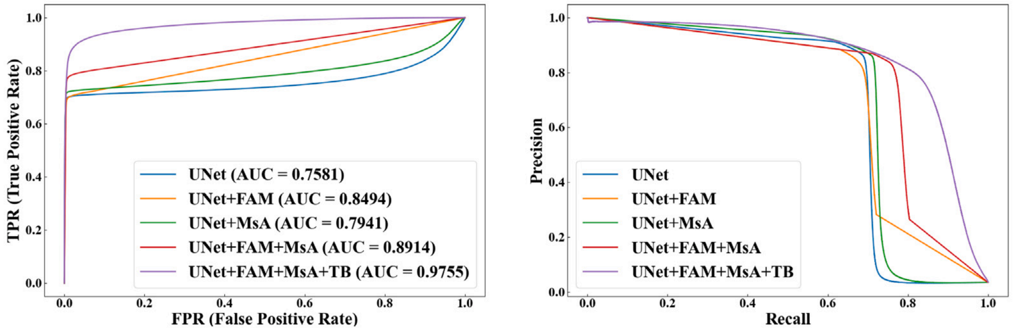
(a) Visualization results on COVID-19-CT-Seg-Test

(b) Visualization results on UESTC-COVID-19-Test

(c) Visualization results on MosMed-Test

Fig. 9. Visual comparison of lung infection segmentation with different models. The blue and yellow areas indicate detected and undetected true infected areas respectively. Red areas indicate false infected areas that were detected in error.

in the same image. The blue and yellow areas indicate undetected and detected true infected areas respectively. Red areas indicate false infected areas that were detected in error. According to Fig. 9, the predicted results of our model are most similar to the ground truth. For the small infection areas, our model can easily predict and achieve the highest accuracy in terms of location and similarity. The ability to capture more accurate location information is due to the addition of auxiliary branches that allow the network to capture long-range



(a) ROC Curve (b) PR Curve

Fig. 10. Performance comparisons of the proposed modules for the COVID-19 infection segmentation using COVID-19-CT-Seg-Test.

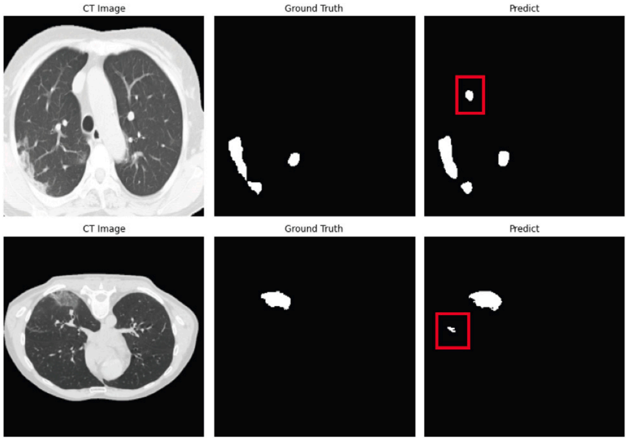


Fig. 11. Some instances of the failure cases for our proposed method.

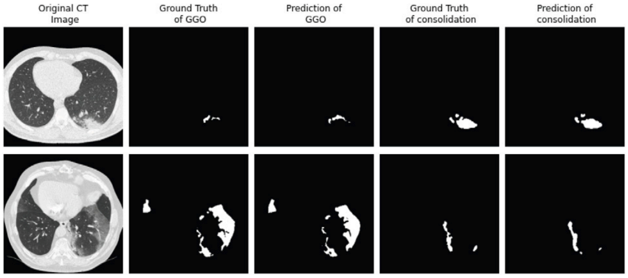


Fig. 12. A visual presentation of multiclass segmentation prediction results.

Table 4
Comparison of the quantitative performance of 10 models in terms of Params and FLOPs.

Model	Params (M)	FLOPs (G)
UNet++[40]	9.163	26.668
FDUNet [41]	11.998	16.920
TransUnet [21]	66.770	24.797
UTNet [42]	13.117	10.012
InfNet [16]	29.951	5.237
CopleNet [33]	42.056	33.924
D2AUNet [43]	42.201	64.118
JCS [44]	45.199	56.311
SCOATNet [45]	10.209	29.992
Ours	18.987	45.592

Table 5
Ablation analysis of proposed model, where FAM denotes Feature Activate Module, MsA denotes Multi-scale Attention, TB denotes the Transformer Branch. Sen is sensitivity, Spec is specificity and Prec is precision.

No.	Model	DSC	IoU	Sen	Spec	Prec
1	UNet [28]	0.736	0.585	0.662	0.995	0.839
2	UNet+FAM	0.759	0.615	0.681	0.996	0.866
3	UNet+MsA	0.763	0.621	0.713	0.995	0.830
4	UNet+FAM +MsA	0.772	0.633	0.712	0.996	0.852
5	UNet+FAM +MsA+TB	0.803	0.673	0.772	0.995	0.842

dependencies from the first layer. And for the larger scale targets, the proposed model is also able to segment them more approximately than compared models. The capability to accurately segment targets at different scales is attributed to the introduction of the MsA. The combination of FAM with MsA contributes to the accuracy of our model in segmenting the infection boundary. The visualization results from UESTC-COVID-19-Test and MosMedData-Test show that in cases where the sizes of infection areas are extremely small, with careful differentiation, we can also see that the proposed model achieves precise positioning and segmentation results.

4.5.3. Ablation experiments

To prove that is effective for us to the components added in COV-TransNet, several ablation experiments are carried out in COVID-19-CT-Seg-Test, due to it is feature-rich. The specific experimental results are shown in Table 5. We take the two experiments including No. 1 and No. 2 as the baseline. And then the MsA module is introduced to the base experiments. From the experimental results we can see that both DSC and IoU are improved in appearance, which proves that the MsA module helps us to improve the performance of the model. By the same method, the effectiveness of FAM also can be manifested. The combination of MsA and FAM enables a large increase in the capability of the model. After introducing the Transformer Branch, the performance of the proposed model is even more boosted. This also indicates that the transformer architecture has great potential in the field of medical image segmentation.

ROC curves and PR curves are also plotted in the ablation experiments, as depicted in Fig. 10. It is evident that the addition of all modules contributes to the effectiveness of the proposed model.

4.5.4. Failure cases

Some failure cases also appear in our prediction results. Two failure cases have been shown in Fig. 11, and the part of incorrect prediction pixels has been labeled with a red box. The possible reason is that this part shows medical imaging features extremely similar to COVID19 in CT images. Despite there are a few incorrect predictions, our model is still able to segment out the main infected regions.

4.5.5. Robustness verification experiments

To demonstrate the robustness of our proposed model, we participated in Kaggle's "COVID-19 CT Images Segmentation" challenge. The model was trained based on the dataset given in the challenge. Before the model training, we preprocessed the dataset with the approach mentioned in Section 4.2. The shape of CT slice was resized into a size of 256 × 256. Due to the memory size limitation of the GPU, we adjusted the batch size to 4. Finally, we obtained an F1-score of 0.69452 on the leader board as evaluated by the organizers on a separate test dataset. To better illustrate the prediction effect of the model, we also visualized the prediction results, as shown in Fig. 12. According to the experimental results, our proposed model has a strong robustness.

5. Conclusion

In this paper, we proposed a novel dual branch fusion network to segment COVID-19 infection areas from chest CT images, which take full advantage of the global map and local feature information. This network utilizes a parallel structure multi-scale attention module to improve the performance of identifying multiple sizes of infection. Moreover, feature activation module was designed to alleviate the effective information loss during sampling. Also, the preprocessing method effectively reduces the noise of CT images. Extensive experiments on several real CT datasets have demonstrated that our proposed methods perform better than the cutting-edge segmentation networks and improve the most advanced performance. In future work, we aim to further focus on the improvement of the performance of the task of multi-class segmentation. Beyond that, we will continue to optimize our segmentation framework and try to apply it to 3-dimensional data and other related work.

CRediT authorship contribution statement

Yanjun Peng: Conceptualization, Supervision, Investigation, Resources, Formal analysis, Project administration, Methodology, Funding acquisition, Writing – review & editing. **Tong Zhang:** Investigation, Methodology, Software, Data curation, Validation, Writing – original draft, Writing – review & editing. **Yanfei Guo:** Formal analysis, Writing – original draft, Writing – review.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2022.104366.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No. 61976126, Shandong Natural Science Foundation, China under Grant No. ZR2019MF003.

References

[1] F. Pan, T. Ye, P. Sun, et al., Time course of lung changes at chest CT during recovery from Coronavirus disease 2019 (COVID-19), Radiology 295 (3) (2020) 715–721, http://dx.doi.org/10.1148/radiol.2020200370.

[2] Y. Pan, H. Guan, S. Zhou, et al., Initial CT findings and temporal changes in patients with the novel Coronavirus pneumonia (2019-nCoV): A study of 63 patients in Wuhan, China, Eur. Radiol. 30 (6) (2020) 3306–3309, http://dx.doi.org/10.1007/s00330-020-06731-x.

[3] H. Shi, X. Han, N. Jiang, et al., Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study, Lancet Infect. Dis. 20 (4) (2020) 425–434, http://dx.doi.org/10.1016/s1473-3099(20)30086-4.

[4] P. Lyu, X. Liu, R. Zhang, et al., The performance of chest CT in evaluating the clinical severity of COVID-19 pneumonia: Identifying critical cases based on CT characteristics, Invest. Radiol. 55 (7) (2020) 412–421.

[5] Z.Y. Zu, M.D. Jiang, P.P. Xu, et al., Coronavirus disease 2019 (COVID-19): A perspective from China, Radiology 296 (2) (2020) E15–E25.

[6] Y. Xie, Y. Xia, J. Zhang, et al., Knowledge-based collaborative deep learning for Benign-Malignant lung nodule classification on chest CT, Ieee Trans. Med. Imag. 38 (4) (2019) 991–1004, http://dx.doi.org/10.1109/tmi.2018.2876510.

[7] N. Lessmann, C.I. Sanchez, L. Beenen, et al., Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence, Radiology 298 (1) (2021) E18–E28, http://dx.doi.org/10.1148/radiol.2020202439.

[8] F. Shan, Y. Gao, J. Wang, et al., Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction, Med. Phys. 48 (4) (2021) 1633–1645, http://dx.doi.org/10.1002/mp.14609.

[9] T.A. Soomro, L. Zheng, A.J. Afifi, et al., Artificial intelligence (AI) for medical imaging to combat Coronavirus disease (COVID-19): A detailed review with direction for future research, Artif. Intell. Rev. (2021) 1–31, http://dx.doi.org/10.1007/s10462-021-09985-z.

[10] Y. Bouchareb, P.M. Khaniabadi, F. Al Kindi, et al., Artificial intelligence-driven assessment of radiological images for COVID-19, Comput. Biol. Med. (2021) 104665.

[11] M. Moezzi, K. Shirbandi, H.K. Shahvandi, et al., The diagnostic accuracy of artificial intelligence-assisted CT imaging in COVID-19 disease: A systematic review and meta-analysis, Inform. Med. Unlocked 24 (2021) 100591, http://dx.doi.org/10.1016/j.imu.2021.100591.

[12] J. Wang, Y. Bao, Y. Wen, et al., Prior-attention residual learning for more discriminative COVID-19 screening in CT images, IEEE Trans. Med. Imaging 39 (8) (2020) 2572–2583, http://dx.doi.org/10.1109/TMI.2020.2994908.

[13] A.A. Ardakani, A.R. Kanafi, U.R. Acharya, et al., Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks, Comput. Biol. Med. 121 (2020) 103795, http://dx.doi.org/10.1016/j.compbiomed.2020.103795.

[14] K. Hu, Y. Huang, W. Huang, et al., Deep supervised learning using self-adaptive auxiliary loss for COVID-19 diagnosis from imbalanced CT images, Neurocomputing 458 (2021) 232–245, http://dx.doi.org/10.1016/j.neucom.2021.06.012.

[15] L. Zhou, Z. Li, J. Zhou, et al., A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis, Ieee Trans. Med. Imag. 39 (8) (2020) 2638–2652, http://dx.doi.org/10.1109/tmi.2020.3001810.

[16] D.-P. Fan, T. Zhou, G.-P. Ji, et al., Inf-Net: Automatic COVID-19 lung infection segmentation from CT images, Ieee Trans. Med. Imag. 39 (8) (2020) 2626–2637, http://dx.doi.org/10.1109/tmi.2020.2996645.

[17] N. Mu, H. Wang, Y. Zhang, et al., Progressive global perception and local polishing network for lung infection segmentation of COVID-19 CT images, Pattern Recognit. 120 (2021) 108168, http://dx.doi.org/10.1016/j.patcog.2021.108168.

[18] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[20] S. Zheng, J. Lu, H. Zhao, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[21] J. Chen, Y. Lu, Q. Yu, et al., TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[22] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, 2021, arXiv e-prints, arXiv:2102.10662.

[23] J. Hu, L. Shen, S. Albanie, et al., Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023, http://dx.doi.org/10.1109/tpami.2019.2913372.

[24] A.G. Roy, N. Navab, C. Wachinger, Concurrent spatial and ‘Channel Squeeze & Excitation‘ in fully convolutional networks, in: 21st International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) / 8th Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM) / International Workshop on Computational Diffusion MRI, CDMRI, in: Lecture Notes in Computer Science, vol. 11070, Springer, 2018, pp. 421–429, http://dx.doi.org/10.1007/978-3-030-00928-1_48.

[25] S. Woo, J. Park, J.-Y. Lee, et al., CBAM: Convolutional Block Attention Module, in: 15th European Conference on Computer Vision, ECCV, in: Lecture Notes in Computer Science, vol. 11211, 2018, pp. 3–19, http://dx.doi.org/10.1007/978-3-030-01234-2_1.

[26] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2016, arXiv preprint arXiv:1511.07122.

[27] L.C. Chen, G. Papandreou, I. Kokkinos, et al., DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848, http://dx.doi.org/10.1109/TPAMI.2017.2699184.

[28] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[29] H. Liu, F. Liu, X. Fan, D. Huang, Polarized self-attention: Towards high-quality pixel-wise regression, 2021, arXiv preprint arXiv:2107.00782.

[30] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9, http://dx.doi.org/10.1109/cvpr.2015.7298594.

[31] H. Wang, Y. Zhu, B. Green, et al., Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 108–126.

[32] M. Jun, G. Cheng, W. Yixin, et al., COVID-19 CT Lung and Infection Segmentation Dataset, 2020, http://dx.doi.org/10.5281/zenodo.3757476, Zenodo.

[33] G. Wang, X. Liu, C. Li, et al., A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images, Ieee Trans. Med. Imag. 39 (8) (2020) 2653–2663, http://dx.doi.org/10.1109/tmi.2020.3000314.

[34] S. Morozov, A. Andreychenko, I. Blokhin, et al., MosMedData: Data set of 1110 chest CT scans performed during the COVID-19 epidemic, Digit. Diagn. 1 (1) (2020) 49–59, http://dx.doi.org/10.17816/dd46826.

[35] MedSeg, COVID-19 segmentation dataset, 2020, Artificial Intelligence A/S, URL http://medicalsegmentation.com/covid19/.

[36] N. Otsu, A threshold selection method from gray - scale histograms, IEEE Trans. SMC SMC-9 (1) (1979) 62–66.

[37] S. Wang, B. Kang, J. Ma, et al., A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19), Eur. Radiol. (2021) 1–9.

[38] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014.

[39] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: 8th International Workshop on Machine Learning in Medical Imaging, MLMI, in: Lecture Notes in Computer Science, vol. 10541, 2017, pp. 379–387, http://dx.doi.org/10.1007/978-3-319-67389-9_44.

[40] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2020) 1856–1867.

[41] S. Guan, A.A. Khan, S. Sikdar, et al., Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal, IEEE J. Biomed. Health Inf. 24 (2) (2020) 568–576.

[42] Y. Gao, M. Zhou, D.N. Metaxas, UTNet: A hybrid transformer architecture for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 61–71.

[43] X.Y. Zhao, P. Zhang, F. Song, et al., D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution, Comput. Biol. Med. 135 (2021) http://dx.doi.org/10.1016/j.compbiomed.2021.104526.

[44] Y.-H. Wu, S.-H. Gao, J. Mei, et al., JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation, IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 30 (2021) 3113–3126, http://dx.doi.org/10.1109/tip.2021.3058783.

[45] S. Zhao, Z. Li, Y. Chen, et al., SCOAT-Net: A novel network for segmenting COVID-19 lung opacification from CT images, Pattern Recognit. 119 (2021) http://dx.doi.org/10.1016/j.patcog.2021.108109.