Computers in Biology and Medicine 169 (2024) 107858

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

FDR-TransUNet: A novel encoder-decoder architecture with vision transformer for improved medical image segmentation

Zhang Chaoyang, Sun Shibao *, Hu Wenmao, Zhao Pengcheng

School of Information Engineering, HeNan University of Science and Technology, Luoyang, 471023, China

ARTICLE INFO

Keywords:
Medical image segmentation
U-Net
Vision transformer
COVID-19
Deep-supervision

ABSTRACT

The U-shaped and Transformer architectures have achieved exceptional performance in medical image segmentation and natural language processing, respectively. Their combination has also led to remarkable results but still suffers from enormous loss of image features during downsampling and the difficulty of recovering spatial information during upsampling. In this paper, we propose a novel encoder-decoder architecture for medical image segmentation, which has a flexibly adjustable hybrid encoder and two expanding paths decoder. The hybrid encoder incorporates the feature double reuse (FDR) block and the encoder of Vision Transformer (ViT), which can extract local and global pixel localization information, and alleviate image feature loss effectively. Meanwhile, we retain the original class-token sequence in the Vision Transformer and develop an additional corresponding expanding path. The class-token sequence and abstract image features are leveraged by two independent expanding paths with the deep-supervision strategy, which can better recover the image spatial information and accelerate model convergence. To further mitigate the feature loss and improve spatial information recovery, we introduce successive residual connections throughout the entire network. We evaluated our model on the COVID-19 lung segmentation and the infection area segmentation tasks. The mIoU index increased by 1.5 points and 3.9 points compared to other models which demonstrates a performance improvement.

1. Introduction

The COVID-19 epidemic has hit the world hard, leading to tremendous attention to medical imaging in the field of artificial intelligence (AI). This increased focus can be attributed to the critical role that medical imaging techniques such as X-ray and computed tomography (CT) play in the diagnosis and treatment of COVID-19 in clinical practice [1,2]. Therefore, the increasing application of medical image segmentation and diagnosis has many positive implications for the fight against COVID-19 [3], such as assisting radiologists in focusing on areas of infection and making diagnostic decisions.

The development of image segmentation epitomizes the history of Artificial Intelligence. The most representative CNN models are Fully Convolutional Networks (FCN) [4] and U-Net [5]. Their encoder-decoder architecture, augmented by multi-scale feature skip-connection, has become the de-facto standard for image segmentation tasks, especially in the medical domain, which suffers from the drawbacks of low contrast, low quantity, high noise, etc. [6]. Moreover, medical image segmentation is an end-to-end and pixel-wise task where each pixel has a corresponding label, and the label has the same size as the image [7]. The encoder-decoder structure is perfect for such tasks as it downsamples images from high to low resolution to extract image representations, and then upsamples the lowest-level abstract semantic features to recover image spatial feature information. The multi-scale feature skip connection between high-level feature maps from the encoder and low-level feature maps from the decoder allows the decoder to obtain more detailed spatial information, which enables precise pixel classification [5].

But two shortcomings limit the performance of CNNs in medical image segmentation. One of them is the consecutive downsampling, which leads to the loss of important texture information of the images and is rarely recovered during upsampling. In addition, CNNs tend to ignore the global relationships between pixels in images. These shortcomings are fatal for medical image segmentation, leading to incomplete organ segmentation and edge aliasing. The Transformer, originally proposed by Ref. [8] for natural language processing (NLP) tasks, has sparked a brand-new research direction in Deep Learning (DL). The self-attention mechanism in the Transformer allows the network to focus

* Corresponding author.
E-mail address: sunshibao@haust.edu.cn (S. Shibao).

https://doi.org/10.1016/j.compbiomed.2023.107858
Received 27 July 2023; Received in revised form 30 November 2023; Accepted 11 December 2023
Available online 14 December 2023

more on global relations and reach the state of the art in many NLP tasks. Therefore, several studies [9–12] have attempted to combine CNN with Transformer to alleviate suffering from these shortcomings [8]. Vision Transformer (VIT) [12] was the first successful hybrid CNN-transformer image classification model that broke the boundaries between computer vision (CV) and NLP. Furthermore, TransUNet [10], the first medical image segmentation model that merged UNet and Vision Transformer, established a powerful encoder that simultaneously extracts global and local image features and reached the state of the art in the field of medical image segmentation.

Our novel approach builds on TransUNet. In this paper, we proceed to alleviate the feature loss of downsampling and improve the recovery of spatiality and localization information, experimented on two COVID-19 medical image datasets, and outperformed related networks in terms of segmentation accuracy and quality. Our main contributions are summarized as follows:

1. We propose a novel medical image segmentation model, which has a flexibly adjustable encoder depth. The encoder consists of ViT and feature double reuse (FDR) blocks, FDR is indeed an amalgamation of concepts from Densenet and Resnet. Compared with Resnet50 in TransUNet, FDR has less feature loss during the downsampling process.
2. We retain and leverage the class-token sequence from ViT and serve as the image representation. We upsample class-token sequence and the original abstract image features through two independent expanding paths.
3. We aggregate the outputs of two expanding paths with a deep supervision strategy. The original image features will compensate for the non-discriminative and noisy localization result caused by one class-token sequence, and the class-token sequence will bring global semantic information of the image. Further, the deep supervision strategy will accelerate model convergence and alleviate the problem of gradient vanishing.
4. We introduced successive residual connections throughout the network from head to toe to further minimize image feature loss in the process of downsampling and upsampling.
5. We conduct comparative experiments and ablation studies on two public COVID-19 lung X-ray datasets. The results show that our model has better segmentation accuracy and generalization ability than most related works.

The rest of this paper is summarized as follows: In Section 2, we presented the related medical image segmentation works by pure CNN and hybrid architectures based on CNN and Transformer methods. In Section 3, we introduced the overview and modules of our model. In Section 4, we showed two COVID-19 datasets, training details, results, qualitative comparisons, and ablation studies. Finally, we gave the conclusions of this paper and prospects.

2. Related works

Pure CNN Methods. Medical image segmentation is an important branch of computer vision, and its terrific prospects have absorbed numerous researchers. The U-Net, presented by Ref. [5] is a widely-used symmetric standard encoder-decoder architecture with multi-scale feature skip-connection. It effectively captures context and localization information and achieves remarkable results in cell image segmentation. Inspired by U-Net, many successful improvements have been proposed [13,14]. For example, U-Net++ [15] redesigned the skip-connection pathways between the encoder and the decoder, which allows for generating full-resolution feature maps and adjustable network depth, and used the deep supervision strategy [16] between different depth outputs. Another notable contribution is Attention U-Net [17], which integrated a soft-attention gate strategy into the skip-connection path. The role of the attention gate mechanism allows the network to neglect irrelevant features and focus on the region of interest.

The above successful methods have also been applied in COVID-19 chest radiograph segmentation for diagnosis [18]. proposed Adaptive Activation Function-based U-Net (AAF-U-Net) chest X-ray segmentation method for the next COVID-19 classification task, resulting in better results from the CNN classifier [19]. proposed a modified version of standard U-Net architecture for lung X-ray segmentation and compared it with other image enhancement techniques. In contrast to high-quality CT lung images (3D), X-ray images have lower tissue contrast. Therefore, more studies [20–23] focus on CT lung image segmentation.

Hybrid Architecture based on CNNs and Transformer. The Vision Transformer (ViT) [12] pioneered the Transformer in computer vision by transforming 2D images into multiple feature sequences as input of the Transformer. Then, Chen et al. proposed TransUNet [10], which combined the U-Net and Vision Transformer in medical image segmentation for the first time. With the spread of the COVID-19 epidemic [24], proposed a novel Swin-UNet [34] that replaced certain modules of the original Swin-UNet and effectively improved segmentation results on COVID-19 CT datasets [25]. proposed a Swin-transformer-based U-shape network to improve the problem of long-range feature dependencies of CNN [26]. has very similar ideas to ours that proposed a densely connected Swin-UNet with multiscale information aggregation for medical image segmentation [27]. combined CNN and Transformer for segmentation of COVID-19 lesions with semi-supervised learning [28]. incorporates a lightweight ViT with a U-shaped network to extract multiscale and global image information [30]. proposed a lightweight CNN-transformer segmentation network that has only 1 M parameters [31]. presented a new framework named Transformers for Fully Convolutional denseNets (TFCNs) to tackle the problems of inherent magnification and distortion in medical images. TDD-UNet [32] combined the decoder of UNet with the multi-head self-attention (MHSA) in the Transformer to enlarge the receptive field of the network. Dhamija et al. [33] also combined the encoders of Transformers and UNet to extract the global and local image features. In Ref. [29], CNN-based blocks and ViT-based blocks were used separately to extract local features and global features, and then combine both features to obtain richer semantic information. This is not a segmentation model but a classification model for COVID-19 CT images. We classify it as related work because the feature extraction ideas are similar. All related works are summarized at Table 1.

Due to the inadequate quality of lung X-ray images, it is generally considered less sensitive and more difficult to segment, therefore, more studies have focused on chest CT, despite X-ray images serving as the standard initial imaging modality for the fight against COVID-19. Additionally, X-rays are more cost-effective and widely used in many developing countries and regions. Thus, segmenting chest X-rays has a significant impact on medical image fields. Our model is a CNN-transformer-based segmentation model of COVID-19 chest X-ray images and achieved satisfactory segmentation accuracy and generalization ability.

3. Method

Compared to TransUNet, our model has four design improvements. First, we replace the original ResNet50 block of TransUNet with the feature double reuse block, enhancing the utilization of image features. Second, we retain and leverage the class-token sequence in the Transformer encoder, serving as image representations, and introduce two independent expanding paths for class-token features and abstract image features to recover global semantic information and localization information. Third, we aggregate the outputs of the two expanding paths with a deep supervision strategy, which can combine the advantages of the class-token sequence and original image features. Finally, to minimize image feature loss, we introduce successive residual connections throughout the network from head to toe. For an overview of our model, please refer to Fig. 1.

Table 1
Related medical image segmentation works.

Pure CNN Methods	Modality	Highlights
U-Net [5]	Microscopy	De-facto standard U-shaped architecture
Recurrent residual U-Net [13]	Microscopy/ CT	U-Net with residual/recurrent residual units
U-Net3+ [14]	CT	Full-scale skip connection
U-Net++ [15]	Microscopy/ CT	Dense skip pathways. Adjustable network depths
Attention U-Net [17]	CT	U-Net with attention gates
AAF-U-Net [18]	X-Ray	COVID-19 segmentation and classification
Modified U-Net [19]	X-Ray	Compare other enhancement techniques with modified U-Net
Inf-Net [20]	CT	Lung infection area segmentation with less training data
U-Net/SegNet [21]	CT	Discuss the effect of U-Net and SegNet on COVID-19 dataset
3D V-Net with shape prior [22]	CT	Integrate 3D V-Net with shape priors
3D U-Net [23]	CT	An infection segmentation pipeline for COVID-19
Hybrid Methods		
TransUNet [10]	CT	The first Transformer based medical image segmentation
Novel Swin-UNet [24]	CT	Redesigned Swin-UNet and new loss function
SwinE-UNet3+ [25]	CT	Improve the long-range feature dependencies of CNN
DCS-UNet [26]	MRI	Multiscale information aggregation
Ccat-Net [27]	CT	Semi-Supervised based on hybrid CNN and Transformer
TranSegNet [28]	OCT	Combine a lightweight ViT with a U-shape network
FCF-Net [29]	CT	Hybrid CNN and ViT to COVID-19 classification
MMViT-Seg [30]	CT	A light weight CNN-transformer segmentation network
Tfcns [31]	CT	Proposed a general attention module
TDD-UNet [32]	CT/X-Ray	Hybrid UNet and multi-head self-attention (MHSA)
USegTransformer [33]	CT	Combine the encoders of UNet and Transformer

3.1. Feature double reuse block

Feature Double Reuse. Reference [35,36]. We integrated residual connection *(Eq. (1))* and dense connection *(Eq. (2))* as follows:

$$X_L = H_L(X) + X_{L-1} \tag{1}$$

$$X_L = H_L([X_0, X_1, \ldots X_{l-1}]) \tag{2}$$

$$X_L = H_L([X_0, X_1, ..., X_{l-1}]) + X_{L-1} \tag{3}$$

Where $H_L(\bullet) + X_{L-1}$ refers to the non-linear transformation and a shortcut connection that skips from X_0 to X_{l-1}. $[X_0, X_1, \ldots, X_{l-1}]$ denotes the concatenation of feature maps produced by different layers. The channel number of X_i is a hyperparameter that refers to the "growth rate" of the FDR block, and in our model, we set it to 32 and 64, the effects of different growth rates can be seen in Fig. 4. Further, each FDR block is the pre-activation architecture that sequentially consists of 1 × 1, 3 × 3, 1 × 1 convolution layers and corresponding BN and ReLU layers. The experiment has shown that the FDR block can effectively alleviate the influence of gradient vanishing and feature loss.

Encoder and Decoder. Our encoder consists of four Feature Double Reuse (FDR) modules and one Transformer (Vaswani) module. The former downsamples the image features from a 256 × 256 resolution to 16 × 16. Depending on different growth rates, each FDR module has a different number of blocks and parameters. We extract the output of the four modules to establish skip connections with corresponding upsampling features. The latter receives 16 × 16 image features and pads a class-token sequence, which represents the global semantic information of the entire image. Then, the Transformer module will output the class-token sequence and original abstract image features separately. The decoder has two independent expanding paths and can upsample abstract image features and class-token sequence to the original image size, respectively. Additionally, dense skip pathways are applied between each downsampling and upsampling layer, represented by the dotted line in Fig. 1.

3.2. Vision Transformer

Patch Embedding. The Transformer block accepts 1D sequences but not 2D images. Thus, the original 2D images are used as the input to the FDR block, and then the output image features of the FDR block are projected linearly into flattened 2D sequence patches $X_P \in \mathbb{R}^{N \times P^2 \bullet C}$ as the input of the transformer block. Where C is the channel number of the image whose resolution is (H, W), P is the resolution of each small patch, and N is the number of patches calculated according to the formula HW/P^2.

Learnable Class Embedding. Unlike TransUNet, we retain the learnable class-token sequence whose length is constant $D = P^2C$ across all transformer encoder layers. The class-token is considered as image representation and trained with the image patch features. To make the best use of the class-token sequence, we extracted it from the outputs of the transformer block and concatenated its N copies, then reshaped it from $Z_L \in \mathbb{R}^{N \times D}$ to $\frac{H}{P} \times \frac{W}{P} \times D$ as a branch of two expanding paths.

Position Embedding. Position embedding [37] also is a learnable sequence similar to class embedding, it contains positional information and is incorporated into the image patch features as follows:

$$z_0 = \left[\boldsymbol{x}_{\text{class}}; \boldsymbol{x}_p^1 \boldsymbol{E}; \boldsymbol{x}_p^2 \boldsymbol{E}; \cdots \boldsymbol{x}_p^N \boldsymbol{E}\right] + \boldsymbol{E}_{pos}, \mathbf{E} \in \mathbb{R}^{\left(P^2, C\right) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{4}$$

Transformer encoder. The Transformer encoder in our model follows the same structure as the original Vision Transformer (ViT) encoder, which consists of L layers. L is a hyperparameter, each layer contains a Multihead Self-Attention (MSA) block and an MLP block (Eq. *(5)*, *(6)*). The residual connection and dropout are applied after each block, and the layer normalization (LN) is in front. The MLP block consists of two linear layers and a GELU activation function applied between them. The formulas are as follows:

$$z'_\ell = \text{MSA}(LN(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1})) + LN\, z_{\ell-1}, \ell = 1 \ldots L \tag{5}$$

$$z_\ell = \text{MLP}\left(LN\left(z'_\ell\right)\right) + z'_\ell, \ell = 1 \ldots L \tag{6}$$

3.3. FDR-TransUNet

Deep Supervision. The class-token is used to represent the global semantic information of the entire image. It captures the overall features in the image, not just local features. This helps the model understand the content and context of the entire image. But only one class-token sequence will capture semantic information from the background and other irregular objects, and cause non-discriminative and noisy localization [38]. To precisely recover the semantic and spatial information of our model, we employed the deep supervision strategy [16] to aggregate the outputs of class-token sequence and image features upsampling. Adding a supervision path not only accelerates model convergence and alleviates the problem of gradient vanishing but also combines the advantages of class-token image representation and image features. Two independent expanding paths mean two model outputs, The parameters are updated by summing the losses of the two outputs, which ultimately leads to improved performance.

Residual Upsampling. The image segmentation task is pixel-wise and the corresponding label has the same resolution as the image. We have built two independent expanding paths into the decoder to upsample the image features and the class-token sequence. Each expanding path

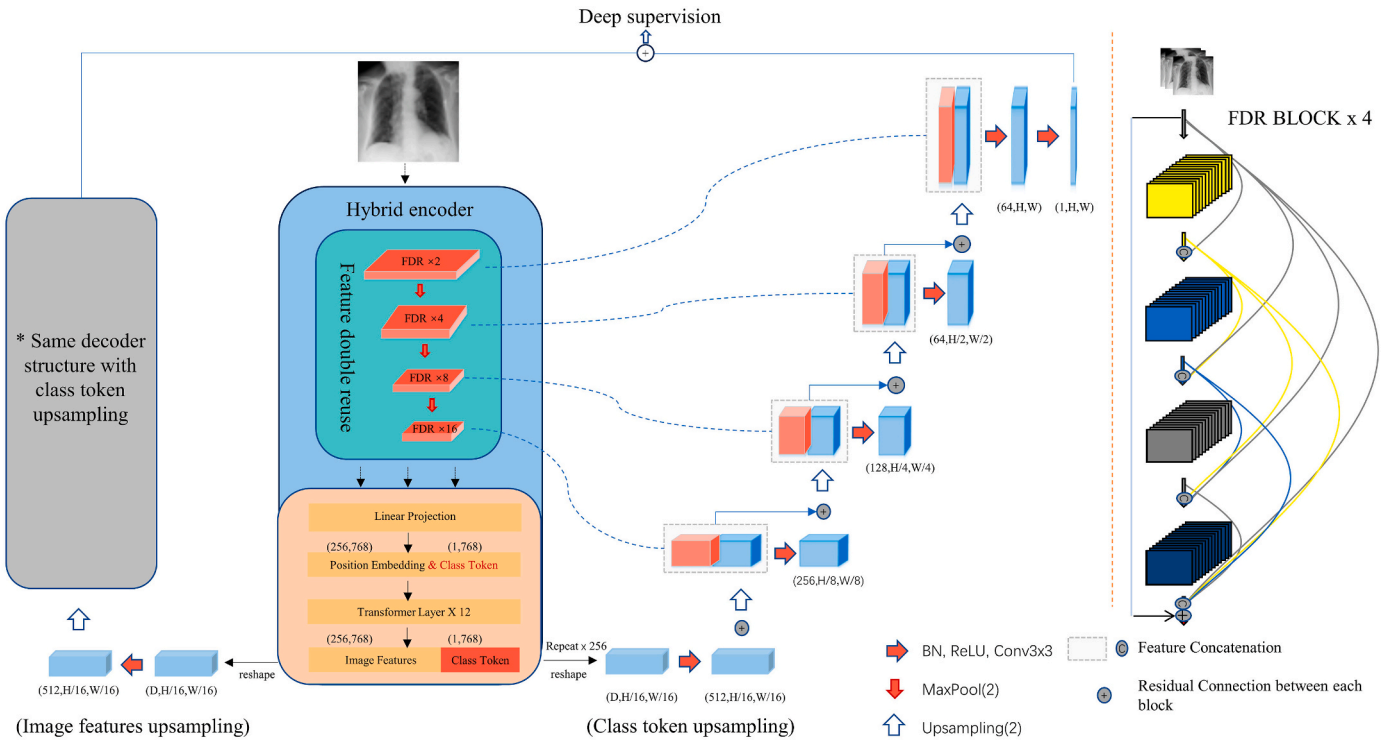

Fig. 1. The left part is the overview of FDR-TransUNet. The right part is the overview of the Feature Double Reuse block.

contains multiple modules, each module consisting of two consecutive 3 × 3 convolutions with ReLU activation function and 2-scale upsampling layers. Shortcut connections are used between each block to preserve important features and support upsampling.

4. Experiments

Datasets. We used two public datasets for our experiments. The first dataset is the COVID-19 Radiography Database,[1] which contains 4 classes: 3616 COVID-19 images, 10,192 normal images, 6012 non-COVID lung infection images, and 1345 viral pneumonia images, along with their respective masks. The images are 299 × 299 resolution with RGB channels, while the masks are 256 × 256 resolution with a single channel. We selected subsets from this dataset, containing 200 images for each class. The second dataset is the COVID-Qu-Ex Dataset,[2] which focuses on lung infection area segmentation, the images and masks are 256 × 256 resolutions. We utilized the COVID-19 infection segmentation data with 2913 images from this dataset. For our experiments, the datasets were divided into a 6:2:2 ratio for training, validation, and testing with a similar data distribution.

Training Details. In our model experiments, we used the Adam optimizer where the hyperparameters were defaulted, except for the learning rate and the weight decay. The learning rate was set to 3e-4 and applied the learning rate scheduler with cosine annealing in which T_0 equals 8 and T_mult equals 2. On the other hand, we used the BCE with Logits Loss as the loss function and set the batch size to 4. Each batch applied simple data augmentation, including Normalize, Random rotation (degree = 10), and random horizontal flip (p = 0.4) during training, but not during inference. During the training and validation processes, we used the mean Intersection over Union (mIoU) of the validation set to supervise the parameter updates rather than loss values. Additionally, we implemented an early-stop-count with a patience threshold of more than 10 epochs of no improvement in mIoU to halt the training. The pixel classification threshold was set to 0.5, where predicted pixels greater than or equal to 0.5 were designated as 1, and pixels less than 0.5 were labeled as 0. In addition, we examined the effects of the experimental results of using 32 and 64 growth rates in the FDR block. All experiments were conducted using a Nvidia GeForce GTX 3090.

As Table 2 shows, our model performed best compared to other models in the COVID-19 chest X-ray segmentation task. The mIoU index increased by 1.5 points and 3.9. These superior results benefit from the feature double reuse block, which can effectively reduce the loss of image features in the process of the encoder. On the other hand, abundant low-resolution features and the two independent expanding paths with deep supervision strategy can recover more accurate localization information. Comparing FDR-TransUNet32 and FDR-TransUNet64, the latter has slightly better results and fewer parameters. The larger growth rate has fewer FDR block depths based on the same input and output.

To demonstrate the generalization ability of our model, we also tested it on the COVID-19 infection segmentation dataset, as illustrated in Table 3. In many cases, lung infection segmentation is a class-imbalanced task, with the lung infection region being a small portion. The mIoU (mean Intersection over Union) represents the average

Table 2
The experimental results on the Covid Radiography Database. FDR-TransUNet32 means that the growth rate is 32.

Method	Lung Segmentation Task				
	mIoU	Dice	Precision	Recall	Paras
U-Net	0.861	0.966	0.979	0.954	28.9 M
UNet++	0.862	0.967	0.983	0.953	47.1 M
UNet3+	0.871	0.968	0.966	0.970	25.4 M
Swin-UNet	0.831	0.959	0.962	0.958	27.2 M
Atten-UNet	0.866	0.968	**0.985**	0.953	40.1 M
TransUNet	0.885	0.971	0.980	0.963	105.3 M
FDR-TransUNet32	0.898	0.974	0.973	**0.974**	104.3 M
FDR-TransUNet64	**0.900**	**0.975**	0.981	0.969	101 M

[1] https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.
[2] https://www.kaggle.com/datasets/anasmohammedtahir/covidqu.

Table 3
The experimental results on the infection segmentation subset of the Covid-Qu-Ex Dataset.

Method	Infection Segmentation Task				
	mIoU	Dice	Precision	Recall	Paras
U-Net	0.724	0.960	0.961	0.960	28.9 M
UNet++	0.715	0.959	0.965	0.955	47.1 M
UNet3+	0.709	0.957	0.955	0.961	25.4 M
Swin-UNet	0.659	0.951	0.961	0.944	27.2 M
Atten-UNet	0.711	0.960	0.973	0.949	40.1 M
TransUNet	0.710	0.959	0.963	0.957	105.3 M
FDR-TransUNet32	0.694	0.961	**0.974**	0.950	104.3 M
FDR-TransUNet64	**0.731**	**0.962**	0.965	**0.962**	101 M

Intersection over Union (IoU) for all classes, and it can be influenced by the number and size of classes. Our model focuses on infection region, and will ignore some normal lung tissue and other chest organs. Thus, the predictions on the normal lung tissue and other chest organs, which are the majority, have a significant impact on mIoU, leading to a lower mIoU percentage. As a result, the lung infection segmentation task yields a lower mIoU result.

The Dice coefficient is used to measure the accuracy of segmentation by calculating the ratio of the intersection between the predicted results and the ground truth to their union. Dice is more sensitive in handling class imbalance based on its definition, making it particularly suitable for evaluating the segmentation performance of minority classes. The results in Table 3 show that our model outperforms other models and demonstrates good generalization ability.

To demonstrate the improved segmentation results of our model intuitively, we conducted qualitative comparisons with COVID-19, lung opacity, normal and viral pneumonia images and shown in Fig. 2. From right to left are original images, ground truth, FDR-TransUNet with 64 growth rate, FDR-TransUNet with 32 growth rate, TransUNet, UNet++, U-Net and Attention UNet. The qualitative comparison results clearly demonstrate that our model produces a more flexible lung margin curve and more lung tissue than other existing models.

We also introduce the Receiver Operating Characteristic (ROC) curve to assess model performance. It measures the True Positive Rate and False Positive Rate, helping to evaluate the classification performance of our model. Fig. 3 evaluates the capabilities of related models based on COVID-19 lung segmentation dataset and COVID-19 infection segmentation dataset.

To further demonstrate the effectiveness of the FDR block and class-token sequence, we visualize the intermediate feature maps and the output feature maps of the model generated by FDR-TransUNet and TransUNet in Fig. 4. The lowest-feature maps (16 × 16 resolution) are placed under the original image for picture typesetting optimization.

The FDR block has 4 intermediate feature maps and the Resnet50 block in the TransUNet has 3 intermediate feature maps. First, we can see that all the intermediate feature maps of both FDR-TransUNet focus more on the lung area and produce highlighted lung pixel values than TransUNet. Even on the lowest abstract image features (the middle and bottom of a column), our model can extract lung features. This shows that the FDR block can retain more image features than the Resnet50 block in the encoder process. The e column indicated that the output feature maps of both FDR-TransUNet generate more abundant lung edge tissue, which benefits from the extra upsampling path of the class-token sequence.

Ablation Studies. We conducted ablation studies on the infection segmentation dataset built upon TransUNet. The results were shown in Table 4.

The original TransUNet consists of Resnet50 block, ViT block, and Cascaded Upsampler (CUP) decoder block (Resnet50-ViT-CUP). First, we use the Feature Double Reuse (FDR) block with different growth rate to replace the original Resnet50 module. We can find that FDR32-VIT-CUP and FDR64-VIT-CUP improve the mIoU index by 1.6 and 1.2 points, respectively, compared to TransUNet. On the other hand, we keep the Resnet50 and CUP block, add a class-token sequence in VIT and the corresponding expanding path (CEP), the mIoU increases by 0.6 points. Finally, to demonstrate the effectiveness of the global residual, we add the residual connection in two expanding paths based on CEP and obtain a 1.0 mIoU improvement (CEPR).

The investigation of growth rate. The feature double reuse block has adjustable depth through different growth rates and leads to a different number of parameters. We further investigate the influence of growth rate on the COVID-19 lung infection segmentation dataset. The x-axis is the number of growth rates. As Fig. 5 shows, a larger growth rate has fewer parameters, but a too-large growth rate gives poor results because too large growth rate leads to the loss of important image features. 64 is a suitable choice of growth rate according to our experimental results.

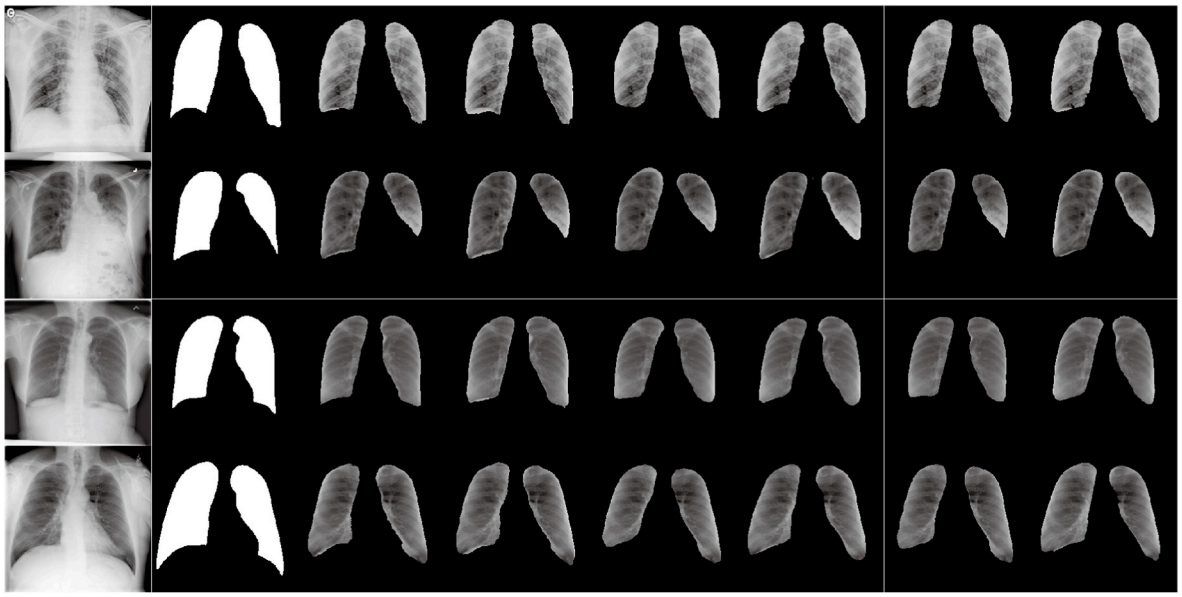

Fig. 2. Qualitative comparison based on (a) Image, (b) Ground Truth, (c) FDR-64, (d) FDR-32, (e) TransUNet, (f) UNet++, (g) U-Net, (h) Attention UNet from left to right. COVID, lung opacity, normal and viral pneumonia from top to bottom.

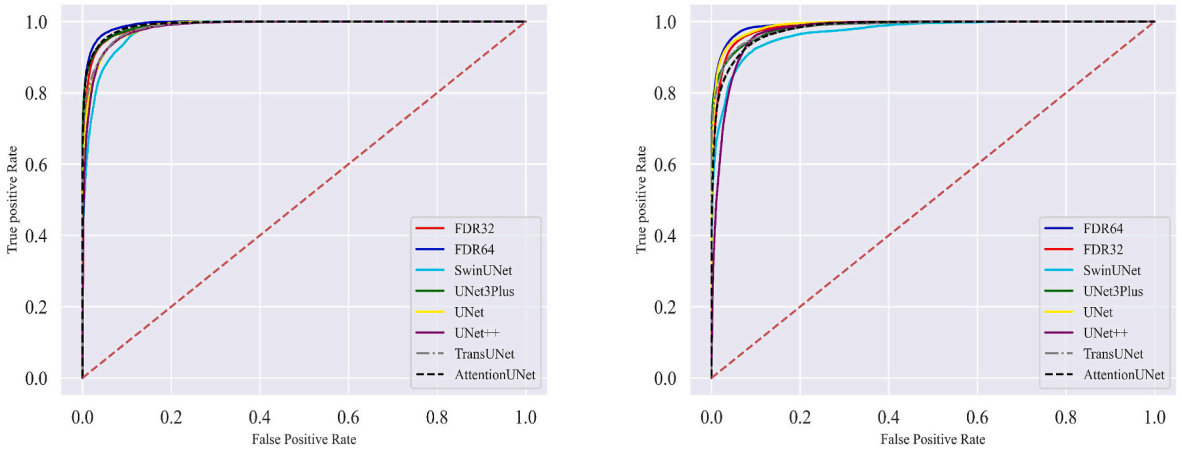

Fig. 3. ROC curves based on COVID-19 lung segmentation dataset (Left) and COVID-19 infection segmentation dataset (Right).

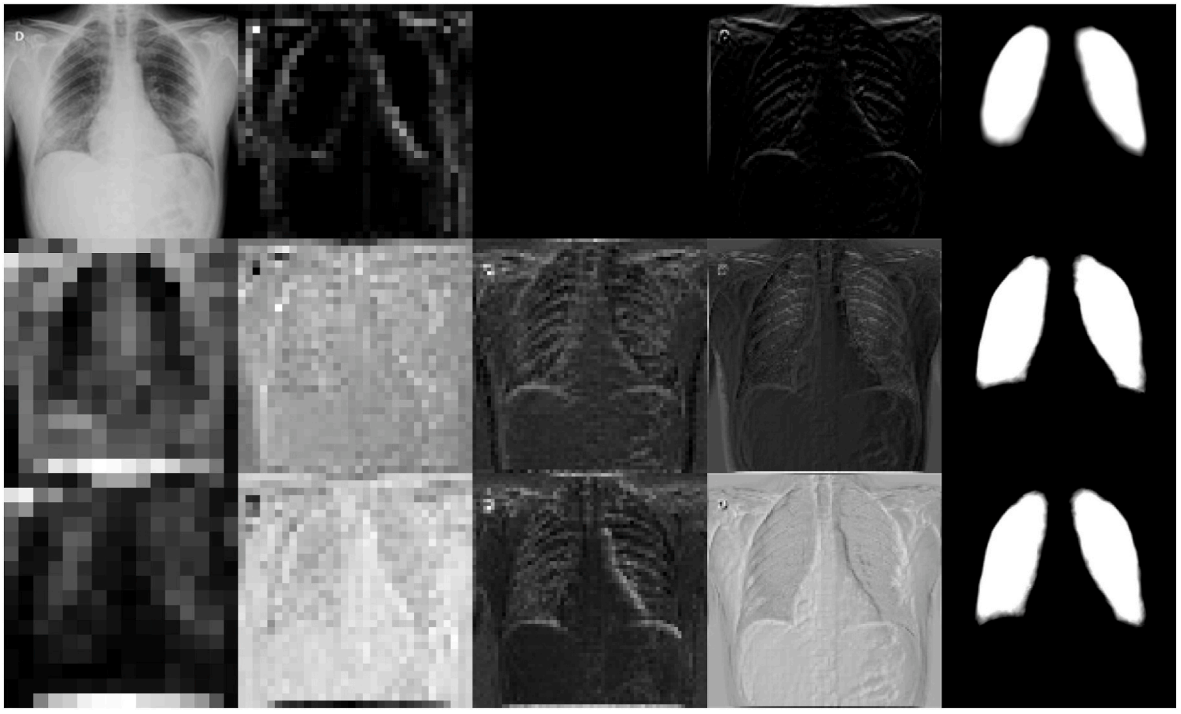

Fig. 4. The visualization of feature maps. From top to bottom, they are TransUNet, FDR-TransUNet32, FDR-TransUNet64. (a) Top is the original image and (a) middle and bottom are 16 × 16 feature maps, (b) are 32 × 32 feature maps, (c) are 64 × 64 feature maps, (d) are 128 × 128 feature maps, and (e) are the output feature maps of the models.

Table 4
Ablation Studies on the different module implementation.

Method	Ablation Studies				
	mIoU	Dice	Precision	Recall	Paras
TransUNet	0.710	0.959	0.963	0.957	105.3 M
FDR32-VIT-CUP	0.726	0.961	0.967	0.957	101.1 M
FDR64-VIT-CUP	0.722	**0.962**	0.974	0.952	97.9 M
TransUNet-CEP	0.716	0.961	0.969	0.955	105.3 M
TransUNet-CEPR	0.720	0.962	0.967	0.960	105.6 M
FDR-TranUNet32	0.694	0.961	**0.974**	0.950	104.3 M
FDR-TransUNet64	**0.731**	**0.962**	0.965	**0.962**	101 M

5. Conclusion

In this study, we proposed a more powerful encoder-decoder architecture. The encoder has flexibly adjustable block depths through different growth rates, and can effectively reduce the loss of image features. By integrating FDR block, ViT, dual-path decoder, deep supervision strategy, and successive residual connections, we achieved significant improvements in pixel localization information extraction and spatial information recovery. Our model is a CNN-transformer-based segmentation model of COVID-19 chest X-ray images and achieves satisfactory segmentation accuracy and generalization ability. The experimental results on two COVID-19 datasets proved that our model can produce a more flexible lung margin curve and more lung tissue. We also explored the application effects of each innovative block and different growth rate hyperparameters.

However, we acknowledge that our model also has some potential

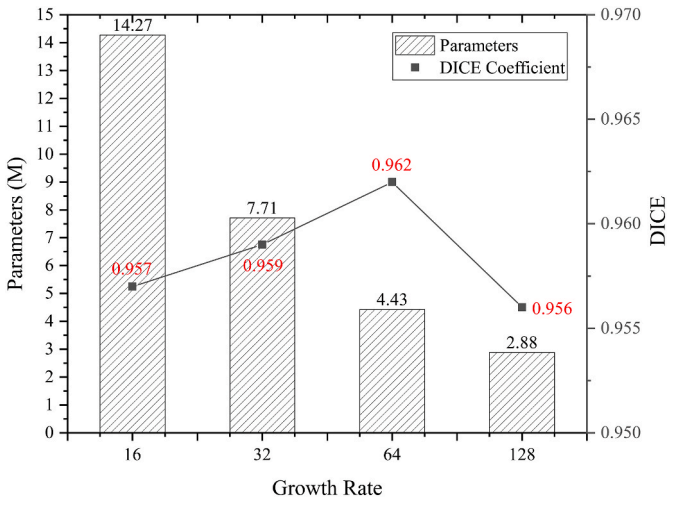

Fig. 5. The number of parameters of FDR blocks with different growth rates, and corresponding experimentational results on COVID-19 lung infection segmentation.

limitations and challenges. It is sensitive to data variation and requires too much computational resources. The choice of hyperparameters is critical to performance. On the other hand, our model needs further adaptation to other medical imaging tasks. Moving forward, we plan to continue refining our model to address these limitations and challenges. We will expand our model to handle multi-modal medical images, such as CT scans and MRI medical images. Additionally, we will further research effective methods to enhance the robustness of our model, and extend to real-time medical image segmentation applications, such as aiding in diagnosis. etc.

Credit author contributions

Zhao Pengcheng: Writing – review & editing, Visualization. Zhang Chaoyang: Writing – original draft, Software, Methodology, Conceptualization. Sun Shibao: Resources, Project administration, Funding acquisition. Hu Wenmao: Validation, Investigation, Data curation

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Longmen laboratory for Exploratory Research Project. [No.MQYTSKT034].

References

[1] J.P.J.R. Kanne, Chest CT Findings in 2019 Novel Coronavirus (2019-nCoV) Infections from Wuhan, China: Key Points for the Radiologist, 2020, pp. 16–17. Radiological Society of North America.
[2] I.D. Apostolopoulos, T.A.J.P. Mpesiana, E.S.I. medicine, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks 43 (2020) 635–640.
[3] F. Shi, et al., Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19 14 (2020) 4–15.
[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[5] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, Proceedings, Part III 18. 2015, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Springer, Munich, Germany, October .
[6] R. Wang, et al., Medical image segmentation using deep learning: A survey 16 (5) (2022) 1243–1267.
[7] S. Akcay, T.J.P.R. Breckon, Towards Automatic Threat Detection: A Survey of Advances of Deep Learning within X-Ray Security Imaging, vol. 122, 2022, 108245.
[8] A. Vaswani, et al., Attention Is All You Need, vol. 30, 2017.
[9] Cao, H., et al. Swin-unet: unet-like pure transformer for medical image segmentation. in Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III. 2023. Springer.
[10] J. Chen, et al., Transunet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021.
[11] A. Hatamizadeh, et al., Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.
[12] A. Dosovitskiy, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020.
[13] M.Z. Alom, et al., Recurrent residual U-Net for medical image segmentation 6 (1) (2019) 14006.
[14] H. Huang, et al., Unet 3+: a full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020.
[15] Z. Zhou, et al., Unet++: redesigning skip connections to exploit multiscale features in image segmentation 39 (6) (2019) 1856–1867.
[16] C.-Y. Lee, et al., Deeply-supervised nets, in: Artificial Intelligence and Statistics, 2015 (Pmlr).
[17] O. Oktay, et al., Attention U-Net: Learning where to Look for the Pancreas, 2018.
[18] A.J.M.T. Das, Applications, Adaptive UNet-based lung segmentation and ensemble learning with CNN-based deep features for automated COVID-19 diagnosis 81 (4) (2022) 5407–5441.
[19] T. Rahman, et al., Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images 132 (2021), 104319.
[20] D.-P. Fan, et al., Inf-net: automatic covid-19 lung infection segmentation from ct images 39 (8) (2020) 2626–2637.
[21] A. Saood, I.J.B.M.I. Hatem, COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet 21 (1) (2021) 1–10.
[22] C. Zhao, et al., Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images 119 (2021), 108071.
[23] D. Müller, I.S. Rey, F. Kramer, in: Automated Chest Ct Image Segmentation of Covid-19 Lung Infection Based on 3d U-Net, 2020 arXiv preprint arXiv: 2007.04774.
[24] Z.-J. Gao, Y. He, Y.J.I.A. Li, in: A Novel Lightweight Swin-Unet Network for Semantic Segmentation of COVID-19 Lesion in CT Images, vol. 11, 2022, pp. 950–962.
[25] P. Zou, J.-S. Wu, SwinE-UNet3+: swin transformer encoder network for medical image segmentation, 1, in: Progress in Artificial Intelligence, vol. 12, 2023, pp. 99–105.
[26] Z. Wang, et al., Densely connected swin-UNet for multiscale information aggregation in medical image segmentation, in: 2023 IEEE International Conference on Image Processing (ICIP), 2023 (IEEE).
[27] M. Liu, et al., Ccat-net: a novel transformer based semi-supervised framework for covid-19 lung lesion segmentation, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022 (IEEE).
[28] Y. Zhang, et al., TranSegNet: hybrid CNN-vision transformers encoder for retina segmentation of optical coherence tomography, Life 13 (4) (2023) 976.
[29] S. Liang, et al., FCF: Feature Complement Fusion Network for Detecting COVID-19 through CT Scan Images, vol. 125, 2022, 109111.
[30] Y. Yang, et al., MMViT-Seg: a lightweight transformer and CNN fusion network for COVID-19 segmentation, Comput. Methods Progr. Biomed. 230 (2023), 107348.
[31] Z. Li, et al., Tfcns: a cnn-transformer hybrid network for medical image segmentation, in: International Conference on Artificial Neural Networks, Springer, 2022.
[32] X. Huang, et al., TDD-UNet: transformer with double decoder UNet for COVID-19 lesions segmentation, Comput. Biol. Med. 151 (2022), 106306.
[33] T. Dhamija, et al., Semantic segmentation in medical images through transfused convolution and transformer networks, Appl. Intell. 53 (1) (2023) 1132–1148.
[34] H. Cao, et al., Swin-unet: unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022.
[35] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[36] G. Huang, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[37] J. Devlin, et al., Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
[38] L. Xu, et al., Multi-class token transformer for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

Chaoyang Zhang. He received his B.S. degree in computer science from Zhengzhou Business University, China in 2019. Currently, he is pursuing a master's degree in computer technology at Henan University of Science and Technology in China. His research interests include medical image processing, artificial intelligence and deep learning.

Shibao Sun. He received his B.S. degree from the School of Information Engineering, Zhengzhou University in 2004. He received his Ph.D. degree from the School of Information Science and Technology, Southwest Jiaotong University in 2008. Currently, he is the Dean of the School of Software at Henan University of Science and Technology. His research interests include artificial intelligence (big data analysis, intelligent applications, industrial Internet, etc.) and graphic image processing areas of research.

Wenmao Hu. He received his B.S. degree from Henan University of Science and Technology, China, in 2018. He is currently working toward the master's degree in computer technology at Henan University of Science and Technology in China. His research interests includes computer vision, image segmentation.

Pengcheng Zhao. (SM'20) received his B.S. degree in Automation from Luoyang Institute of Technology, China, in 2017 and his M.S. degree in Software Engineering from Henan University of Science and Technology, China, in 2020. Currently, he is pursuing his PhD in Control Theory Engineering at Henan University of Science and Technology, China. His research interests include data security and privacy, cyberspace security.