International Journal of Computational Intelligence Systems (2024) 17:21
https://doi.org/10.1007/s44196-023-00401-8

RESEARCH ARTICLE

Segmentation of Lung Lesions through Bilateral Learning Branches to Aggregating Contextual and Local Characteristics

Hao Niu[1] · Linjing Li[2] · Bo Yuan[3] · Min Zhu[1] · Xiuyuan Xu[1] · Xi Lu[4] · Fengming Luo[5] · Zhang Yi[1]

Received: 8 December 2023 / Accepted: 28 December 2023

Abstract

Detecting and analyzing lung lesion regions using artificial intelligence is of great significance in the medical diagnosis of lung CT images, which can substantially improve the efficiency of doctors. However, segmentation of the inflammatory region in the CT image of the lung remains challenging due to the varied sizes, blurry local details, irregular shapes, and limited sizes of datasets. Faced with these challenges, this paper proposes a novel lung lesion segmentation network that incorporates two feature extraction branches to achieve a balance of speed and accuracy. We first design a context branch (CB) to preserve the scale-invariant global context information by the transformer-like module. Besides, a shallow detail branch (DB) based on a deep aggregation pyramid (DAP) module is designed to provide detailed information. Extensive experiments are conducted on two datasets, including the public COVID-19 dataset and a private dataset. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods. Moreover, the trade-off between accuracy and inference speed is achieved.

Keywords Lung lesion segmentation · Contextual and local characteristics · Bilateral learning

1 Introduction

Interstitial pneumonia is a diffuse pulmonary disease primarily caused by environmental factors such as fungal spores, organic dust, and systemic lupus erythematosus [1]. The damage caused by interstitial pneumonia to the human lungs can be divided into three stages. In the early stage, it primarily leads to alveolar wall and pulmonary interstitial inflammation. During the middle stage, diffuse interstitial fibrosis becomes prominent. As the disease progresses to the late stage, patients develop fibrosis of the alveolar wall, resulting in symptoms like poor breathing, acidosis, and hypoxia. In severe cases, respiratory failure can occur, leading to the patient's death [2, 3]. In the early stages of interstitial pneumonia, lung function remains normal or only slightly damaged. Timely anti-fibrosis treatment at this stage can effectively prevent or even reverse the pathological process of interstitial pneumonia. Therefore, timely and effective screening for interstitial pneumonia holds great significance in preserving the patient's lung function to the fullest extent [4].

Different patients have varied types of clinical manifestations, imaging, and pathological features, which leads to the difficulty of diagnosis. To accurately evaluate the symptoms of patients, medical staff from multiple disciplines, including the rheumatology department, respiratory department, and radiology department, need to participate in the diagnosis [5]. Many studies have shown that computed tomography (CT) can reflect the abnormal degree of lung parenchyma and stroma. Observing the tissue structure of the pulmonary lobule is an effective approach for the diagnosis of early pneumonia lesions [6].

✉ Fengming Luo
fengmingluo@outlook.com

✉ Zhang Yi
zhangyi@scu.edu.cn

[1] College of Computer Science, Sichuan University, Chengdu 610065, China

[2] Southwest University of Science and Technology, Mianyang 621010, China

[3] General Practice Ward/International Medical Center Ward, West China Hospital, Sichuan University, Chengdu 610044, China

[4] Department of Radiology, West China Hospital, Sichuan University, Chengdu 610044, China

[5] Department of Pulmonary and Critical Care Medicine, West China Hospital, Sichuan University, Chengdu 610044, China

Published online: 05 February 2024

Recently, multiple convolutional neural networks (CNNs) have been applied to segment the lesions of interstitial pneumonia. U-Net [7] is one of the most representative models among the numerous segmentation networks. U-Net uses an encoder composed of multiple convolution and pooling operators to extract abstract features from the input. A decoder consisting of multiple transposed convolution or upsampling operators attempts to identify the target from the features. Moreover, shortcut connections between the encoder and decoder are designed to facilitate the information flow. Based on U-Net, V-Net [8] extends the dimensionality of the operators from two-dimensional to three-dimensional and adds the residual connections, effectively alleviating the problem of gradient disappearance caused by the deepening of network depth. To reduce the semantic differences between different abstract levels of features, Fan et al. [9] proposed a network named Inf-Net for segmenting CT images of new coronary pneumonia. The network uses a set of implicit reverse attention modules and explicit edge attention modules to establish the relationship between regions and boundaries. Besides, Gu et al. [10] designed a CA-Net based on the attention mechanism and conditional random field to regulate the flow of information for the segmentation of brain glioma.

Although the methods mentioned above delivered an impressive performance for segmenting lesions in the medical images, few of them have considered the discrepancy between the global and local information due to the locality inherent in the convolution operator [11]. The self-attention mechanism in transformer [12] is prevalent for modeling global information, and it has been widely applied in natural language processing tasks. Recently, transformers have achieved or even surpassed the performance of advanced convolution-based models [13] in tasks such as image classification, object detection, and segmentation. However, it is acknowledged that the successful application of transformers relies on large-scale datasets, posing challenges for medical image segmentation tasks. Based on this consideration, we propose a novel bilateral learning mechanism that balances global and local information. A context branch (CB) based on the dilated convolution is designed to preserve the scale-invariant global information, while a detail branch (DB) that leverages the deep aggregation pyramid module (DAP) is proposed to learn the local information. Multiple experiments verified the effectiveness of the proposed method. In summary, the main contributions of this paper are as follows:

(1) The global and local information of lesions in the CT images are balanced through the proposed CB and DB branches.

(2) Multiple experimental results demonstrate that the proposed method achieves higher accuracy than commonly used methods.

(3) Besides the performance improvement, experimental results show that the proposed method can significantly promote inference speed.

2 Related Works

2.1 Medical Image Segmentation base on U-shaped CNNs

In recent years, convolutional networks have achieved tremendous success in medical image segmentation tasks. Long et al. [14] proposed a fully convolutional network (FCN) for end-to-end segmentation of natural images, achieving a breakthrough from conventional handcrafted feature-based methods. Later, the U-Net [7] won first place in the ISBI cell segmentation challenge, showcasing the powerful capacity of CNNs with succinct architecture. After the proposition of U-Net, the U-Net paradigm became the main approach for medical image segmentation tasks [15]. For example, a U-Net combined with the neural memory Ordinary Differential Equation (nmODE) [16] exhibits impressive performance on the organs-at-risk segmentation of radiotherapy [17]. However, both the convolutional and pooling operators in the U-Net are designed to extract local features, implying their constrained capability for global representation. This limitation motivates the use of the transformer model embedded with the self-attention mechanism, which can learn the long-range relationship between features. Variants and applications of the transformer are illustrated as follows:

2.2 Transformers and MLPs

Recent research on transformers such as ViT [13] and MLP-Mixer [18] has exemplified their great potential as alternatives to advanced CNNs. For example, the MLP-Mixer accepts a sequence of linearly projected image patches. The token-mixing MLP blocks in MLP-Mixer allow communication between each image patch from different spatial locations, while channel-mixing MLP blocks enable interactions among different channels. The main component of each MLP block is the fully connected layer, whose size is directly related to the number of image patches. The transformer shows advantages in the modeling of global context information [19]. However, in order to achieve promising performance, a small image patch size and a large patch number are required, which increases the computational cost dramatically. Moreover, it has been acknowledged that the successful application of the transformer necessitates a large amount of annotated data, which is difficult for tasks with a limited dataset, e.g., medical image segmentation.

3 Methodology

The proposed method is a pixel-level segmentation network. Figure 1 shows the architecture of the network that mainly consists of two components, including the context branch (CB) and detail branch (DB). The CB is composed of the lightweight RepVGG network and the MLP-based feature fusion (MLPF) modules. The last three feature maps in the RepVGG (i.e., 1/8, 1/16, 1/32 resolution) are fused through the MLPF to represent global information. The DB is built upon the DAP module for extracting feature maps at the same resolution, specifically designed to capture local information. The feature maps extracted from the CB and DB branches are then fused to the same resolution with respect to the input image. Finally, a 1×1 convolution layer followed by a sigmoid function is used to generate a probability map of the lung lesion region. The details of each component are illustrated in the following sub-sections.

3.1 Improved RepVGG

In order to achieve the balance between speed and accuracy in the segmentation network, dilated convolution is used in the RepVGG [20] to extract feature maps with multi-scales. The improved block of RepVGG is compared with the conventional convolution as shown in Fig. 2. Figure 2A shows the original residual block of ResNet [21], which consists of a residual branch with two 3×3 convolutional layers and a shortcut branch. The two branches are then summed as the output of the block. The residual structure significantly alleviates the problem of gradient vanishing in the deep network and accelerates the convergence of the network during training. Figure 2B represents the proposed block in the RepVGG during training, which consists of a residual branch with one 3×3 convolutional layer, a shortcut branch with one 1×1 convolutional layer, and a shortcut connection branch. The three branches are summed to obtain the output of the proposed block. The proposed structure has multiple paths to propagate gradient information, indicating a better solution for the gradient vanishing problem. Figure 2C represents the proposed block in our RepVGG during inference, which consists of only two 3×3 convolutional layers reconstructed from those three branches in the training phase. By discarding the multi-branches during inference, the proposed method owns better memory utilization and computational efficiency.

It is acknowledged that obtaining large receptive fields is of great importance to the task of segmenting lesions with large size changes and irregular shapes, such as pneumonia lesions. In order to extract features with global properties, a 5×5 dilated convolutional layer with stride 2 is used to replace the standard 3×3 convolutional layer with stride=1. The dilated convolution is based on the vanilla convolution, except for the enlarged receptive field. The vanilla convolution operator can be formulated as follows:

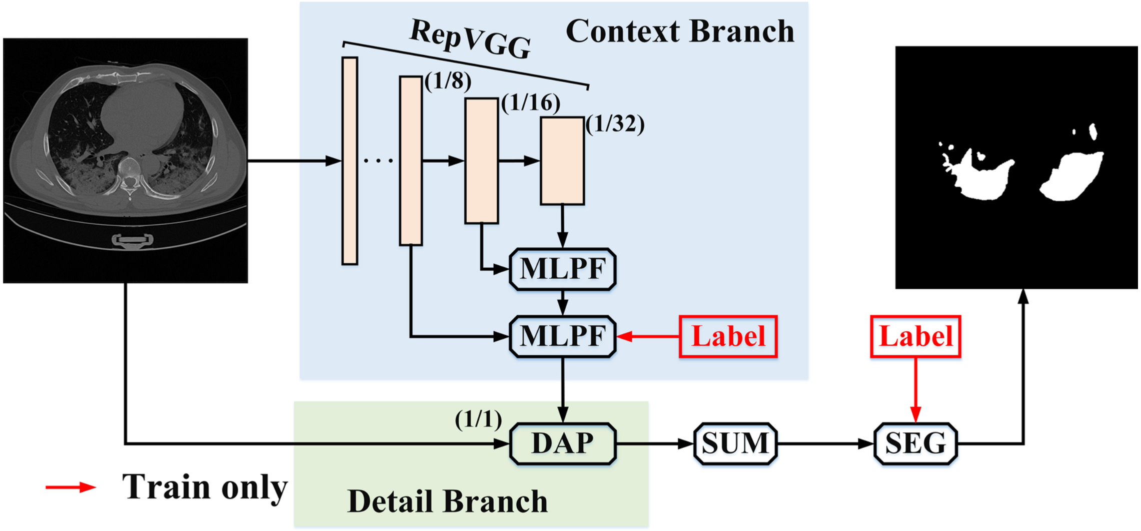

Fig. 1 An overview of the proposed method. Each orange box in the center denotes a multi-channel feature map. The numbers in parentheses represents the spatial ratio of the feature map with respect to the input. **MLPF** stands for the fusion of two feature maps based on MLP. **DAP** stands for the deep aggregation pyramid module. The red characters **Label** represent the supervision we take during the training phase

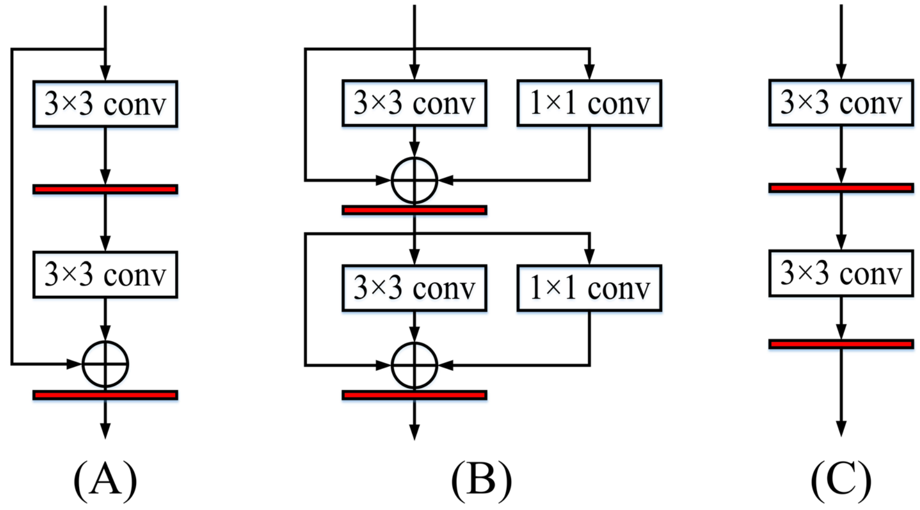

Fig. 2 Comparison of conventional residual block and the proposed method. From left to right are the residual block in the ResNet, training, and inference architecture of the proposed method. The small red box represents the batch normalization layer, and ⊕ represents the addition operator

$$z_{v,u}^{l+1} = \sum_{c=0}^{C^l-1}\sum_{p=0}^{P^l-1}\sum_{q=0}^{Q^l-1} a_{v+p,u+q,c}^{l} \cdot w_{p,q,c}^{l}, \tag{1}$$

where P, Q, and C denote the dimensions of the convolution kernel, and the lower-case letters p, q, and c correspond to the indices in kernel w^l. $z_{v,u}^{l+1}$ represents the input of location (v, u) in the $l+1$ layer, computed through the convolution operation with kernel w^l and the output from the previous layer a^l. Based on the vanilla convolution, the dilated convolution introduces the dilation rate r to control the spacing in the kernel. The dilated convolution can be regarded as inserting $r-1$ zeros between elements in the convolutional kernel. The computational principle of dilated convolution is:

$$z_{v,u}^{l+1} = \sum_{c=0}^{C^l-1}\sum_{p=0}^{P^l-1}\sum_{q=0}^{Q^l-1} a_{v+r\cdot p,u+r\cdot q,c}^{l} \cdot w_{p,q,c}^{l}. \tag{2}$$

By comparing Eq. 1 and Eq. 2, it can be seen that the dilated convolution is degraded to standard convolution when $r = 1$.

3.2 MLP-based Context Branch

The last three feature maps extracted by the proposed method are further processed by a feature fusion module constructed by the MLPF module. The application of MLPF is shown in Fig. 1, where two MLPF are used to fuse the three feature maps. The resolution of the output is 1/8 with respect to that of the input. The computational principle of the MLPF is demonstrated in Fig. 3. During the fusion process, the smaller feature map is up-sampled (UPS) twice based on bilinear interpolation to match the size of the larger one. Then two 3×3 convolutional layers are used sequentially to reduce the number of channels of two feature maps to the smaller one. The two feature maps of the same size are concatenated in the dimension of the channel. Finally, two MLPs are used to fuse the feature maps of the same size on the token dimension and the

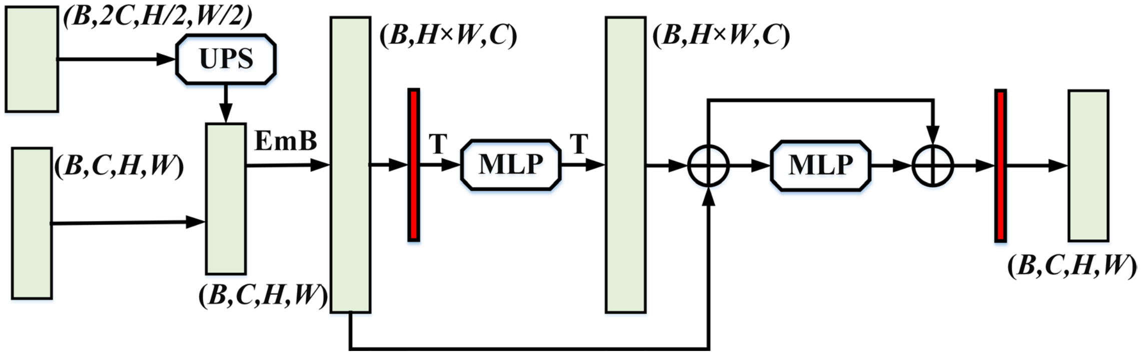

Fig. 3 Computational principle of the proposed **MLPF** module. The small red box stands for batch normalization, and **T** represents the transpose between channel and token. **UPS** stands for the bilinear interpolation up-sampling operation. **EmB** indicates the function that extends the size of (B, C, H, W) to (B, H×W, C) along the spatial dimension

channel dimension, respectively. The MLP layer is based on the transformer block, which owns the advantages in the modeling of long-range relationships among features. The computational principle of the transformer block can be formulated as follows:

$$A' = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3}$$

Suppose the input feature is denoted as A, then the symbol Q, K, and V in Eq. 3 represent the transformed representation of feature A. Here, d_k represents the dimension of feature A, and the division by $\sqrt{d_k}$ is intended to enhance numerical stability. Taking the representation of Q as an example, its computational principle can be summarized as $Q = W^Q A$, where W^Q denotes the learnable parameter that transforms the presentation of A into Q through matrix multiplication. A similar principle can also be found for the representation of K and V. The softmax activation function is designed to introduce nonlinearity into the transformation and accomplish the normalization.

3.3 DAP-based Detail Branch

Local detail information is essential to achieve precise segmentation. The feature extracted by CB is in low resolution and lacks detailed information. Thus, we propose a DB branch based on a deep aggregation pyramid (DAP) module [22] to extract features with high resolution, compensating for the limitation of the CB branch. The architecture of the DB branch is shown in Fig. 4. The structure of the DB branch consists of five paths. First, three average pooling operators are used to reduce the feature maps to different sizes. Second, convolutional layers are used to extract multi-scale local details. Third, these feature maps are up-sampled to the same resolution with respect to the input. Finally, the five paths are further processed by 1x1 convolution and concatenated along with the channel dimension. Note that each convolution unit is composed of the convolutional layer, batch normalization, and ReLU activation function.

3.4 Deep Supervision with Loss

In order to improve the representation capability of the feature from different branches, this paper designs a deep supervision module to guide the learning of the network. The architecture of the deep supervision module is shown in Fig. 5.

In each deep supervision module, the feature map is processed by a 1×1 convolutional layer followed by the sigmoid function. Then, the loss is calculated with respect to the ground truth. During the training phase, two deep supervision modules are used to supervise the prediction from CB and the final probability map. During the inference phase, the first auxiliary supervision module is omitted, and the final result is regarded as the predicted result. Semantic segmentation tasks usually use cross-entropy loss [23] to measure the dissimilarity between the prediction and groundtruth. However, a category imbalance issue arises in lesion segmentation tasks, where the background constitutes the majority of the CT image, thereby slowing down the learning process of the network. Furthermore, there are a large number of challenging voxels in the background, such as textures and nodules, which increase the complexity of the segmentation task. Therefore, the dice loss is employed during the training phase, as it can better address the class imbalance problem compared to cross-entropy. The dice coefficient (DSC) is a metric that assesses the similarity of two sets [24]. The DSC between two sets is given as follows:

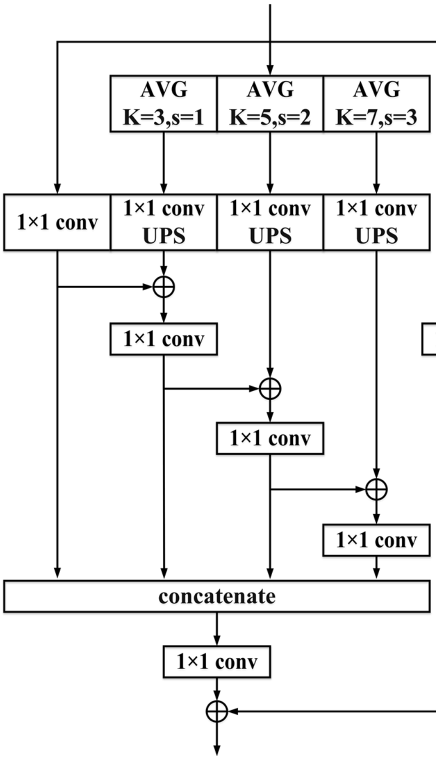

Fig. 4 Architecture of the proposed DAP module. **AVG** stands for the average pooling operators. The letters K and s in the **AVG** represent the kernel size and stride, respectively. **concatenate** denotes the concatenation of multiple feature maps along the channel dimension

$$DSC = 2\frac{|P \bigcap Y|}{|P| + |Y|}, \tag{4}$$

where P and Y denote the volume of the model's segmentation results and ground truth, respectively. The DSC is bounded to [0, 1], where a value of 1 indicates a perfect match between the prediction and ground truth, and

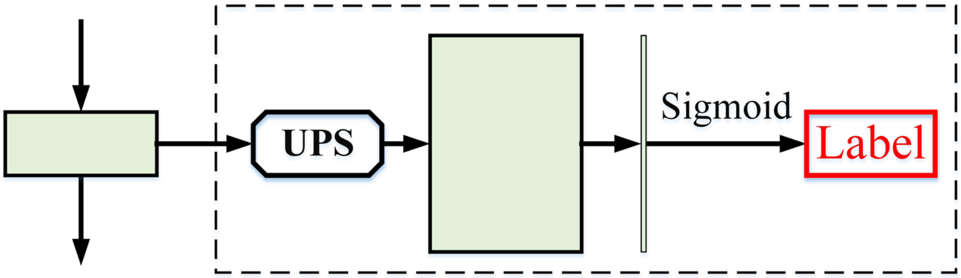

Fig. 5 Architecture of the deep supervision module. **Sigmoid** stands for the sigmoid function. **Label** represents the ground truth

vice versa. Based on the DSC, the dice loss can be described as follows:

$$L_{DSC} = 1 - 2\frac{|P \bigcap Y|}{|P| + |Y|}. \tag{5}$$

According to the proposed deep supervision modules, the loss function is composed of two parts, namely, the main loss used to report the segmentation performance and the auxiliary loss designed to stabilize the training. The two losses are designed as DSC losses that are weighted as follows:

$$Loss = L_f + \alpha L_m, \tag{6}$$

where L_f represents the main loss between the final result and ground truth. L_m represents the auxiliary loss. The two losses are weighted by the hyper-parameter α, which is set to 0.4 in the experiment.

4 Experiments

We conduct various experiments on two datasets to evaluate the proposed method. The first subsection introduces the two datasets and data augmentation methods. The second subsection presents details of the experiment environment. The experimental results and ablation experiments are presented in the last two subsections.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

The proposed method is validated using a self-labeled CT image dataset of interstitial pneumonia provided by West China Hospital of Sichuan University. This dataset is labeled in detail by professional doctors. The dataset comprises 600 gray images with a resolution of 512×512. Among them, 400 images are used as the training dataset, while the remaining 200 images constitute the validation dataset. The COVID-19 dataset introduced in [25] is also used in the experiments. This dataset has 1812 CT images, each with a size of 512× 512 pixels and accompanied by voxel-level binary label images. In our experiments, the dataset is divided into the training and validation parts that contain 1681 and 131 images, respectively.

4.1.2 Data Augmentation

The generalization capability of the model can be improved by using data augmentation [26]. In light of our previous research and analysis, there is a significant difference in noise distribution in the lung lesion images. Based on this consideration, horizontal flip and geometric distortion are used to increase the diversity of the training datasets.

According to the irregular shape of lung lesion regions, the trigonometric function is applied to change the position of each pixel to introduce geometric distortion. This geometric augmentation method can be described as:

$$\begin{cases} \hat{h} = h - 10sin(\frac{2\pi\omega}{152} + 10) \\ \hat{\omega} = \omega - 10sin(\frac{2\pi h}{152} + 10), \end{cases} \tag{7}$$

where $\hat{h}$ and $\hat{\omega}$ denote the coordinates of pixel after distortion.

4.1.3 Evaluation Metrics

All experiments are carried out on the two datasets, and four commonly used metrics for semantic segmentation tasks are used to evaluate the proposed method. Let TP, FN, and FP denote the true positive, false negative, and false positive classes. The recall, precision, F1, and IoU are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

$$IoU = \frac{TP}{TP + FN + FP}. \tag{11}$$

Besides the performance metrics, the number of parameters (Par.) and the floating point operations per second (Gfl.) are used to evaluate the computational cost of the model during the inference process. The time (ms) is also used to measure the milliseconds taken by the model to complete one inference task.

4.2 Implementation Details

The proposed network and control methods are implemented by using the PyTorch library. All CNNs-based models are trained and validated on a computer with NVIDIA GTX TITAN XP with 12GB of RAM, and i7-7770 CPU @ 4.20GHz. We utilize the Adam optimizer with a learning rate of 0.0001. The batch size is set to 4 for optimizing all segmentation networks in the training stage. The model with the minimum average loss value on validation datasets is chosen as the final model during the training process of 100 epochs. During the inference stage, voxels with an output probability value greater than 0.5 are identified as lung lesions, while the rest are considered as background.

4.3 Comparison with Other Models

The proposed network is compared to five modern segmentation networks on two datasets using the same loss function. The compared networks include SegNet [27], Deeplab v3 [28], U-Net [7], U-Net++ [29], and TransUNet [30]. The decoder in SegNet uses pooling indices computed in the max-pooling step of the encoder to perform nonlinear up-sampling that eliminates the need for learning to up-sample and improves the memory versus accuracy trade-off. Deeplab v3 uses multiple dilated convolutional operators to detect characteristics on multiple scales. The architecture of U-Net consists of a decoder, an encoder, and multiple shortcut connections between them. U-Net++ eliminates the semantic gap between different levels of feature maps. TransUNet uses the transformer structure to improve the feature fusion module based on U-Net.

The above results presented in Table 1 demonstrate that the proposed network achieves the best performance compared with other segmentation networks in terms of IoU and F1 on two datasets. These improvements are also evident in Fig. 6, which compares the prediction results against the groundtruth. Specifically, the networks based on U-Net excel in identifying tiny cracks and are more accurate than other networks. The prediction results of the proposed network for lung lesion regions exhibit better continuity and integrity than those of U-Net, which can be attributed to the contextual branch (CB) introducing a larger receptive field. The detailed branch (DB) of the proposed method provides rich detailed information that contributes to clearer details and smoother boundaries.

The network's segmentation speed is directly related to the algorithm's application in engineering. We compare the speed and accuracy with the input size of 512×512. The detailed results are shown in Table 1. Our network achieved the best accuracy and real-time speed, with a processing time of 17.9 ms for each image. The amount of parameters and computation cost of the proposed method is also the lowest among all methods. The above analysis indicates that our network can be applied to the practical segmentation task.

4.4 Ablation Study

In this subsection, the effectiveness of the two modules proposed in this paper is analyzed in our private dataset. The feature fusion module based on MLPF, the context branch, and deep supervision are separately added on the basis of RepVGG to verify their effectiveness. There are five ablation experiments, namely RepVGG-32, RepVGG-8, context branch, without deep supervision, and ours. Table 2 reports the contributions of each module. Experiments are carried out on the private interstitial pneumonia dataset provided by West China Hospital of Sichuan University. The implementation details are kept the same with Sect. 4.2. From the results shown in Table 2, it can be observed that all of the modules lead to the improvement of performance, demonstrating the effectiveness of the modules in the proposed method.

Table 1 Comparison of different networks on the private dataset and COVID-19 dataset

Network	Par.(M)	Gfl.(GMac)	Time.(ms)	Our Dataset		COVID-19	
				IoU(%)	*F1*(%)	*IoU*(%)	*F1*(%)
SegNet	29.4	16	24.5	56.9	72.6	55.0	70.9
Deeplab v3	15.3	15.9	26.9	69.7	82.2	71.9	83.7
UNet	17.3	16	25.4	79.1	88.4	79.2	88.4
UNet++	26.9	15.1	21.9	80.9	89.5	80.4	89.1
TransUnet	66.8	129.4	32.9	81.2	89.6	80.9	89.5
ours	**7.3**	**12.7**	**17.9**	**81.9**	**90.1**	**81.5**	**89.9**

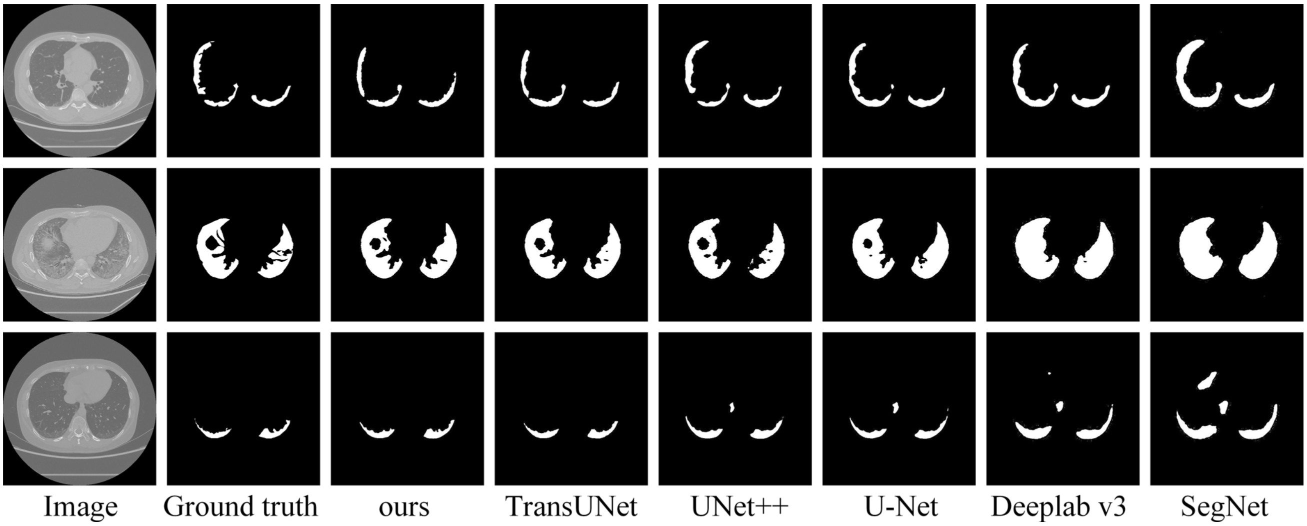

Fig. 6 Comparison of segmentation results from different networks

Table 2 Comparison of segmentation networks with different modules

Network	*IoU*(%)	*F1*(%)
RepVGG-32	65.9	79.5
RepVGG-8	77.9	87.4
context branch	78.3	88.0
without deep supervision	81.4	89.8
ours	**81.9**	**90.1**

5 Conclusion

Aiming at the characteristics of the global context, varied size change, small local details, and irregular shapes of the lung lesion regions, a novel deep network composed of a context feature extraction branch and a detailed feature extraction branch is proposed in this paper. In the proposed method, an improved lightweight network based on RepVGG is used to extract global context feature maps with multi-scale receptive fields. Feature fuse modules are used to efficiently fusing feature maps. At the same time, a detailed feature extraction branch composed of a deep aggregation pyramid module is used to provide local detailed information. The proposed network is verified on two datasets. Experimental results show that the proposed network achieves the best accuracy. In future research, we will validate the proposed method on additional medical image segmentation tasks. Additionally, we plan to incorporate the currently prevalent prompt-based mechanism to develop a universal medical image segmentation model.

Author Contributions Conceptualization, Zhang Yi and Fengming Luo; methodology, Zhang Yi and Min Zhu; software, Xiuyuan Xu and Xi Lu; validation, Linjing Li and Bo Yuan; writing-original draft preparation, Hao Niu; writing-review and editing, Linjing Li and Bo Yuan; All authors have read and agreed to the published version of the manuscript.

Funding This work is supported by the National Major Science and Technology Projects of China under Grant 2018AAA0100201.

Availability of data and material The dataset used in this paper is available upon request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Jee, A.S., Sheehy, R., Hopkins, P., Corte, T.J., Grainge, C., Troy, L.K., Symons, K., Spencer, L.M., Reynolds, P.N., Chapman, S., et al.: Diagnosis and management of connective tissue

disease-associated interstitial lung disease in australia and new zealand: a position statement from the thoracic society of australia and new zealand. Respirology **26**(1), 23–51 (2021)

2. Choi, Y., Liu, T.T., Pankratz, D.G., Colby, T.V., Barth, N.M., Lynch, D.A., Walsh, P.S., Raghu, G., Kennedy, G.C., Huang, J.: Identification of usual interstitial pneumonia pattern using rna-seq and machine learning: challenges and solutions. BMC Genom. **19**, 147–159 (2018)
3. Plantier, L., Cazes, A., Dinh-Xuan, A.-T., Bancal, C., Marchand-Adam, S., Crestani, B.: Physiology of the lung in idiopathic pulmonary fibrosis. Eur. Respir. Rev. **27**, 147 (2018)
4. Hashisako, M., Fukuoka, J.: Pathology of idiopathic interstitial pneumonias. Clin. Med. Insights **9**, 23320 (2015)
5. Oliveira, R., Ribeiro, R., Melo, L., Grima, B., Oliveira, S., Alves, J.: Connective tissue disease-associated interstitial lung disease. Pulmonology **28**(2), 113–118 (2022)
6. Alhamad, E.H., Cosgrove, G.P.: Interstitial lung disease: the initial approach. Med. Clin. **95**(6), 1071–1093 (2011)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241. Springer, Cham (2015)
8. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, pp. 565–571 (2016)
9. Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: automatic covid-19 lung infection segmentation from ct images. IEEE Trans. Med. Imaging **39**(8), 2626–2637 (2020)
10. Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging **40**(2), 699–711 (2020)
11. Sang, H., Zhou, Q., Zhao, Y.: Pcanet: pyramid convolutional attention network for semantic segmentation. Image Vis. Comput. **103**, 103997 (2020)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 55 (2017)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
15. Wang, L., Zhang, C., Zhang, Y., Li, J.: A method of lung organ segmentation in ct images based on multiple residual structures and an enhanced spatial attention mechanism. Mathematics **11**(21), 4483 (2023)
16. Zhang, Y.: nmODE: neural memory ordinary differential equation. Artificial Intelligence Review, 1–36 (2023)
17. Hu, J., Yu, C., Zhang, Y., Zhang, H.: Enhancing robustness of medical image segmentation model with neural memory ordinary differential equation. Int. J. Neural Syst. **5**, 2350060–2350060 (2023)
18. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: an all-mlp architecture for vision. Adv. Neural. Inf. Process. Syst. **34**, 24261–24272 (2021)
19. Wang, J., Zhang, H., Yi, Z.: Cctrans: Improving medical image segmentation with contoured convolutional transformer network. Mathematics **11**(9), 2082 (2023)
20. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Hong, Y., Pan, H., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085 (2021)
23. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
24. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855 (2019)
25. Zhou, T., Canu, S., Ruan, S.: An automatic covid-19 ct segmentation network using spatial and channel attention mechanism. arXiv preprint arXiv:2004.06673 (2020)
26. Wu, Q., Dai, P., Chen, P., Huang, Y.: Deep adversarial data augmentation with attribute guided for person re-identification. SIViP **15**, 655–662 (2021)
27. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
28. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
29. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)
30. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.