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.

Computers in Biology and Medicine 135 (2021) 104526

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution

Xiangyu Zhao [a], Peng Zhang [a], Fan Song [a], Guangda Fan [a], Yangyang Sun [a], Yujia Wang [a], Zheyuan Tian [a], Luqi Zhang [a], Guanglei Zhang [a,b,*]

[a] School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China
[b] Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China

ARTICLE INFO

Keywords:
Dual attention strategy
COVID-19
Deep learning
Hybrid dilated convolution
Segmentation

ABSTRACT

Coronavirus Disease 2019 (COVID-19) has become one of the most urgent public health events worldwide due to its high infectivity and mortality. Computed tomography (CT) is a significant screening tool for COVID-19 infection, and automatic segmentation of lung infection in COVID-19 CT images can assist diagnosis and health care of patients. However, accurate and automatic segmentation of COVID-19 lung infections is faced with a few challenges, including blurred edges of infection and relatively low sensitivity. To address the issues above, a novel *dilated dual attention U-Net* based on the dual attention strategy and hybrid dilated convolutions, namely *D2A U-Net*, is proposed for COVID-19 lesion segmentation in CT slices. In our *D2A U-Net*, the dual attention strategy composed of two attention modules is utilized to refine feature maps and reduce the semantic gap between different levels of feature maps. Moreover, the hybrid dilated convolutions are introduced to the model decoder to achieve larger receptive fields, which refines the decoding process. The proposed method is evaluated on an open-source dataset and achieves a Dice score of 0.7298 and recall score of 0.7071, which outperforms the popular cutting-edge methods in the semantic segmentation. The proposed network is expected to be a potential AI-based approach used for the diagnosis and prognosis of COVID-19 patients.

1. Introduction

COVID-19 pandemic caused by SARS-nCov-2 continues to spread all over the world [1], and most of the countries have been affected in this unprecedented public health event. By March 2021, more than 116 million of cases of COVID-19 have been reported and more than 2,580,000 people died [2] of COVID-19 infection. Due to the strong infectivity of SARS-nCov-2, identification of people infected by COVID-19 is significant to cut off the transmission and slow down virus spread. Reverse transcriptase-polymerase chain reaction (RT-PCR) is considered as the gold standard of diagnosis [3] for its high specificity, but it is time-consuming and laborious. Also, the capacity of RT-PCR tests can be rather insufficient in the less-developed regions, especially during the pandemic. Computed tomography (CT) imaging is one of the most commonly used screening methods to detect lung infection and has proved to be efficient in the diagnosis and follow-up prognosis of COVID-19.

Compared with chest X-ray images, CT imaging is more sensitive, especially in the early stage of infection. Ground glass pattern is the most common finding in COVID-19 infections, usually in the early stage, while pulmonary consolidation can be observed in the later stage. Pleural effusion can also be observed in pathological CT slices. These typical features of COVID-19 lung infection are shown in Fig. 1.

Thus, chest CT imaging is regarded as a convenient, fast and accurate approach to diagnose COVID-19. The evaluation of the localization and geometric features of the infection area could provide adequate information on disease progression and help physicians make better treatments [5–7]. However, manual annotation of the infection regions is a time-consuming and laborious work. Also, the annotation made by radiologists may be subjective and biased due to personal judgements.

Recently, numerous deep learning algorithms using convolutional neural networks (CNNs) have been proposed to detect COVID-19 infection. For instance, Wang and Wong [8] have developed a COVID-Net to perform ternary classification among healthy people, COVID-19 patients and people infected with other pneumonia in chest X-ray images, which achieves an overall accuracy of 93.3%. In terms of

---

* Corresponding author. School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China.
E-mail address: guangleizhang@buaa.edu.cn (G. Zhang).

https://doi.org/10.1016/j.compbiomed.2021.104526
Received 7 March 2021; Received in revised form 24 May 2021; Accepted 24 May 2021
Available online 2 June 2021
0010-4825/© 2021 Elsevier Ltd. All rights reserved.

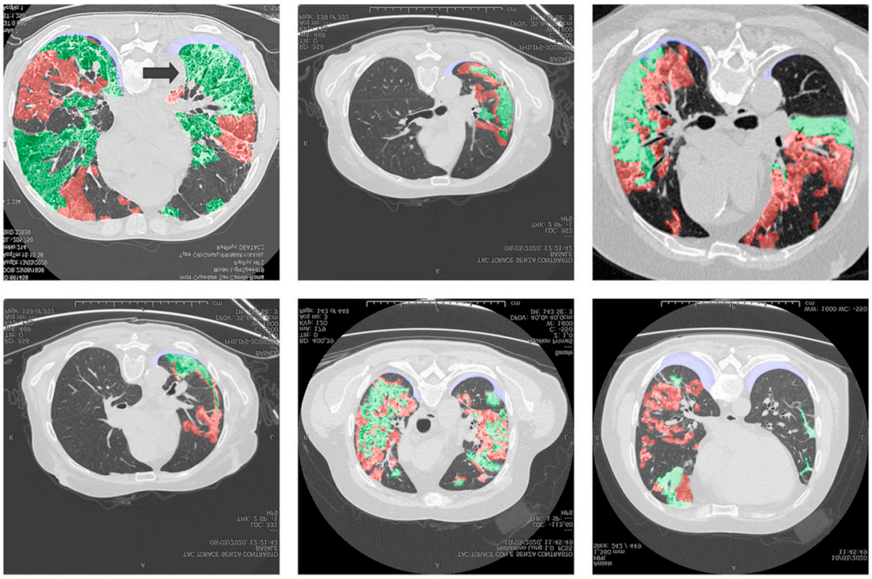

Fig. 1. Example of COVID-19 CT slices, where the red, green and blue masks denote the ground glass, consolidation and pleural effusion respectively. The images are collected from Ref. [4].

deep learning algorithms for CT imaging, Zhou and Canu [9] have proposed an automatic network facilitated with attention mechanism to segment the infection area from CT slices. Fan et al. [10] developed an Inf-Net and corresponding semi-supervision algorithm to perform CT segmentation. Zheng et al. [11] proposed a weakly-supervised deep learning method to detect the COVID-19 infection in CT volumes. Xi et al. [12] presented a dual-sampling attention network to diagnose COVID-19 from community acquired pneumonia. However, the detection of the lung infections caused by COVID-19 in CT images remains challenging, because infection regions vary in shape, position and texture, and the boundaries between lesions and normal tissues can be rather blurred. These features increase the difficulty of COVID-19 detection and limit the model performance, especially in terms of sensitivity.

To address the issues above, we proposed a *dilated dual attention U-Net* (*D2A U-Net*) framework to automatically segment the lung infection in COVID-19 CT slices. Since the infected tissues can be hardly distinguishable from the normal tissues, we introduce a dual attention strategy consisting of a *gate attention module* (GAM) and a *decoder attention module* (DAM) to refine feature maps and produce more informative feature representation. The proposed GAM is utilized by fusing features and semantic-rich gate signals to refine the skip connections in the network. The proposed DAM is introduced to the model decoder to improve the decoding quality, especially when segmenting the blurred lesions. As COVID-19 infection varies in position and size, we utilize hybrid dilated convolutions with different dilation rate in the model decoder to obtain larger receptive fields and balance the segmentation performance on both large and tiny objects, which thus provides better segmentation results. The sensitivity for infection segmentation has been improved significantly due to these refinements, which leads to better segmentation performance.

The paper is organized as follows: Section 2 offers a review of related works on CT segmentation. Section 3 describes the overview of this work and details our proposed model. Section 4 presents the details of our experiments and provides both quantitative and qualitative segmentation results. Section 5 discusses the proposed method and concludes our work.

2. Related works

In this section, we will go through 4 types of most related works, which includes chest CT segmentation, attention mechanism, dilated convolution and AI-based COVID-19 segmentation systems.

2.1. Chest CT segmentation

Chest CT imaging is one of the most popular screening methods for lung disease diagnosis [13]. Segmentation of organs and lesions provides crucial information for the diagnosis and prognosis of many diseases. However, since manual segmentation remains time-consuming, laborious and subjective, automatic CT segmentation has gained much popularity in the research fields. Recent researches upon automatic CT segmentation mainly focus on utilizing machine learning techniques. Related works most feature a pixel-wise classifier to infer from extracted features and make dense predictions. For example, Mansoor et al. [14] proposed a texture-based feature classifier for pathological lung segmentation in the CT images. Yao et al. [15] utilized texture analysis and support vector machine to segment infections in the lung tissues. These algorithms have realized automatic segmentation in the chest CT images but several issues remain unsolved, including subjective bias in feature extraction and difficulties in segmenting nodule regions. Deep learning algorithms feature powerful fitting capacity and require no laborious preprocessing. Most cutting-edge segmentation algorithms are based on deep learning approaches. For example, Shaziya et al. [16] used U-Net to segment lung tissues in the chest CT scans. Zhao et al. [17] proposed a fully convolutional neural network with multi-instance and conditional adversary loss for pathological lung segmentation.

2.2. Attention mechanism

Attention plays an important role in human perception and visual cognition [18]. One significant property in human perception is that humans hardly process visual information as a whole. Instead, humans usually process visual information recurrently, where top information is utilized to guide bottom-up feedforward process [19]. Inspired by this principle, attention mechanism has been widely used in computer vision, especially in the image classification [20–22]. Related algorithms typically refine feature maps in the spatial dimension, channel dimension or both. For example, Hu et al. [20] introduced a Squeeze-and-Excitation module, where global average pooling is performed on the input features to produce channel-wise attention. Woo et al. [21] proposed a convolutional block attention module (CBAM) to

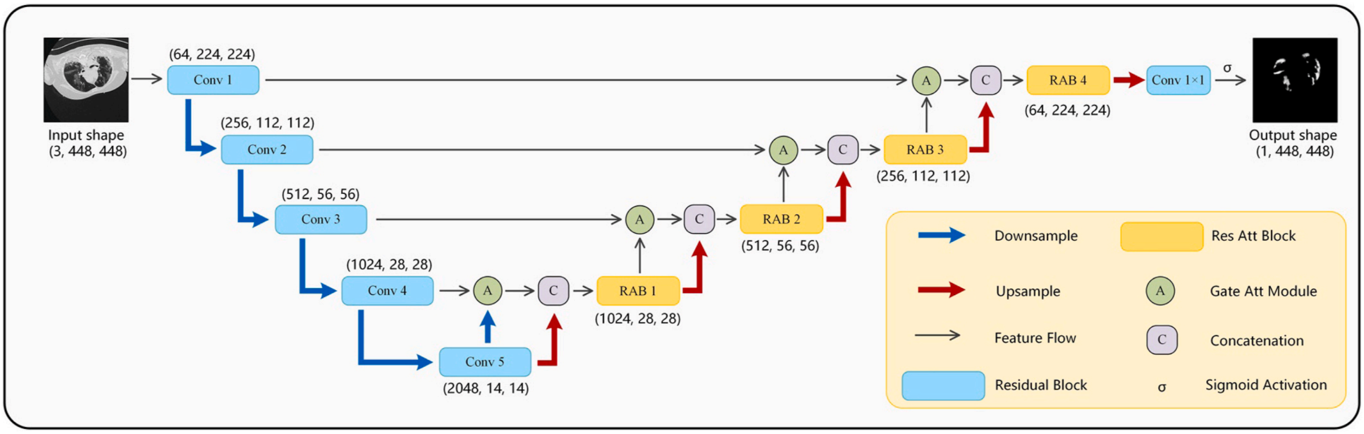

Fig. 2. The proposed *D2A U-Net* architecture with a ResNeXt-50 (32 × 4d) backbone, which takes a CT slice as input and outputs infection region predictions.

introduce a fused attention consisting of channel attention and spatial attention. Wang et al. [22] presented a residual attention network, which contains an attention module featuring an encoder-decoder architecture. Attention mechanism has also been utilized in semantic segmentation tasks to make more accurate dense predictions. For instance, Li et al. [23] proposed a Pyramid Attention Network to exploit the impact of global contextual information in semantic segmentation.

These typical algorithms resemble in some aspects. Certain operations, such as global pooling, convolution, and the combination of downsampling and upsampling, are utilized to enhance the informative regions in the feature maps and suppress irrelevant information, which allows the network to learn more generalized visual structures and improve the robustness against noisy inputs.

2.3. *Dilated convolution*

Traditional deep convolutional networks usually involve strided convolution or pooling operations to improve the receptive fields, in which the input images are downsampled. However, these operations often lead to the loss of global information in dense predictions, such as semantic segmentation and object detection. Yu and Koltun [24] introduced dilated convolution to deep networks, which has proved to be useful in dense predictions. The basic idea of dilated convolution is to insert “holes” (zeros) in the convolution kernels to obtain larger receptive fields without downsampling. Dilated convolution avoids information loss during downsampling and has been widely used in the semantic segmentation tasks [25–27]. However, it has been observed that simply stacking dilated convolution in CNNs may cause grid effects [24], which could lead to severe performance deterioration. Wang et al. [28] proposed a hybrid dilated convolution (HDC) framework to avoid grid effects, which improves the segmentation performance on both large and tiny objects.

2.4. *AI-based COVID-19 segmentation systems*

Artificial intelligence (AI) has been widely utilized in fighting against COVID-19. We mainly focus on AI-based semantic segmentation systems upon CT scans. Many works focus on learning robust and noise-insensitive representations from limited or noisy inputs. For example, Xie et al. [29] proposed a RTSU-Net for segmenting pulmonary lobes in the CT scans. A non-local neural network module was introduced to learn both visual and geometric relationships among the feature maps to produce self-attention. Wang et al. [30] presented a noise-robust framework for COVID-19 lesion segmentation. They utilized a noise-robust Dice loss and an adaptive self-ensembling strategy to learn from noisy labels. Chen et al. [31] proposed a residual attention U-Net which introduced aggregated residual transformations and soft attention mechanism to learn robust feature representations. Also, researchers have investigated segmentation schemes that achieve both high speed and accuracy. For example, Zhou et al. [32] developed a rapid, accurate and machine-agnostic segmentation and quantification method for automatic segmentation of COVID-19 lesions. The innovation of their work lies in the first CT scan simulator for COVID-19 and a novel network architecture which solves the large-scene-small-object problem. Qiu et al. [33] developed a parameter-efficient framework to achieve fast segmentation of COVID-19 lung infection with relatively low computational cost.

3. Methods

In this section, we will go through the details of the proposed *D2A U-Net* architecture. In the first part, we will offer an overview of the proposed network. We then provide details about the proposed attention modules. Finally we introduce our proposed model decoders with hybrid dilated convolutions.

3.1. *Overview of network architecture*

Basically, our proposed network is based on the U-Net [34] architecture, which is quite popular in medical image segmentation. Compared with the original U-Net, dilated convolutions and a novel combination of attention mechanism are integrated in our framework to obtain better feature representation. We integrate the dual attention strategy in the model decoder. A gated attention module is inserted inside the skip connections to utilize feature representations from different levels and reduce the semantic gap between the encoder and the decoder. Also, we introduce another fused attention mechanism in the model decoder to refine feature maps after upsamling. Specifically, a hybrid dilated convolution module [28] is utilized as the basic block of the model decoder to enlarge receptive fields and produce better dense predictions. For the model encoder, both VGG-style encoder proposed in the original U-Net [34] and ResNeXt-50 (32 × 4d) [35] pretrained on ImageNet are utilized. The network scheme is shown in Fig. 2.

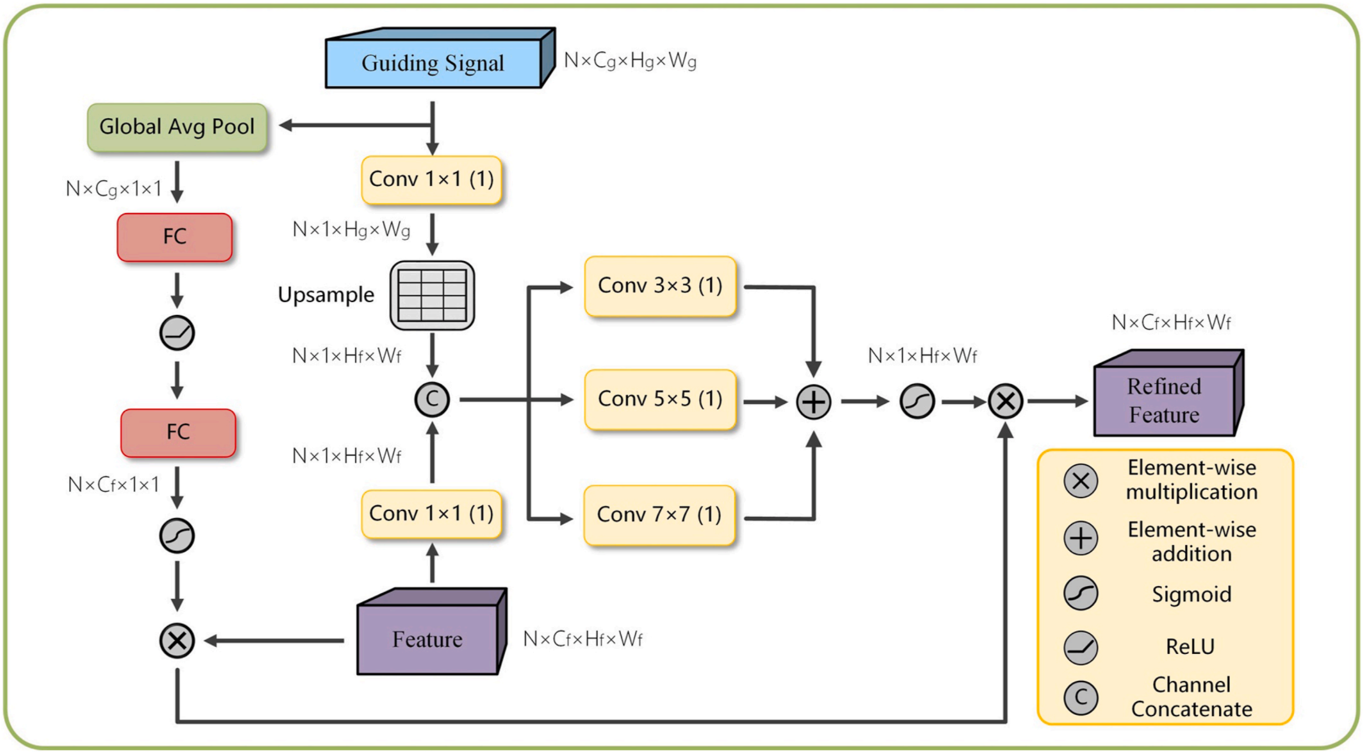

Fig. 3. The proposed *gate attention module*, which takes guiding signal and features as input to generate fused attention. The number shown in the parentheses inside conv block means the number of outchannels. The shape of tensors are also shown in the figure, where *N* denotes the batch size, *C* denotes the number of channels, *H* denotes the height and *W* denotes the width.

3.2. Dual attention

We introduce a *dual attention* strategy composed of a *gate attention module* (GAM) and a *decoder attention module* (DAM) to our network. The motivation behind utilizing *dual attention* strategy instead of single attention module is to further highlight the infection area and suppress false positives. GAM is utilized to refine the features extracted by the model encoder and to reduce the semantic gap by fusing high and low level feature maps, which highlights potential infection regions and improves the sensitivity to COVID-19 infection. DAM is inserted in the model decoder to refine the feature representations after upsampling, which is used to suppress the noise that may be introduced during upsampling and inhibit false positives.

3.2.1. Gate attention module

Feature concatenation from the encoder to the decoder is the typical topological structure in U-Net, where the combination of high-resolution features in the encoder and upsampled features in the decoder enables better localization of segmentation targets [34]. However, not all visual representations in the encoder feature maps contribute to precise segmentation. In addition, the semantic gap between the encoder and the decoder can limit the performance of the model. Therefore, we introduce a *gate attention module* prior to concatenation to refine the features from the encoder and reduce the semantic gap.

Oktay et al. [36] proposed an attention gate to refine the encoder features with attention mechanism. But in their proposed attention gate, only spatial attention mechanism is implemented to refine features. However, the introduction of both channel attention and spatial attention will improve the efficiency of attention mechanism. Thus, inspired by the global attention upsample module proposed in pyramid attention network [23] and CBAM [21], we propose a novel design of a *gate attention module* to enable both channel attention and spatial attention. Detailed scheme of the proposed GAM is shown in Fig. 3. Two feature maps are fed into the attention module. The guiding signal refers to the feature map from the model decoder (or the last convolution block in the model encoder), and the feature refers to the feature map fed to the skip connections. $\mathbf{G} \in \mathbb{R}^{C_g \times H_g \times W_g}$ denotes the guiding signal and $\mathbf{F} \in \mathbb{R}^{C_f \times H_f \times W_f}$ denotes the feature.

In the U-shaped mesh structure, **G** contains more deep semantic information which is encoded in the channel dimension compared with **F**. We utilize a global average pooling operation followed by a multilayer perception (MLP) to create the channel attention map $Z_c(\mathbf{F}) \in \mathbb{R}^{C_f \times 1 \times 1}$. The output size of the MLP is smaller than the input size, which enables the suppression of irrelevant feature representations in the channel dimension. In short, we compute the channel attention as follows:

$$Z_c(\mathbf{F}) = \sigma\left(MLP\left(P_{avg}(\mathbf{G})\right)\right) = \sigma\left(W_{C_f}\left(ReLU\left(W_{C_g/r}\left(P_{avg}(\mathbf{G})\right)\right)\right)\right) \quad (1)$$

where σ denotes sigmoid activation, P_{avg} denotes global average pooling, $W_{C_g} \in \mathbb{R}^{C_g/r \times C_g}$ and $W_{C_f} \in \mathbb{R}^{C_f \times C_g/r}$, r denotes reduce ratio and in our experiments it is set to 16.

Spatial attention is guided by both the guiding signal and the input feature itself. We use convolution operation with one filter to squeeze the channel dimension of **G** and **F**. Then the reduced feature map from **G** is upsampled to match the size of **F**. A combination of convolution operation with different kernel size is utilized to produce spatial attention $Z_s(\mathbf{F}) \in \mathbb{R}^{1 \times H_f \times W_f}$. In short, we compute spatial attention as:

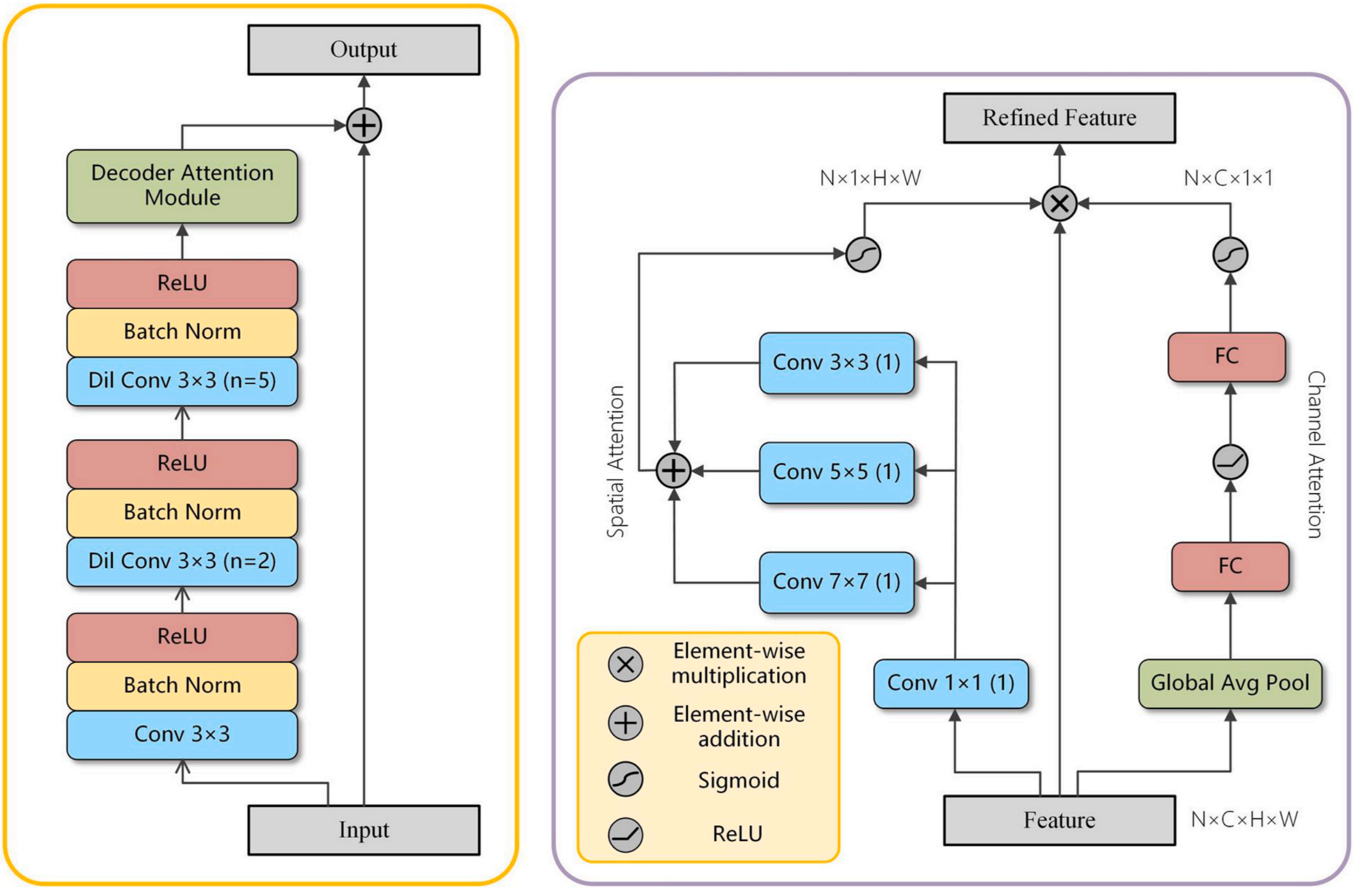

Fig. 4. The proposed *residual attention block* (left) and *decoder attention module* (right). RAB integrates a hybrid dilated convolution module and a DAM; *n* in the parentheses refers to dilation rate. DAM is utilized to refine post-upsample features; the number shown in the parentheses inside conv block means the number of outchannels. The shape of tensors are also shown in the figure, where *N* denotes the batch size, *C* denotes the number of channels, *H* denotes the height and *W* denotes the width.

$$\begin{aligned} &Z_s(\mathbf{F}) = \sigma(f_{3\times3}([\mathbf{F_r},\mathbf{G_r}]) + f_{5\times5}([\mathbf{F_r},\mathbf{G_r}]) + f_{7\times7}([\mathbf{F_r},\mathbf{G_r}])) \\ &\quad where \quad \mathbf{F_r} = f^r_{1\times1}(\mathbf{F}),\ \mathbf{G_r} = upsample(f^r_{1\times1}(\mathbf{G})) \end{aligned} \tag{2}$$

where σ denotes sigmoid activation, $f_{3\times3}$, $f_{5\times5}$ and $f_{7\times7}$ denote convolution operation with corresponding kernel size. $f^r_{1\times1}$ is used to squeeze channel dimension.

Then we use element-wise multiplication to combine spatial and channel attention to produce the fused attention $Z(\mathbf{F})$:

$$Z(\mathbf{F}) = \mathbf{F} \circ Z_s(\mathbf{F}) \circ Z_c(\mathbf{F}) \tag{3}$$

where $\circ$ denotes element-wise multiplication.

3.2.2. Decoder attention module

In semantic segmentation, high-resolution visual representations in the encoder need to be upsampled to make dense predictions. Transposed convolution and interpolation are both popular solutions to image upsampling, but both have their drawbacks. Compared with interpolation, transposed convolution is trainable and offers more nonlinearity to deep networks, which improves the model capacity. However, grid effects are hard to avoid if hyperparameters are not properly configured, and this drawback can be more troublesome when stacking more than one transposed convolution layer. Thus we propose a combination of bilinear interpolation and following convolution to upsample the feature maps. However, as interpolation is not trainable, it is inevitable to introduce irrelevant information or noise to the upsampling process. Thus, we introduce a *decoder attention module* to solve this issue. A fused attention mechanism is utilized to refine the post-upsampling feature maps in both channel and spatial dimensions. The scheme is shown in Fig. 4. Compared with the proposed GAM, DAM is more simplified and only takes one input, but the implementation of both channel and spatial attention is quite similar. We use $Z_c(\mathbf{F}) \in \mathbb{R}^{C\times1\times1}$ to denote channel attention, $Z_s(\mathbf{F}) \in \mathbb{R}^{1\times H\times W}$ to denote spatial attention and $Z(\mathbf{F})$ to denote fused attention. In short, DAM is computed as follows:

$$Z_c(\mathbf{F}) \quad = \sigma(MLP(P_{avg}(\mathbf{F}))) = \sigma\big(W_1\big(ReLU\big(W_0\big(P_{avg}(\mathbf{F})\big)\big)\big)\big) \tag{4}$$

where σ denotes sigmoid activation, P_{avg} denotes global average pooling, $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$, r denotes the reduce ratio and it is set to 16 in our experiments.

$$Z_s(\mathbf{F}) = \sigma\big(f_{3\times3}(f^r_{1\times1}(\mathbf{F})) + f_{5\times5}(f^r_{1\times1}(\mathbf{F})) + f_{7\times7}\big(f^r_{1\times1}(F)\big)\big) \tag{5}$$

where σ denotes sigmoid activation, $f_{3\times3}$, $f_{5\times5}$ and $f_{7\times7}$ denote

Table 1
Dataset description.

Num	Dataset	Description	Split
1	COVID-19 CT Segmentation Dataset [4]	110 slices with 100 containing annotations.	Test Set
2	Segmentation Dataset nr. 2 [4]	9 CT volumes (373 out of the total of 829 slices have been evaluated by a radiologist as positive and segmented.)	Training Set
3	COVID-19 CT Lung and Infection Segmentation Dataset [37]	20 CT volume (Left lung, right lung, and infections are labeled by two radiologists and verified by an experienced radiologist, and 1844 out of the total of 3520 slices contains infection regions.)	Training Set

convolution operation with corresponding kernel size. And $f^r_{1\times 1}$ is used to squeeze channel dimension.

$$Z(\mathbf{F}) = \mathbf{F}\circ Z_s(\mathbf{F})\circ Z_c(\mathbf{F}) \tag{6}$$

where $\circ$ denotes element-wise multiplication.

3.3. Residual attention block

Standard convolution hardly reaches a large receptive field with a fixed kernel size. Such drawback in traditional U-Net based networks may limit the segmentation performance. Inspired by the design of hybrid dilated convolution [28], we proposed a *residual attention block* (RAB) as the basic module in the model decoder. We explore to use dilated convolutions in the decoder to capture multiscale patterns of the upsampled feature maps. The stem of RAB is a stack of dilated convolutions with a kernel size of 3 and dilation rate of [1, 2, 5]. Such dilation rate settings acquires larger receptive fields and also avoids grid effects of vanilla dilated convolutions [28]. Then the RAB is followed by a *decoder attention module*. The scheme is shown in Fig. 4.

We assume initial receptive field as 1 × 1. The equivalent kernel size of dilated convolution is computed as follows:

$$K = k + (k-1)(n-1) \tag{7}$$

where K denotes the equivalent kernel size, k denotes the actual kernel size, and n denotes the dilation rate.

Thus, the equivalent kernel sizes of dilated convolutions with kernel size 3 and dilation rate [1, 2, 5] are 3, 5, 11, respectively. According to the definition of receptive field, such design of stacked dilated convolution obtains a receptive field of 17 × 17, which enables the capture of global information. Also, dilated convolution with different dilation rate can capture multiscale information in the feature maps, which can contribute to the accurate segmentation on both large and small objects.

In addition, we utilize residual connections in the RAB to avoid gradient vanishing. Hybrid dilated convolutions are followed by a DAM to refine upsampled features and produce fused attention maps. In short, the output of our RAB is computed as follows:

$$\mathbf{Y} = \mathbf{X} + DAM(HDC(\mathbf{X})) \tag{8}$$

where X denotes the input feature maps, Y denotes the output feature maps, DAM denotes the proposed *decoder attention module*, and HDC denotes the hybrid dilated convolutions.

4. Experiments

4.1. CT segmentation dataset

CT axial slices used in our experiments consist of 3 independent datasets [4,37]. The details about the datasets used in our experiments are shown in Table 1. Dataset 1 contains 100 axial CT slices from more than 40 patients, which have been rescaled to 512 × 512 pixels and grayscaled. All slices are segmented by a radiologist using three labels: ground-glass opacity, consolidation and pleural effusion. Dataset 2 contains 9 axial CT volumes, where 373 out of the total 829 slices have been evaluated by a radiologist as positive and segmented using 2 labels including ground-glass opacity and consolidation. Dataset 3 contains 20 CT axial volumes, which have been segmented by two radiologists and verified by an experienced radiologist.

Dataset 2 and Dataset 3 contain 29 CT volumes in total, but not all slices contain infection regions. We choose to discard all slices containing no COVID-19 infection and use slices with annotations only. As annotations in Dataset 3 do not distinguish ground-glass opacity and consolidation, we take both ground-glass opacity and consolidation in Dataset 2 as COVID-19 lesions and do not distinguish them as well, thus creating a binary segmentation dataset. An intensity normalization has been applied on both datasets and all slices have been rescaled to 512 × 512 pixels to match Dataset 1. We take all ground-glass, consolidation and pleural effusion in Dataset 1 as COVID-19 lesions, just the same as what we have done to Dataset 2.

We do not choose to combine processed Dataset 1 to 3 together and then split them randomly, because in this way slices of one subject may exist in both training and test datasets, which could be regarded as data leakage and cause a virtual-high model performance. Since Dataset 1 contains the largest number of subjects (40 subjects), which hence best suits to be the independent test set, we finally obtain 1645 processed slices from processed both Dataset 2 and Dataset 3 and use these slices as our final training dataset, and then we use the 100 axial slices from Dataset 1 as our final test dataset. Such data split can best evaluate model generalization capacity.

4.2. Implementation details

4.2.1. Model hyperparameters and settings

Model encoder is a ResNeXt-50 (32 × 4d) pretrained on ImageNet-1K. We remove the global average pooling and full connection layers from original network. The number of output channels is 64, 256, 512, 1024, 2048, respectively, which are the same as the original paper of ResNeXt. Convolution operations in model decoder are padded and without stride, if not specified. Bilinear interpolation is utilized to upsample feature maps, and scale factor is set to 2. Dice loss is widely utilized in semantic segmentation, but the differential of Dice loss is sometimes numerically unstable and may lead to oscillation in training process. The combination of Dice loss and cross-entropy could avoid this issue. Thus we combine Dice loss $\mathcal{L}_d$ and binary cross-entropy loss $\mathcal{L}_c$ as our final loss function:

$$\mathcal{L}_{seg} = \mathcal{L}_d + \alpha\mathcal{L}_c \tag{9}$$

where $\alpha = 1$ in our experiments.

Table 2
Quantitative analysis of U-Net based models on our dataset, including U-Net, Attention U-Net, U-Net + + and the proposed *D2A U-Net*. Metrics include Dice score, pixel error and recall score.

Model	Param.	FLOPs	Dice	Pix Err	Recall
U-Net	7.85 M	43.13G	0.6384	0.0332	0.5512
Att. U-Net	8.12 M	43.78G	0.6646	0.0390	0.6470
U-Net++	9.16 M	106.81G	0.6830	0.0332	0.6417
D2A U-Net	8.95 M	53.19G	**0.7047**	**0.0323**	**0.6626**

4.2.2. Training details

Our model is implemented using PyTorch on an Ubuntu 16.04 server. We use a NVIDIA RTX 2080 Ti GPU to accelerate our training process. Data augmentation is utilized in our training process to reduce overfitting and improve the generalization capacity. First all input images are rescaled to 560 × 560, followed by random flip, random rotation, random gamma and log transform. Finally images are randomly cropped to 448 × 448 and fed into the network. The model is optimized by an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e - 8$. The L_2 regularization is utilized to reduce overfitting as well. We set model weight decay to $1e - 8$. Monte Carlo cross-validation is utilized to find the optimal hyper-parameters (i.e., the initial learning rate and number of epochs) during the training phase. Initial learning rate is set to 1e-4 and is reduced when faced with plateau, with reduce factor being 0.1 and patience being 10. The batch size is set to 6 and we perform evaluation on test set after 30 epochs. The training process takes approximately 140 min.

4.3. Evaluation metrics

We use Dice similarity coefficient and pixel error as the main metrics to evaluate the segmentation performance of our *D2A U-Net*. Dice is a statistic used to gauge the similarity of two samples, and has been widely used to evaluate the performance in semantic segmentation. Pixel error measures the number of pixels predicted falsely in the image, which shows the global segmentation accuracy of the proposed models. Compared to the Dice score or recall score, pixel error is easier to interpret and more intuitive. Both metrics measure segmentation performance in a global way. In addition, we calculate recall score of infection regions, as recall score measures model's sensitivity to lung infection, which is rather significant in terms of COVID-19 infection. We use G to denote ground truth, P to denote dense predications, TP to denote true positive, FP to denote false positive, TN to denote true negative and FN to denote false negative. These metrics are calculated as follows:

$$Dice \quad = \frac{2|G \cap P|}{|G| + |P|} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

$$Pixel\ Error = \frac{FP + FN}{TP + TN + FP + FN} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

4.4. Comparison with cutting-edge methods

In this section, the proposed *D2A U-Net* is compared with other cutting-edge methods to evaluate the effectiveness of the proposed model. Two groups of model comparison have been conducted in the experiments to provide a fair comparative observation of the model performance from different angles of view.

First, the proposed *D2A U-Net* has been compared with popular U-Net family models including U-Net [34], Attention U-Net [36] and U-Net + + [38]. Models listed above are all trained from scratch and share the same backbone structure, i.e. the VGG-style backbone, which refers to the encoder design proposed in the original U-Net paper [34]. Such experimental settings provide the most fair comparison of those U-Net based models, as they share the same model backbone and training strategies.

In addition, utilizing backbone pretrained on ImageNet to accelerate convergence and improve segmentation results has been popular in the CV tasks of natural images. Thus, we also introduce a pretrained *D2A U-Net* with ResNeXt-50 (32 × 4d) backbone to further improve the segmentation performance. The pretrained version is compared with 2 cutting-edge models widely used for natural image segmentation, including FCN8s [39] and DeepLab v3 (output stride = 8) [40], both of which contain a pretrained ResNet-101 backbone.

Apart from model performance comparison, model parameters and computational costs (FLOPs) are also compared in our experiments.

To better evaluate the performance, all the metrics listed in Table 2 and Table 3 are averaged in 5 reduplicate experiments to report a fair and reliable result.

4.5. Segmentation results

4.5.1. Quantitative analysis

Detailed comparison among different models in our experiments is shown in Table 2 and Table 3. As shown in Table 2, our proposed network outperforms U-Net, Attention U-Net and U-Net + + in terms of Dice, pixel error and recall. As these models are identical in the encoder, it is clear that the proposed dual attention strategy and RAB contribute significantly to the infection segmentation. The utilization of attention mechanism aids the model to detect infected tissues more accurately, which reduces the number of false positives and improves recall score. Also, RAB in the decoder captures both large and tiny visual structures, which is helpful to segment infection lesions with different size. In addition, it should be noted that the proposed *D2A U-Net* with VGG-style backbone outperforms U-Net + + with comparably lower model parameters and computational costs, which could prove the balance of efficiency and performance in our models.

Utilizing pretrained backbone could also improve model performance. As can be seen, our *D2A U-Net* with pretrained ResNeXt-50 (32 × 4d) backbone outperforms other networks in terms of Dice, pixel error and recall by a large margin and yields the best results on our dataset. Also, our *D2A U-Net* with pretrained ResNeXt-50 (32 × 4d) backbone takes fewer computational resources than FCN8s and DeepLab v3 (output stride = 8). As can be seen from Table 3, pretrained encoder could offer a better initialization of the parameters and reduce overfitting, especially when the data amount is insufficient. Overall, the proposed architecture performs better than the existing cutting-edge models.

Table 3
Quantitative analysis of CV models on our dataset, including FCN-8s, DeepLab v3 (os = 8) and the proposed *D2A U-Net*. Backbone *ResNet-101* and *ResNeXt-50 (32 × 4d)* are pretrained on ImageNet-1K. Metrics include Dice score, pixel error and recall score.

Model	Backbone	Param.	FLOPs	Dice	Pix Err	Recall
DeepLab v3	ResNet-101	58.63 M	185.00G	0.7095	0.0323	0.6780
FCN8s	ResNet-101	51.94 M	165.67G	0.6825	0.0315	0.6348
D2A U-Net	ResNeXt-50	90.05 M	149.97G	**0.7298**	**0.0311**	**0.7071**

4.5.2. *Qualitative analysis*

We visualized segmentation results, as shown in Fig. 5. It can be seen from the visualization that our proposed model outperforms other models noticeably. U-Net and Attention U-Net are the least sensitive to COVID-19 lesions, and the background pixels have much stronger activation compared with other models. U-Net + + obtains more accurate segmentation results, but it is still not promising because some tiny lesions or lesions with blurred edge are poorly segmented. *D2A U-Net* with VGG-style backbone produces most accurate segmentation masks compared with other U-Net based models mentioned above, and when backbone is switched to ResNeXt-50 (32 × 4d), *D2A U-Net* achieves the best segmentation results, which is comparably more sensitive to blurred or tiny lesions than other models.

Table 4
Comparison with the latest researches in the field of COVID-19 CT segmentation.

Literature	Method	Dice
Wang et al. [41]	3D U-Net	0.704
Yan et al. [42]	COVID-SegNet	0.7026
Ma et al. [43]	3D U-Net	0.673
Fan et al. [10]	Semi-Inf-Net	0.739
Ours	*D2A U-Net*	0.7298

4.5.3. *Comparison with latest researches*

Apart from common models in the field of computer vision, we also conducted the comparison with latest researches, as shown in Table 4. Our proposed *D2A U-Net* yields top performance compared with the

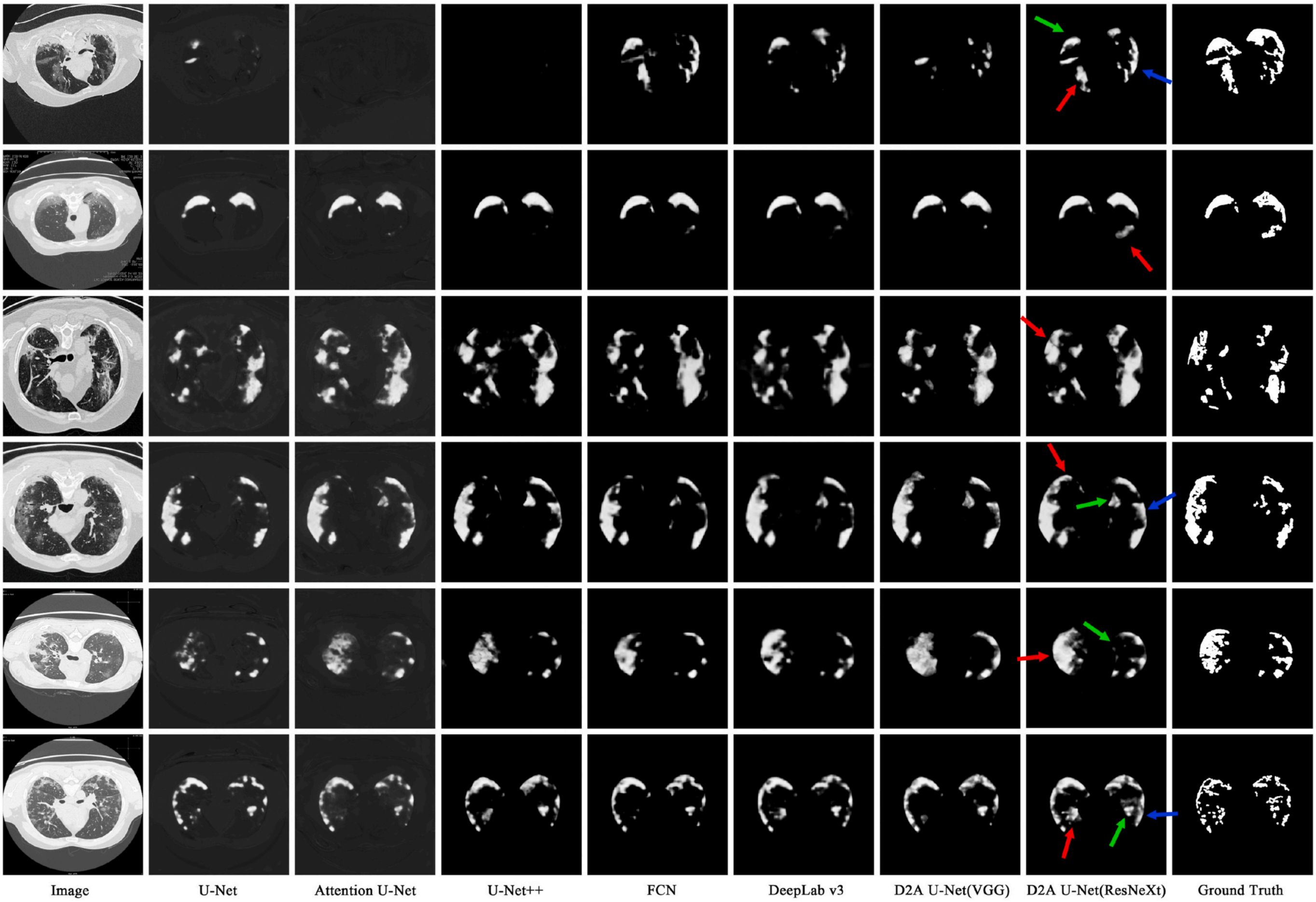

Fig. 5. Visual comparison of COVID-19 lesions segmentation results.

Table 5
Ablation analysis of proposed *D2A U-Net*, where GAM denotes *gate attention module*, RAB denotes *residual attention block* and PB denotes pretrained backbone.

Method	Dice	Pixel Error	Recall
(No.1) U-Net	0.6384	0.0332	0.5512
(No.2) U-Net + GAM	0.6771	0.0343	0.6445
(No.3) U-Net + RAB	0.6579	0.0354	0.6154
(No.4) U-Net + RAB + GAM	0.7047	0.0323	0.6626
(No.5) U-Net + RAB + GAM + PB	**0.7298**	**0.0311**	**0.7071**

latest advances in the field of COVID-19 CT segmentation. The performance of our proposed *D2A U-Net* attributes its success to the development of our proposed *dual attention* strategy and the utilization of hybrid dilated convolution blocks.

4.5.4. *Ablation Study*

Several ablation experiments are conducted to evaluate the performance of components presented in our model, as shown in Table 5 and Fig. 6. In addition, we have visualized feature maps to further demonstrate the effectiveness of the proposed network components.

4.5.4.1. Effectiveness of proposed GAM. To evaluate the validity of the proposed GAM in our experiments, we design two baselines shown in Table 5, including No.1 (U-Net only) and No.2 (U-Net + GAM). Feature maps have been shown in Fig. 7 to provide an intuitive demonstration of the effectiveness of the proposed GAM. Experimental results have shown that introducing GAM to the U-Net model can highlight the potential infection region and thus boost the performance, which leads to a better Dice score and recall.

4.5.4.2. Effectiveness of proposed RAB. We conducted similar experiments (No.1 and No.3) to explore the effectiveness of the proposed RAB, which includes a hybrid dilated convolution block and a decoder attention module. From Fig. 8, it is indicated that the introduction of hybrid dilated convolution block into the decoder improves the recall score of segmentation, and the following decoder attention module further highlights the infection regions and also suppresses false positives. By introducing RAB to our model, the proposed network yields better results than the vanilla version.

4.5.4.3. Effectiveness of combining GAM, RAB and PB. As can be seen from Table 5, in No.4, introducing GAM and RAB together (proposed *D2A U-Net*) yields the best results in our experiments, and the performance boost exceeds the simple addition of each module's performance boost. Such experimental results indicate that introducing GAM and RAB together promotes the performance mutually. Also, in No.5, the pretrained backbone offers better parameter initialization, and therefore could improves the performance further.

5. Conclusion

In this paper, we proposed a novel segmentation network, *D2A U-Net*, for COVID-19 CT segmentation. In order to refine the feature maps and improve segmentation performance, especially in terms of recall score, we present a *dual attention* strategy consisting of a *gate attention module* and a *decoder attention module*. *Gate attention module* is proposed to produce a fused attention map on the features extracted by the encoder. *Decoder attention module* is introduced to the model decoder, which helps refine the upsampled feature maps after convolution operations. Also, *hybrid dilated convolution*, combined with *decoder attention module*, referred to as *residual attention block*, has been introduced as the basic block of the model decoder. *Hybrid dilated convolution* is utilized in the decoder to increase receptive field and improve the quality of feature representation. Experimental results indicate that the proposed network is capable of segmenting COVID-19 lesions from CT slices automatically, and achieves the best results among the popular cutting-edge models evaluated in our experiments. But our work is still limited to some degree, as only binary segmentation is performed in our experiments, which can limit model's potential use in both diagnosis and health care. Multi-class segmentation is expected in the future to further evaluate the performance of the proposed model. Also, despite the significantly better performance of our *D2A U-Net* with ResNeXt-50 (32

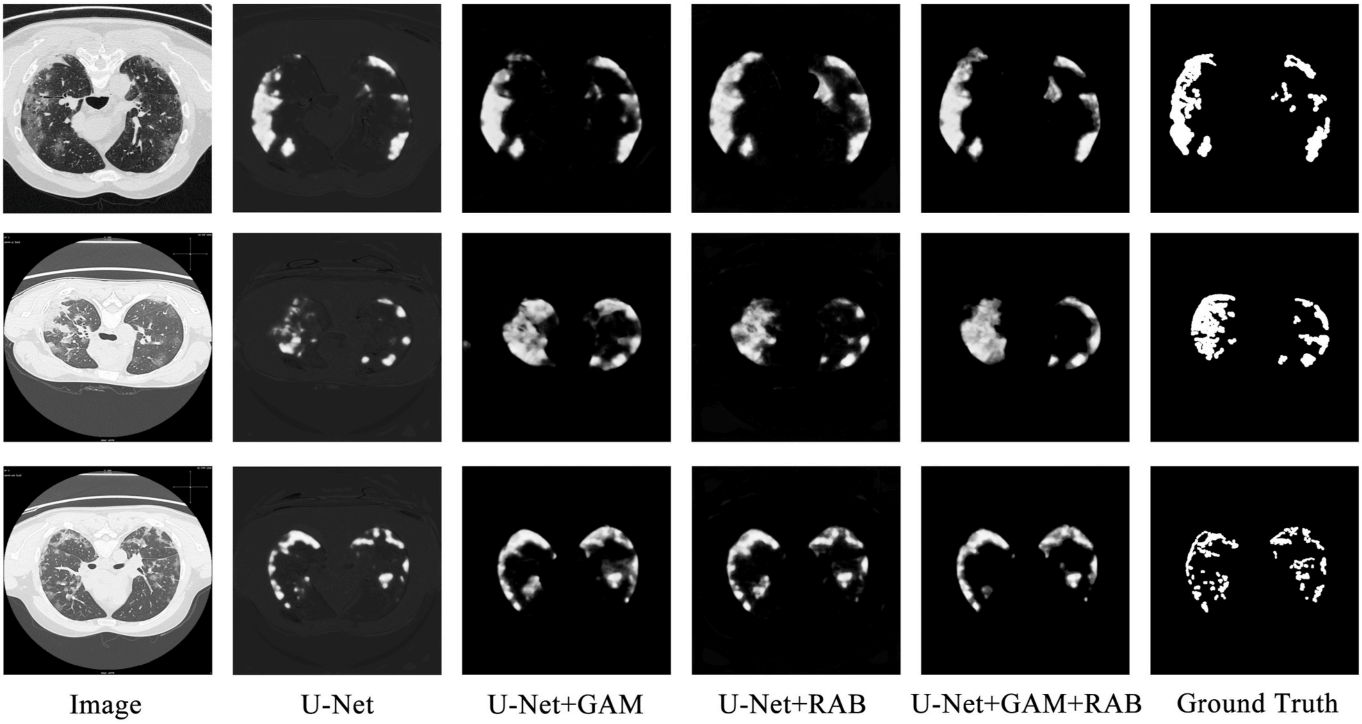

Fig. 6. Visualization of ablation study.

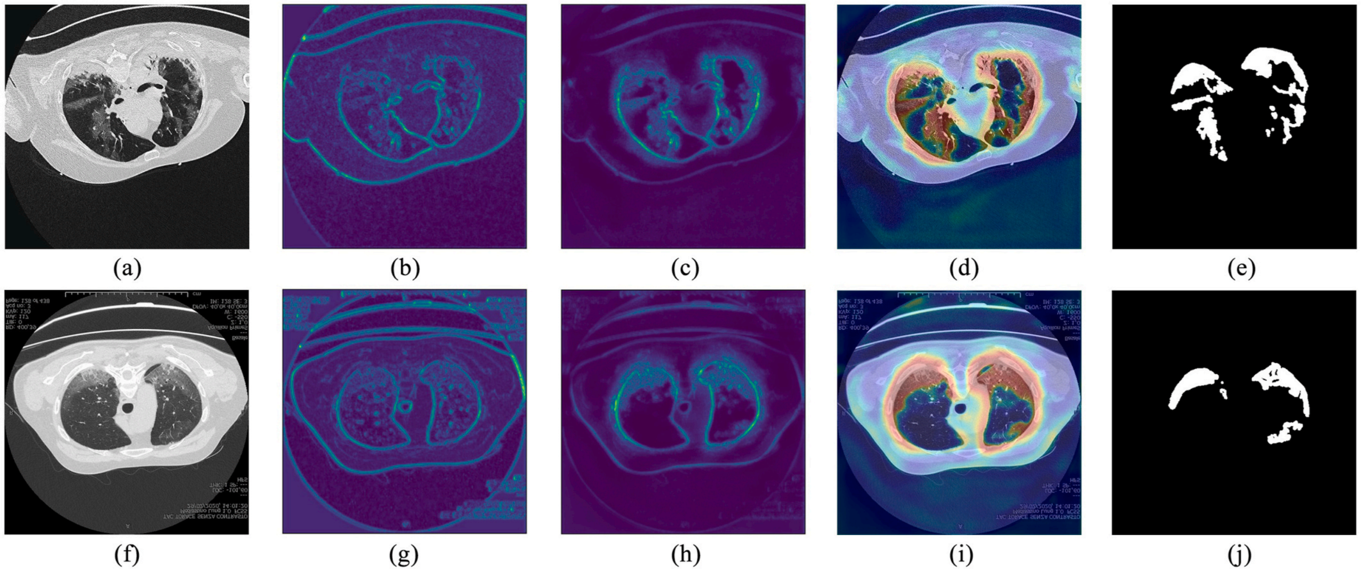

Fig. 7. From left to right(a-e, f-j): Views of a CT scan, feature maps before and after GAM, attention coefficients and the ground truth.

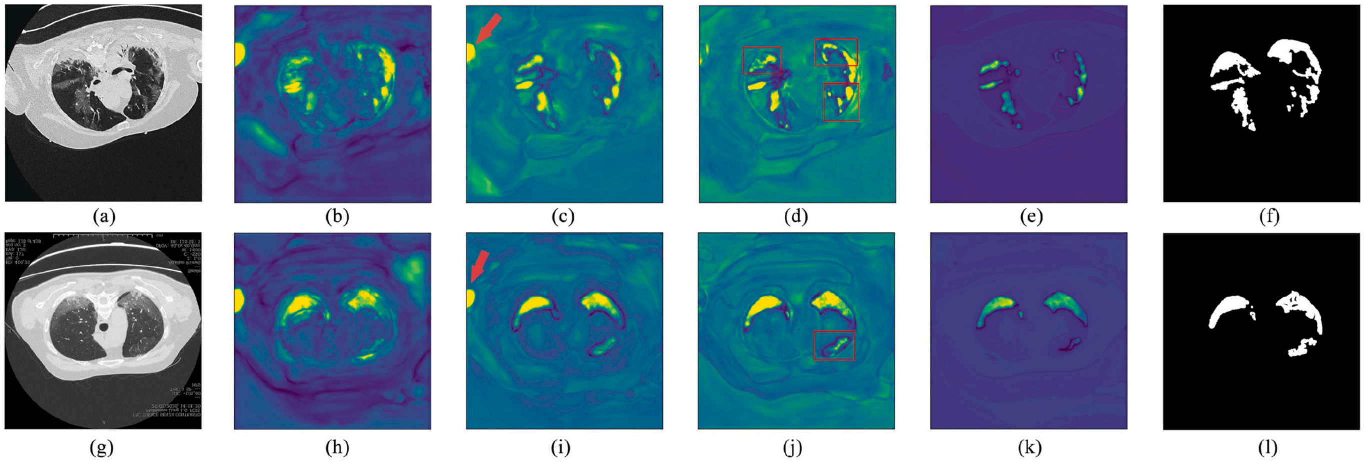

Fig. 8. From left to right(a-f, g-l): Views of a CT scan, feature maps before RAB, feature maps after hybird dilated convolutions, feature maps segmented by the RAB, feature maps segmented by U-Net and the ground truth. Segmentation missed when using U-Net architectures have been highlighted with red bounding boxes in (d) and (j). False positives suppressed by decoder attention module has been highlighted with red arrows in (c) and (i).

× 4d) backbone, the model has much more model parameters than other architectures with similar backbones (FCN8s and DeepLab v3). It is believed that as ResNet family models have a large number of channels (eg. 1024 and 2048 in the last two layers), the parameters of the decoder becomes extremely large. Such problem might be addressed by introducing so-called Bottleneck in ResNets to the decoder of *D2A U-Net* to reduce the number of channels and thus model parameters.

Declaration of competing interest

We confirm that neither the manuscript nor any parts of its content are currently underconsideration or published in another journal. All authors listed have contributed to thismanuscript and agreed to submit to your journal. The authors declare that there is no conflictof interest regarding the publication of this paper.

Acknowledgements

This work was partially supported by the Fundamental Research Funds for Central Universities, the National Natural Science Foundation of China (No. 61871022, 61601019), the Beijing Natural Science Foundation (No. 7202102), and the 111 Project (No. B13003).

References

[1] Chen Wang, Peter W. Horby, Frederick G. Hayden, George F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (10223) (2020) 470–473.
[2] Covid-19 Global Cases, johns hopkins university, 2020. Accessed August 28, https://coronavirus.jhu.edu/map.html/.
[3] Wenling Wang, Yanli Xu, Ruqin Gao, Roujian Lu, Kai Han, Guizhen Wu, Wenjie Tan, Detection of sars-cov-2 in different types of clinical specimens, Jama 323 (18) (2020) 1843–1844.
[4] Covid-19 Ct Segmentation Dataset, 2020. Accessed August 28, https://medicalsegmentation.com/covid19/.
[5] Junqiang Lei, Junfeng Li, Xun Li, Xiaolong Qi, Ct imaging of the 2019 novel coronavirus (2019-ncov) pneumonia, Radiology 295 (1) (2020) 18, 18.
[6] Ming-Yen Ng, Elaine YP. Lee, Yang Jin, Fangfang Yang, Li Xia, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, et al., Imaging profile of the covid-19 infection: radiologic findings and literature review, Radiology: Cardiothoracic Imaging 2 (1) (2020), e200034.
[7] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L. Hesketh, Lian Yang, et al., Time Course of Lung Changes on Chest Ct during Recovery from 2019 Novel Coronavirus (Covid-19) Pneumonia, Radiology, 2020, p. 200370.

[8] Linda Wang, Alexander Wong, Covid-net: A Tailored Deep Convolutional Neural Network Design for Detection of Covid-19 Cases from Chest X-Ray Images, 2020 arXiv preprint arXiv:2003.09871.
[9] Tongxue Zhou, Stéphane Canu, Ruan Su, An Automatic Covid-19 Ct Segmentation Network Using Spatial and Channel Attention Mechanism, 2020 arXiv preprint arXiv:2004.06673.
[10] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, Ling Shao, Inf-net: automatic covid-19 lung infection segmentation from ct images, IEEE Trans. Med. Imag. (2020).
[11] Chuansheng Zheng, Xianbo Deng, Qing Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, Xinggang Wang, Deep Learning-Based Detection for Covid-19 from Chest Ct Using Weak Label, medRxiv, 2020.
[12] Xi Ouyang, Jiayu Huo, Liming Xia, Fei Shan, Jun Liu, Zhanhao Mo, Fuhua Yan, Zhongxiang Ding, Yang Qi, Bin Song, et al., Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia, IEEE Trans. Med. Imag. (2020).
[13] Bhawana Kamble, Satya Prakash Sahu, Rajesh Doriya, A review on lung and nodule segmentation techniques, in: Advances in Data and Information Sciences, Springer, 2020, pp. 555–565.
[14] Awais Mansoor, Ulas Bagci, Ziyue Xu, Brent Foster, Kenneth N. Olivier, Jason M Elinoff, Anthony F. Suffredini, Jayaram K. Udupa, Daniel J. Mollura, A generic approach to pathological lung segmentation, IEEE Trans. Med. Imag. 33 (12) (2014) 2293–2310.
[15] Jianhua Yao, Andrew Dwyer, Ronald M. Summers, Daniel J. Mollura, Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification, Acad. Radiol. 18 (3) (2011) 306–314.
[16] Humera Shaziya, K. Shyamala, Raniah Zaheer, Automatic lung segmentation on thoracic ct scans using u-net convolutional network, in: International Conference on Communication and Signal Processing (ICCSP), Pages 0643–0647, IEEE, 2018, 2018.
[17] Tianyi Zhao, Dashan Gao, Jiao Wang, Zhaozheng Tin, Lung segmentation in ct images using a fully convolutional neural network with multi-instance and conditional adversary loss, in: IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 505–509, 2018.
[18] Maurizio Corbetta, Gordon L. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nat. Rev. Neurosci. 3 (3) (2002) 201–215.
[19] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., Recurrent models of visual attention, in: Advances in Neural Information Processing Systems, 2014, pp. 2204–2212.
[20] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[21] Sanghyun Woo, Jongchan Park, , Joon-Young Lee, Kweon In So, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 3–19.
[22] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Li Cheng, Honggang Zhang, Xiaogang Wang, Xiaoou Tang, Residual attention network for image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[23] Hanchao Li, Pengfei Xiong, Jie An, Lingxue Wang, Pyramid Attention Network for Semantic Segmentation, 2018 arXiv preprint arXiv:1805.10180.
[24] Yu Fisher, Vladlen Koltun, Multi-scale Context Aggregation by Dilated Convolutions, 2015 arXiv preprint arXiv:1511.07122.
[25] Zhengyang Wang, Shuiwang Ji, Smoothed dilated convolutions for improved dense prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2486–2495.
[26] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, Hannaneh Hajishirzi, Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 552–568.
[27] Hyojin Park, Youngjoon Yoo, Geonseok Seo, Dongyoon Han, Sangdoo Yun, Nojun Kwak, Concentrated-comprehensive Convolutions for Lightweight Semantic Segmentation, 2018 arXiv preprint arXiv:1812.04920.
[28] Panqu Wang, Pengfei Chen, Ye Yuan, Liu Ding, Zehua Huang, Xiaodi Hou, Garrison Cottrell, Understanding convolution for semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1451–1460.
[29] Weiyi Xie, Colin Jacobs, Jean-Paul Charbonnier, Bram van Ginneken, Relational modeling for robust and efficient pulmonary lobe segmentation in ct scans, IEEE Trans. Med. Imag. 39 (8) (2020) 2664–2675.
[30] Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Li Kang, Ning Huang, Shaoting Zhang, A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images, IEEE Trans. Med. Imag. 39 (8) (2020) 2653–2663.
[31] Xiaocong Chen, Lina Yao, Yu Zhang, Residual Attention U-Net for Automated Multi-Class Segmentation of Covid-19 Chest Ct Images, 2020 arXiv preprint arXiv: 2004.05645.
[32] Longxi Zhou, Zhongxiao Li, Juexiao Zhou, Haoyang Li, Yupeng Chen, Yuxin Huang, Dexuan Xie, Lintao Zhao, Ming Fan, Shahrukh Hashmi, 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.
[33] Yu Qiu, Yun Liu, Jing Xu, Miniseg: an Extremely Minimum Network for Efficient Covid-19 Segmentation, 2020 arXiv preprint arXiv:2004.09750.
[34] Olaf Ronneberger, Philipp Fischer, Thomas 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.
[35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[36] Ozan Oktay, Schlemper Jo, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Y. Nils, Hammerla, Bernhard Kainz, et al., Attention U-Net: Learning where to Look for the Pancreas, 2018 arXiv preprint arXiv:1804.03999.
[37] Jun Ma, Ge Cheng, Yixin Wang, Xingle An, Jiantao Gao, Ziqi Yu, Minqing Zhang, Xin Liu, Xueyuan Deng, Shucheng Cao, Hao Wei, Mei Sen, Xiaoyu Yang, Ziwei Nie, Li Chen, Tian Lu, Yuntao Zhu, Qiongjie Zhu, Guoqiang Dong, Jian He, COVID-19 CT Lung and Infection Segmentation Dataset, April 2020.
[38] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: a nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.
[39] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[40] Liang-Chieh Chen, Papandreou George, Florian Schroff, Hartwig Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, 2017 arXiv preprint arXiv: 1706.05587.
[41] Yixin Wang, Yao Zhang, Yang Liu, Tian Jiang, Zhong Cheng, Zhongchao Shi, Yang Zhang, Zhiqiang He, Does non-covid-19 lung lesion help? investigating transferability in covid-19 ct image segmentation, Comput. Methods Progr. Biomed. 202 (2021) 106004.
[42] Qingsen Yan, Bo Wang, Dong Gong, Chuan Luo, Wei Zhao, Jianhu Shen, Qinfeng Shi, Shuo Jin, Liang Zhang, You Zheng, Covid-19 Chest Ct Image Segmentation–A Deep Convolutional Neural Network Solution, 2020 arXiv preprint arXiv:2004.10987.
[43] Jun Ma, Yixin Wang, Xingle An, Ge Cheng, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al., Toward data-efficient learning: a benchmark for covid-19 ct lung and infection segmentation, Med. Phys. 48 (3) (2021) 1197–1210.