A dual-attention V-network for pulmonary lobe segmentation in CT scans

Shaohua Zheng1 | Weiyu Nie1 | Lin Pan1 | Bin Zheng2 | Zhiqiang Shen1 | Liqin Huang1 | Chenhao Pei1 | Yuhang She3 | Liuqing Chen1

1 College of Physics and Information Engineering, Fuzhou University, Fuzhou, China
2 Thoracic Department, Fujian Medical University Union Hospital, Fuzhou, China
3 College of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China

Correspondence
Lin Pan, College of Physics and Information Engineering, Fuzhou University, No. 2, Xueyuan Road, Shangjie Town, Minhou County, Fuzhou, Fujian Province 350108, China
Email: panlin@fzu.edu.cn

Funding information
Natural Science Foundation of Fujian Province, Grant/Award Number: 2020J01472; Provincial Science and Technology Leading Project, Grant/Award Number: 2018Y0032

Abstract
The reliable and automatic segmentation of pulmonary lobes in computed tomography scans is an important pre-condition for the diagnosis, assessment, and treatment of lung diseases. However, due to the incomplete lobar structures and morphological changes caused by diseases, the lobe segmentation still encounters great challenges. Recently, convolution neural network has exerted a tremendous impact on medical image analysis. Nevertheless, the basic convolution operations mainly obtain local features that are insufficient for accurate lobe segmentation. The idea that the global features are equally crucial especially when lesions appear is considered. Here, a dual-attention V-network named DA V-Net for pulmonary lobe segmentation is proposed. First, a novel dual-attention module to capture global contextual information and model the semantic dependencies in spatial and channel dimensions is introduced. Second, a progressive output scheme is used to avoid the vanishing gradient phenomenon and obtain relatively effective features in hidden layers. Finally, an improved combo loss is devised to address input and output lobe imbalance problem during training and inference. In the evaluation using the LUNA16 dataset and our in-house dataset, the proposed DA V-Net obtains Dice similarity coefficients of 0.947 and 0.934, respectively; these values are superior to those obtained by existing methods.

1 INTRODUCTION

The human lung consists of five compartments called lobes. Between two adjacent lung lobes is a faintly visible fissure that appears as a white crack in lung computed tomography (CT) scans. The left lung is separated by the left oblique fissure into two lobes called the left upper lobe and left lower lobe. The right lung is separated by the right horizontal and right oblique fissures into three lung lobes called the right upper lobe, right middle lobe, and right lower lobe. The structure of the human lung is shown in Figure 1. Each lobe is anatomically independent and has its own tracheal and vascular tree. The axial, coronal, and sagittal CT scans of a lung are shown in Figure 2.

The automatic and reliable segmentation of pathologic pulmonary lobes is important as different lung diseases usually occur in specific lobes. For example, tuberculosis, cystic fibrosis, pulmonary nodules, and centrilobular emphysema mainly appear in the upper lobes [1] while panlobular emphysema and interstitial pneumonia significantly affect the lower lobes [2]. The lobe-wise analysis of diseases can be used to assess and quantify the severity of these diseases and provide guidance for further surgery. During diagnosis, radiologists need to carefully observe the fissures to be able to identify the infected lobes and manually annotate the lobes to aid the analysis. This process is time consuming and cumbersome. Therefore, automatic and accurate lobe segmentation is critical as a pre-condition for radiologists to quickly locate affected lobes with diseases and efficiently examine the CT scans of lungs [3, 4].

However, pulmonary lobe segmentation still remains a challenge. First, lobar fissures may not fully extend to the lobar boundaries and may thus be incomplete [5]. According to statistics, lobar fissures may be over 50% incomplete [6, 7]. Second, other fissures, such as accessory fissures [8] and azygos fissures, may be misinterpreted as lobar fissures. Third, lung lesions...
FIGURE 1  Representations of 3D lung anatomy with lobes and fissures. Each lobe can be seen as an independent functional unit.

may cause morphological changes in lobar fissures. The recent spread of the novel coronavirus (COVID-19) plunged the world into crisis. CT scans of the lungs of COVID-19 patients present ground-glass shadows, and even the lobes or the whole lungs are filled with pleural fluid, thus making parts of the lobar fissures invisible. Finally, no specific standards exist for lung CT scans due to the differences in CT scanners, operators, and specific needs of radiologists. The three points mentioned above will reduce the accuracy of automatic pulmonary lobe segmentation.

Several methods have been proposed to segment the lobes in CT scans and attempted to overcome the aforementioned difficulties. Traditional methods formulate lobe segmentation as a fissure detection task; these methods include atlas registration [9], graph searching with shape constraints [10], watershed [11], and curve fitting [12–14]. There are some reviews on lung segmentation and pulmonary lobe segmentation [15, 16]. These methods achieve relatively good results in cases of visible fissures. For incomplete fissures, existing studies employed interpolation algorithms or utilized the anatomical information of the lung as prior knowledge (i.e. combination of the anatomical characteristics of the fissures, trachea, and vessels to comprehensively generate the segmentation of the lobes).

However, due to the existence of structures similar to lobar fissures, fissure detection is inevitably prone to false positive results, and segmentation results are not ideal, especially given severe pathologies.

In recent years, deep learning, or convolutional neural network (CNN), has exerted tremendous impact on medical image analysis. As the earliest deep learning model for pixel-wise segmentation, fully convolutional networks (FCNs) [17] opened a new era of image segmentation. Relative to traditional segmentation methods, FCNs have features that can be automatically learned from data in an end-to-end manner. Following closely, Ronneberger et al. [18] presented an encoder–decoder structure network with skip connection called U-Net, which has since become the most popular method in 2D medical image segmentation. U-Net was then extended to 3D for volumetric medical images. Several new networks, such as 3D U-Net [19], V-Net [20], and other 3D networks, have since emerged.

A number of CNN-based methods for pulmonary lobe segmentation have been proposed. George et al. [21] used a 2D FCN to detect lobar boundaries and then segmented the lobes via a 3D random walk. The limitation of this method is that it is only applicable to 2D slices and it does not consider 3D contextual information. Ferreira et al. [22] presented a model of lobe segmentation based on V-Net called FRV-Net, which employs additional regularization techniques to mitigate overfitting. Park et al. [23] and Lassen et al. [24] later proposed a lung lobe segmentation strategy that is based on 3D U-Net. Imran et al. [25] introduced a progressive dense V-Net. However, these CNN-based methods mainly capture local features without considering the global contextual information. The actual effective receptive field in a CNN may be much smaller than the results of calculations [26]. Apart from exploiting local features to detect the position of lobar fissures, the relative positions between lobes should also be learned through global features. Especially in cases involving lobar fissures that are fuzzy or incomplete, global information is important as the anatomical information of the trachea, vessels, and lung shape can help guide the segmentation of pulmonary lobes.

Attention is a helpful mechanism that leads a CNN to localize the most prominent area of previous feature maps. The
advantages of this mechanism make it widely used in many classification and segmentation tasks [27–30]. There are also some studies that have used attention mechanisms in the field of medical imaging [31–33]. Inspired by these beneficial works, we consider the idea that attention mechanisms can aid in obtaining the global contextual features of pulmonary lobes. In this work, we use V-Net as the baseline of the segmentation network to generate voxel-wise feature maps. We propose to add a 3D dual-attention (DA) module that can help obtain the spatial and channel global feature dependencies. We call this approach the dual-attention V-network (DA V-Net). Moreover, we propose a progressive output (PO) scheme and an improved combo loss (ICL) to further improve the performance of DA V-Net in pulmonary lobe segmentation. The evaluation of the model using two datasets shows that the experimental results represent high accuracy and robustness.

The main contributions of this work can be summarized as follows:

- To capture the global feature dependencies of lung lobes and solve the limitations of 2D networks, we extend the dual-attention module [29] to 3D scenario, called the 3D DA module, which can adaptively capture the global context of lobes from 3D lung CT scans.
- To address the input and output lobe imbalance problem during training and inference, we propose the ICL, which can help achieve high accuracy and low false positive/false negative rates in lobe segmentation.
- To avoid the vanishing gradient phenomenon and rapidly train the proposed network, we present an enhanced version of deep supervision called the progressive output scheme, which progressively aggregates the information between the final output and the side outputs. This scheme obtains relatively effective features in hidden layers.

2 | METHOD

The overview of the proposed method is shown in Figure 3. We propose DAV-Net for reliable and automatic segmentation of pulmonary lobes. We first specify the core module of DAV-Net: the 3D DA module, in Section 2.1. Then, we present the complete network architecture and describe the progressive output scheme in Section 2.2. Finally, the loss function is discussed in detail in Section 2.3.

2.1 | 3D dual-attention module

Although current medical image segmentation networks, such as U-Net [18] and V-Net [20], can achieve relatively good effects, they suffer from a certain flaw. That is, they ignore the long-range contextual dependencies of medical images. When encountering lung lobes, these networks’ segmentation effect is not ideal. The main reasons are explained as follows. In different lung CT scans, lobes are diverse in scale, position, and views. Moreover, the features coming from a convolution operation only contain a local receptive field, which may lead to different features of pulmonary lobes. These differences will lead to intra-class inconsistencies and poor segmentation results. Hence, expanding the difference between different lobes and background areas in feature maps and enabling different lung lobes to benefit one another remain difficult. To address this problem, we propose the 3D DA module, which could adaptively aggregate long-range contextual information and improve the feature representation of lobes.

The 3D DA module can add further global contextual information to the original network and derive feature representations for voxel-wise segmentation. As illustrated in Figure 4, the 3D DA module consists of two types of attention.

---

**FIGURE 3** Architecture of proposed DAV-Net. The network employs an encoder to extract features and a decoder to generate segmentation results. The 3D DA module is adopted to capture the global contextual features of lobes, and the PO scheme is introduced to progressively fuse the outputs of different decoder layers.
modules, namely, the position attention module (PAM) and channel attention module (CAM). The feature maps produced by previous layers are passed to these two parallel attention modules. Take the PAM in the upper branch of Figure 4 as an example. A convolutional layer is first employed to obtain the initial features of the PAM. Then, new feature maps of spatial long-range contextual information are generated via three steps. First, a spatial attention matrix is produced to model the spatial relationship between any two voxels of feature maps. Second, matrix multiplication is performed on the spatial attention matrix and initial feature matrix. Third, a voxel-wise sum operation is performed on the above multiplied resulting matrix and initial features to obtain the output result of the PAM. The CAM also obtains new channel contextual features through similar operations. The difference between the CAM and the PAM is that the attention matrix it obtains represents the relationship between the channels. Finally, we combine the output of the two attention modules with a voxel-wise sum operation as feature representations for a rich voxel-level prediction.

2.1.1 Position attention module

Feature representation is significant in medical image analysis and is achieved by capturing long-range contextual information. However, related studies [26, 34] denoted that the actual receptive field may be much smaller than the calculated results and that the local features produced by ordinary convolution may lead to the misclassification of targets. Therefore, we introduce the positional attention module to encode contextual information and thereby establish the global feature representation ability.

As illustrated in Figure 4, the PAM models long-range dependencies in the spatial dimension, with local feature map $A \in \mathbb{R}^{H \times W \times D \times C}$ as the input. After passing the $1 \times 1 \times 1$ convolution layer, $A$ is first converted into two feature maps $B_1, B_2 \in \mathbb{R}^{H \times W \times D \times C}$ and are then reshaped into new feature spaces $\mathbb{R}^{N \times C}$, where $N = H \times W \times D$ is the number of voxels. Thereafter, matrix multiplication is performed between $B_1$ and the transpose of $B_2$, and a softmax layer is applied to normalize the position attention maps $\hat{s}_p \in \mathbb{R}^{N \times N}$

$$\hat{s}_p = \frac{\exp(B_{1,i} \cdot B_{2,j})}{\sum_{i=1}^{N} \exp(B_{1,i} \cdot B_{2,j})},$$  

(1)

where $\hat{s}_p$ denotes the similarity of the $i$th and $j$th positions. The more similar the feature representation of the two positions is, the greater the correlation between them will be.

$A$ is also converted into a new feature map $C \in \mathbb{R}^{H \times W \times D \times C}$ through a convolutional layer and is reshaped into $\mathbb{R}^{N \times C}$. Matrix multiplication is performed between $\hat{s}_p$ and $C$, resulting in a new feature attention map $\gamma$

$$\gamma_j = \sum_{i=1}^{N} \hat{s}_{p_{ij}} C_i,$$  

(2)

FIGURE 4 Overview of 3D dua-attention module
where $\gamma_j$ represents the $j$th position feature vector and $\hat{J}_j$ indicates the importance of the $i$th position to $j$th position.

The final feature $Y \in \mathbb{R}^{H \times W \times D \times C}$ of the PAM is a voxel-wise weighted summation result of the attention feature map and initial feature map; that is,

$$Y_j = \eta^p \gamma_j + A_j,$$

where the scale parameter $\eta^p$ is initially set to 0 and gradually learns additional weight through backpropagation. Therefore, the PAM can make the initial feature map further focus on the important position area according to the global context information brought by the position attention map.

### 2.1.2 Channel attention module

Unlike the PAM that emphasizes the relationship between positions, the CAM pays attention to the interrelationship between channel maps. Each channel map can be viewed as a response to a certain category. With the use of the dependencies between channels, the semantic features on different channels promote one another, thereby highlighting the important parts of each channel. We directly calculate the CAM under the feature map $A \in \mathbb{R}^{H \times W \times D \times C}$. $A$ is first reshaped into a new feature space $\mathbb{R}^{N \times C}$. Matrix multiplication is performed between the transpose of $A$ and $A$, and softmax is applied to normalize the channel attention map $\gamma \in \mathbb{R}^{C \times C}$

$$\gamma_{ji} = \frac{\exp (A_i \cdot A_j)}{\sum_{j=1}^{C} \exp (A_i \cdot A_j)},$$

where $\gamma_{ji}$ represents the importance of the $i$th channel to $j$th channel. Matrix multiplication is performed between $\gamma$ and $A$ to obtain a new feature attention map. Thereafter, we reshape the attention map back to the original feature space $\mathbb{R}^{H \times W \times D \times C}$ and multiply it by the feature scale parameter $\eta^c$. A voxel-wise sum operation is then performed on the result and original feature. The final feature of the CAM is $Z \in \mathbb{R}^{H \times W \times D \times C}$

$$Z_j = \eta^c \sum_{i=1}^{C} (\gamma_{ji} A_i) + A_j,$$

where the scale parameter $\eta^c$ gradually learns more weight from 0.

Finally, we exploit a voxel-wise sum operation on the output of the PAM and CAM to form the 3D DA module. The 3D DA module enables DAV-Net to obtain global context information in the position and channel dimensions and thus greatly enriches its feature representation capability.

### 2.2 Network architecture

We propose a high-accuracy DAV-Net, which segments pulmonary lobes in lung CT scans. As shown in Figure 3, our model receives 3D lung CT scans as input images and outputs the voxel predictions of six target classes (background, left upper lobe, left lower lobe, right upper lobe, right middle lobe, right lower lobe). This network mainly consists of two parts, namely, encoder and decoder. The encoder contains five layers (i.e. E1, E2, E3, E4, E5) to obtain deep semantic information via feature extractors. We only use one convolutional layer in the E1 layer to extract initial features. To enlarge the receptive field, we employ a progressive dilated residual block (PDRB) as the basic operation unit in other encoder layers for feature extractors (Figure 5). The idea of the PDRB designed herein is based on [35, 36]; the PDRB contains three convolutional layers with dilatation rate $r = (1,2,3)$. Then, the input and output features are fused in a sum operation as the final result. The PDRB can effectively expand the covered receptive field by gradually increasing the dilatation rate. It can also alleviate the vanishing gradient phenomenon through the residual mechanism and improve network performance without additional parameters.

After passing through the encoder, feature information is inevitably lost due to downsampling. At this time, the decoder restores the features to the original size through deconvolution. The decoder we designed contains four layers (D4, D3, D2, D1). The rich encoding features of the network gradually propagate through the decoder to higher resolution layers. We also use the PDRB as the basic unit of the decoder. At the end of the network, we utilize the softmax activation function to generate the final segmentation results of the lung decoder. In addition, the skip connection is exploited to concatenate the features between the encoder and the decoder. To capture the global contextual features, we add the 3D DA module to the last two layers of the skip connection. We also use the PO scheme to integrate the pulmonary lobe features at different scales and thereby make the learnable network parameters increasingly effective. Two of the main contributions of this work are to improve the effectiveness of the network through the 3D DA module and PO scheme. The functions of these two techniques are as follows.

In Section 2.1, we have detailed the principle of the 3D DA module, which can capture the long-range dependencies of global feature maps in the spatial and channel dimensions and thereby highlight the boundary of pulmonary lobe areas. Inspired by [37], we consider the idea that the encoder features represent simple features as they are calculated in the shallow layers and that the decoder features represent complex semantic features as they are generated in the deep layers. The two sets of features that are merged directly using skip connection may
have a large semantic gap. Thus, we propose to replace the skip connection with the 3D DA module. These additional attention mechanisms are expected to effectively aggregate the features between the encoder and the decoder and provide global pivotal information for pulmonary lobe segmentation. Note that we only employ the 3D DA module on the last two skip connections to balance the tradeoff between the feature representation capacity and the GPU memory.

The PO scheme can produce a refined pulmonary lobe segmentation by reliably aggregating outputs from different network stages. Specifically, the PO scheme is an enhanced version of deep supervision [38, 39]; it not only supervises the hidden layers to guide training by calculating the loss of the side outputs in the intermediate stage but also gradually merges the predictions from different network stages through voxel-wise sum operations. These features allow the predictions of different levels and scales efficiently contribute to the final result. The latter prediction can progressively improve the previous one and thereby promote the integration of multiscale spatial information in the training process.

2.3 Loss function

Inspired by the discussion of the loss function for the imbalanced category [40], we propose an ICL for the pulmonary lobe segmentation task. Overall, our loss function is a weighted sum of modified dice loss and modified cross-entropy. We employ the LUNA16 dataset for training.

\[
\mathcal{L} = \alpha \left( -\frac{1}{CN} \sum_{c=1}^{C} \sum_{n=1}^{N} \beta \left( t_{nc} \ln p_{nc} \right) \right) \\
+ \left[ (1 - \beta) \left[ \left(1 - t_{nc} \right) \ln \left(1 - p_{nc} \right) \right] \right) \\
- (1 - \alpha) \frac{2 \sum_{c=1}^{C} \omega_c \sum_{n=1}^{N} p_{nc} t_{nc} + S}{\sum_{c=1}^{C} \omega_c \sum_{n=1}^{N} \left(p_{nc} + t_{nc} \right) + S},
\]

where \( \alpha \in [0, 1] \) controls the contribution of the dice term to the loss function \( \mathcal{L} \) and \( \beta \in [0, 1] \) controls the level of model punishment for the false positive/negative. When \( \beta \) is set to less than 0.5, the FP is punished more than FN as the term \((1 - t_{nc}) \ln(1 - p_{nc})\) has greater weight, and vice versa. \( C \) refers to the total number of classes, which is equal to six (five lobes and one background class) in our task. \( N \) indicates the total number of voxels in each mini-batch, and \( n \) denotes the index of each voxel. \( t_{nc} \) is the ground truth of voxel \( n \) on category \( c \), and \( p_{nc} \) is the corresponding predicted probability of voxel \( n \) on category \( c \). \( S \) represents a small number to avoid the zero division.

The difference between our loss and the original combo loss is that we additionally consider the different lobe sizes and the frequency of its corresponding class (i.e. the additional probability \( \omega_c = \frac{1}{\sum_{c=1}^{C} t_{nc}} \) for each class).

Similarly, for the \( D \) side output of the lung lobe, the loss \( \mathcal{L}_d \) is shown as follows:

\[
\mathcal{L}_d = \left( \sum_{d=1}^{D} \alpha \right) \left( -\frac{1}{CN} \sum_{c=1}^{C} \sum_{n=1}^{N} \beta \left( t_{nc} \ln p_{nc} \right) \right) \\
+ \left[ (1 - \beta) \left[ \left(1 - t_{nc} \right) \ln \left(1 - p_{nc} \right) \right] \right) \\
- (1 - \alpha) \frac{2 \sum_{c=1}^{C} \omega_c \sum_{n=1}^{N} p_{nc} t_{nc} + S}{\sum_{c=1}^{C} \omega_c \sum_{n=1}^{N} \left(p_{nc} + t_{nc} \right) + S}/D,
\]

Finally, the total loss function used in the entire network training phase is as follows:

\[
\mathcal{L}_{total} = \lambda_1 \mathcal{L} + \lambda_2 \mathcal{L}_d.
\]

3 EXPERIMENT AND RESULT

We validated the effectiveness of our proposed method on two datasets, that is, LUNA16 dataset and our in-house dataset. In this section, we first introduce the datasets. Then, we introduce an evaluation metric to measure the accuracy of lobe segmentation. Next, we specify the implementation details including pre- and post-processing, training parameters, and system settings. After that, we compare the proposed method with some state-of-the-art methods. Finally, we conduct a series of ablation studies to analyze the effect of each component of proposed method.

3.1 Datasets

The CT scans used in this work come from two sources. We refer to the first as the Lung Nodule Analysis 16 (LUNA16) set and to the second as our in-house set.

LUNA16 is a subset of the LIDC/IDRI dataset and was first used for lung nodule detection tasks. We use 50 cases of public lobe annotations [41], 40 cases for training, and the remaining 10 cases for testing. These 3D CT scans are obtained with different scanners, sensitivities, imaging protocols, and reconstruction kernels. The slice thickness ranges from 0.625 to 2.5 mm, and the internal resolution of the slice ranges from 0.545 to 0.778 mm. Some scans are full of noise. In this work, we only employ the LUNA16 dataset for training.

Our in-house dataset is provided by the hospital, and the reference annotations are manually labeled by a radiologist. We randomly select 10 cases from this dataset as the test set to verify the robustness of the model. The slice thickness of the scans is 1 mm, and the internal resolution of the slice ranges from 0.35 to 0.41 mm. This dataset includes healthy and pathological lungs.
3.2 | Evaluation metric

We adopt the Dice similarity coefficient (Dice) to quantitatively evaluate the segmentation performance. The Dice is used to calculate the similarity between the ground truth and the predicted segmentation result [20]. The formula is as follows:

\[
DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|},
\]

where \( A \) denotes the predicted lobe segmentation result, \( B \) indicates the reference of the lobe, and \(|A \cap B|\) is the number of intersecting voxels between \( A \) and \( B \). We calculate the Dice for each lobe and the average Dice for all lobes as the final evaluation criterion.

3.3 | Implementation details

3.3.1 | Pre- and post-processing

To minimize the interference of irrelevant information in the CT scans of lungs, we set the threshold and combine the connected domains to extract the lung parenchyma area. This step is performed in the subsequent training and testing. All training and testing CT scans are trimmed by Hounsfield units to [-1024, -300] and then normalized by z-score standardization. Within this range, the relevant information of lobar fissures could be fully retained. Thereafter, all scans are downsampled to a uniform size of \( 256 \times 256 \times 128 \). Considering the limitation of the GPU memory, we randomly sample from this size to \( 128 \times 128 \times 64 \) patches as the inputs of the network. This approach could also be regarded as a way to augment data, in which case no details are lost.

3.3.2 | Training parameters

We utilize the uniform distribution initialization method [42] to initialize the learnable parameters of each convolutional layer and train all models from scratch. A total of 3000 epochs with 40 patches are trained in each epoch. We use the Adam optimizer [43] by setting \( \beta_1 = 0.9, \beta_2 = 0.999 \). Given the limitation of the GPU memory, we set the batch size to 1. In the experiment, the initial learning rate is \( 10^{-5} \) with a decay of \( 10^{-7} \). In Equations (6) and (7), we set \( \alpha = 0.5, \beta = 0.5 \). The weights of the final output and side output for the loss function in Equation (8) are \( \lambda_1 = 0.75, \lambda_2 = 0.25 \).

3.3.3 | System settings

We model DAV-Net using Keras (v2.4.0) with TensorFlow (v1.12.0) in the backend. All training and testing experiments are run on a workstation with an Intel® Core™ i5-7500 CPU, 16 GB of RAM, and an NVIDIA 1080Ti GPU.

3.4 | Comparison with state-of-the-art methods

To validate the effectiveness of our proposed method, we compare the DA V-Net with previous state-of-the-art methods on the basis of the two datasets (LUNA16 test set and the in-house test set). Specifically, we evaluate the proposed DA V-Net with PTK [44], 3D U-Net [19, 23, 24], PDV-Net [25], and FRV-Net [22]. In our experiments, all methods adopt 3D CNNs under the same training set and hyperparameters, except for the PTK, which is a software suite for the analysis of 3D medical lung images in an unsupervised approach.

Table 1 reports the quantitative results of the proposed DA V-Net and state-of-the-art pulmonary lobe segmentation methods. The corresponding box-whisker plots are depicted in Figure 6. The results for the LUNA16 test set and our in-house test set show that the proposed DA V-Net performs greatly on each of the lung lobes and achieves average per-lobe Dice of 0.947 ± 0.067 and 0.934 ± 0.022, respectively. The standard deviations of the Dice produced by the proposed DA V-Net are also lower than those of other methods, thus indicating the robustness and generalizability of our method. In addition, the Dice of PTK only reaches 0.787 for LUNA16 set and 0.855 for in-house set, which is significantly worse than that of the method based on 3D CNN. Compared to PTK, the Dice performance of CNN-based methods all exceed 0.9. These results confirm the segmentation superiority of the CNN for pulmonary lobes. Moreover, it shows that the CNN can learn effective features in a supervised manner and address the diversity of CT scans with lung disease. For CNN-based methods, the proposed DA V-Net also performs
Ablation studies

The quantitative results of proposed DAV-Net and state-of-the-art methods on the LUNA16 (top) and in-house test sets (bottom) are shown in Table 1.

### Table 1: Quantitative results of proposed DAV-Net and state-of-the-art methods on the LUNA16 (top) and in-house test sets (bottom)

(a) LUNA16 results

| Method         | LUL       | LLL       | RUL       | RML       | RLL       | Overall   |
|----------------|-----------|-----------|-----------|-----------|-----------|-----------|
| PTK [44]       | 0.885 ± 0.208 | 0.863 ± 0.242 | 0.859 ± 0.148 | 0.551 ± 0.404 | 0.775 ± 0.345 | 0.787 ± 0.311 |
| 3D U-Net [19, 23, 24] | 0.923 ± 0.038 | 0.933 ± 0.039 | 0.930 ± 0.070 | 0.842 ± 0.121 | 0.930 ± 0.040 | 0.912 ± 0.078 |
| FRV-Net [22]   | 0.948 ± 0.026 | 0.946 ± 0.038 | 0.940 ± 0.070 | 0.866 ± 0.122 | 0.940 ± 0.052 | 0.928 ± 0.077 |
| PDV-Net [25]   | 0.931 ± 0.028 | 0.945 ± 0.027 | 0.935 ± 0.069 | 0.862 ± 0.121 | 0.950 ± 0.027 | 0.924 ± 0.073 |
| DAV-Net (ours) | 0.959 ± 0.019 | 0.961 ± 0.015 | 0.951 ± 0.065 | 0.905 ± 0.114 | 0.960 ± 0.050 | 0.947 ± 0.067 |

(b) In-house dataset results

| Method         | LUL       | LLL       | RUL       | RML       | RLL       | Overall   |
|----------------|-----------|-----------|-----------|-----------|-----------|-----------|
| PTK [44]       | 0.884 ± 0.123 | 0.878 ± 0.085 | 0.895 ± 0.102 | 0.739 ± 0.294 | 0.879 ± 0.157 | 0.855 ± 0.180 |
| 3D U-Net [19, 23, 24] | 0.903 ± 0.028 | 0.916 ± 0.033 | 0.918 ± 0.022 | 0.860 ± 0.039 | 0.929 ± 0.020 | 0.905 ± 0.038 |
| FRV-Net [22]   | 0.930 ± 0.017 | 0.931 ± 0.025 | 0.930 ± 0.014 | 0.895 ± 0.022 | 0.935 ± 0.020 | 0.924 ± 0.025 |
| PDV-Net [25]   | 0.912 ± 0.020 | 0.931 ± 0.027 | 0.925 ± 0.013 | 0.884 ± 0.027 | 0.951 ± 0.011 | 0.920 ± 0.030 |
| DAV-Net (ours) | 0.939 ± 0.009 | 0.944 ± 0.012 | 0.929 ± 0.011 | 0.901 ± 0.015 | 0.956 ± 0.007 | 0.934 ± 0.022 |

The Dice of each lobe and average lobe are given as mean and standard deviation. LUL, LLL, RUL, RML, RLL: Left upper, left lower, right upper, right middle, right lower lobes. Overall: per-lobe mean.

Figure 7 shows the qualitative segmentation results of different pulmonary lobe segmentation methods and the reference segmentations for comparison from three slice views and a 3D view. We exhibit the LUNA16 and in-house cases accordingly. Lobe segmentation is challenging for the cases with incomplete fissures, with the lobar boundaries being difficult to identify. Furthermore, the PTK method produces unsatisfactory results possibly because of its heavy dependencies on obvious fissure features. Although 3D U-Net, FRV-Net, and PDV-Net show better results than the PTK method, they still suffer from errors in the segmentation of details. We consider the possibility that the former methods might be insufficient to segment lobes effectively and only learn local position features. Nevertheless, the proposed DAV-Net can still accurately segment the lobes with smooth boundaries and utilize the relevant information of the position and channel to infer the shape of the lobes from the global context, thus obtaining satisfactory pulmonary lobe segmentation results.

### Table 2: Ablation study on the LUNA16 dataset for the PAM, CAM, PO scheme, and ICL into the baseline framework

| Method   | PAM | CAM | PO | ICL | Dice   |
|----------|-----|-----|----|-----|--------|
| Baseline | √   | √   |    |     | 0.919 ± 0.080 |
| w/o PAM  | √   | √   |    |     | 0.933 ± 0.074 |
| w/o CAM  | √   | √   |    |     | 0.940 ± 0.070 |
| w/o PO   | √   | √   |    |     | 0.944 ± 0.068 |
| w/o ICL  | √   | √   |    |     | 0.945 ± 0.068 |
| DAV-Net  | √   | √   |    |     | 0.947 ± 0.067 |

#### 3.5 Ablation studies

Table 2 shows the results of the ablation study for our proposed method and highlights the effect of each component applied to the model on the segmentation results. We evaluate the performance of each component by removing the PAM, CAM, PO scheme, and ICL from DAV-Net, respectively. The baseline model used is a V-Net structure with five layers in the encoder, with each layer employing the PDRB, except for the first one. Furthermore, we replace the ICL with the standard Dice to verify the function of our comprehensive weighted loss. To verify the PO scheme, we only supervise the final output, which removes the side outputs to prove the advantage of the progressive fusing features. We exploit the Dice as the evaluation metric and finally report the average and standard deviation of all lobes.

The proposed DAV-Net achieves the best results. The segmentation result of the baseline is relatively poor possibly because of the difficulty in extracting features in the training process and the nonconvergence of the network to the global optimal value. We observe that the PAM is vital for enhancing accurate lobe segmentation. Specifically, the PAM can obtain...
FIGURE 7  Qualitative comparison of segmentation results for LUNA16 (top) and the in-house (bottom) test cases in terms of slice and volume level. From left to right: input image, ground truth (GT), PTK, 3D U-Net, FRV-Net, PDV-Net, and proposed DAV-Net. From top to bottom: axial plane, sagittal planes, coronal planes, and 3D view. Color coding: yellow, LUL; cyan, LLL; pink, RUL; almond, RML; blue, RLL.
global position dependencies by adequately utilizing the relative position information of the trachea, vessels, and lobar fissures to identify the position features of the lung lobes. It also shows a 1.4% increase in the Dice for the test set. Therefore, we believe that the PAM can greatly expand the receptive field and obtain global contextual information effectively. Meanwhile, the CAM plays a crucial role in the model by improving the accuracy by nearly 1% for the test set. It is focused on the relationship between channel maps and can gradually increase the inter-class distance and reduce the intra-class distance to make the network particularly representative. In our research, the PO scheme can slightly increase the accuracy of the network as it could directly supervise the hidden layer with multiscale features and gradually aggregate the final output and side output information. The ICL also contributes to a benign improvement in the network. More importantly, this loss function could be regarded as a regularization technique to alleviate the imbalance of input and output without extra parameters. These results demonstrate that each proposed component of DA V-Net plays a positive role in lung lobe segmentation. Moreover, the 3D DA module contributes the most to the performance improvement of DA V-Net.

4 | CONCLUSION

In this work, we have proposed and evaluated a 3D deep learning network called DA V-Net for segmenting pulmonary lobes in lung CT scans. Inspired by the attention mechanism, the proposed DA V-Net takes the 3D DA module as the core building block. We have leveraged the 3D DA module to combine local and global features efficiently and help capture the global dependencies of the feature maps in the spatial and channel dimensions. These capabilities are conducive to highlighting the positions and categories of pulmonary lobes and are crucial to the network. Furthermore, we have proposed a PO scheme to progressively improve the output accuracy of the network and employed an ICL to address the imbalance problem of inputs and outputs. These results have demonstrated that the proposed DA V-Net has superior capabilities to segment pulmonary lobes in lung CT scans.

More importantly, the proposed DA V-Net exhibits robustness and accuracy even for scans with incomplete fissures or pathology. These contributions prove the prospect of applying deep learning to clinical lung lobe segmentation. We consider that DA V-Net can be migrated to other medical image segmentation tasks. A possible future research direction is to extend the network as a backbone to detect and segment lung nodules, which is an interesting topic in lung CT image analysis. In the future, we will collect more lung CT scans and their annotations from multiple hospitals for clinical application.

ACKNOWLEDGEMENTS

This project was supported by the Natural Science Foundation of Fujian Province, China (Grant No. 2020J01472) and Provincial Science and Technology Leading Project (Grant No. 2018Y0032).

REFERENCES

1. Nemec, S.F., Bankier, A.A., Eisenberg, R.L.: Upper lobe-predominant diseases of the lung. Am. J. Roentgenol. 200(3), W222–W237 (2013)
2. Nemec, S.F., Bankier, A.A., Eisenberg, R.L.: Lower lobe-predominant diseases of the lung. Am. J. Roentgenol. 200(4), 712–728 (2013)
3. Elizabeth, D.S., et al.: Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. IET Image Process. 6(6), 697–705 (2012)
4. Baby, Y.R., Sumathy, V.K.R.: Kernel-based Bayesian clustering of computed tomography images for lung nodule segmentation. IET Image Process. 14(5), 890–900 (2019)
5. Raszeh, B.N., et al.: Radiographic anatomy of the interlobar fissures: A study of 100 specimens. Am. J. Roentgenol. 138(6), 1043–1049 (1982)
6. Gulsun, M., et al.: Variability of the pulmonary oblique fissures presented by high-resolution computed tomography. Surg. Radiol. Anat. 28(3), 293–299 (2006)
7. Azia, A., et al.: High resolution CT anatomy of the pulmonary fissures. J. Thorac. Imaging 19(3), 186–191 (2004)
8. Cronin, P., et al.: Normal and accessory fissures of the lung Evaluation with contiguous volumetric thin-section multiphotodetector CT. Eur. J. Radiol. 75(2), e1–e8 (2010)
9. Zhang, L., Hoffman, E.A., Reinhardt, J.M.: Atlas-driven lung lobe segmentation in volumetric X-ray CT images. IEEE Trans. Med. Imaging 25(1), 1–16 (2006)
10. Zhang, L., Hoffman, E.A., Reinhardt, J.M.: Lung lobe segmentation by graph search with 3D shape constraints. In: Medical Imaging 2001: Physiology and Function from Multidimensional Images, pp. 204–215. San Diego, (2001)
11. Kuhlman, J.M., et al.: Lung lobe segmentation by anatomy-guided 3D watershed transform. In: Medical Imaging 2003: Image Processing, pp. 1482–1490. San Diego (2003)
12. Ross, J.C., et al.: Pulmonary lobe segmentation based on ridge surface sampling and shape model fitting. Med. Phys. 40(12), 121903 (2013)
13. Gu, S., et al.: Robust pulmonary lobe segmentation against incomplete fissures. In: Medical Imaging 2012: Computer-Aided Diagnosis, San Diego (2012)
14. Pu, J., et al.: Pulmonary lobe segmentation in CT examinations using implicit surface fitting. IEEE Trans. Med. Imaging 28(12), 1986–1996 (2009)
15. Mittal, A., Hooda, R., Sofat, S.: Lung field segmentation in chest radiographs: A historical review, current status, and expectations from deep learning. IET Image Process. 11(11), 937–952 (2017)
16. Doel, T., Gavaghan, D.J., Grau, V.: Review of automatic pulmonary lobe segmentation methods from CT. Comput. Med. Imaging Graph. 40, 13–29 (2015)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, pp. 3431–3440. Boston (2015)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Munich (2015)
19. Cieck, O., et al.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432. Athens (2016)
20. Milletari, F., Navab, N., Ahmadi, S.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision, pp. 565–571. Stanford (2016)
21. George, K., et al.: Pathological pulmonary lobe segmentation from CT images using progressive holistically nested neural networks and random walker. In: 3rd International Workshop on Deep Learning in Medical Image Analysis, pp. 195–203. Quebec City (2017)
22. Ferreira, F.T., et al.: End-to-end supervised lung lobe segmentation. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1–8. Rio de Janeiro (2018)
23. Park, J., et al.: Fully automated lung lobe segmentation in volumetric chest CT with 3D U-net: Validation with intra- and extra-datasets. J. Digit. Imaging 33(1), 221–230 (2020)
24. Lassen-Schmidt, B., et al.: Automatic segmentation of the pulmonary lobes with a 3D u-net and optimized loss function. arXiv [preprint], arXiv:2006.00083 (2020)
25. Imran, A., et al.: Fast and automatic segmentation of pulmonary lobes from chest CT using a progressive dense V-network. Comput. Methods Biomech. Biomed. Eng. 8, 509–518 (2019)
26. Luo, W., et al.: Understanding the effective receptive field in deep convolutional neural networks. In: 30th Annual Conference on Neural Information Processing Systems, pp. 4905–4913. Barcelona (2016)
27. Hu, J., et al.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42(8), 2011–2023 (2020)
28. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6458. Honolulu (2017)
29. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition, pp. 3141–3149. Long Beach (2019)
30. Li, H., et al.: Pyramid Attention network for semantic segmentation. arXiv [preprint], arXiv:1805.10180 (2018)
31. Oktay, O., et al.: Attention U-Net: Learning where to look for the pancreas. arXiv [preprint], arXiv:1804.03999 (2018)
32. Guan, Q., et al.: Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv [preprint], arXiv:1801.09927 (2018)
33. Schlemper, J., et al.: Attention-gated networks for improving ultrasound scan plane detection. arXiv [preprint], arXiv:1804.05338 (2018)
34. Peng, C., et al.: Large kernel matters—Improve semantic segmentation by global convolutional network. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1743–1751. Honolulu (2017)
35. Chen, L., et al.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2018)
36. Wang, P., et al.: Understanding convolution for semantic segmentation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 1451–1460. Lake Tahoe (2018)
37. Ibtehaz, N., Rahman, M.S.: MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 121, 74–87 (2020)
38. Lee, C., et al.: Deeply-supervised nets. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, pp. 562–570. San Diego (2015)
39. Dou, Q., et al.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 149–157. Athens (2016)
40. Taghanaki, S.A., et al.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput. Med. Imaging Graph. 75, 24–33 (2019)
41. Tang, H., Zhang, C., Xie, X.: Automatic pulmonary lobe segmentation using deep learning. In: 16th IEEE International Symposium on Biomedical Imaging, pp. 1225–1228. Venice (2019)
42. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 249–256. Sardinia (2010)
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations. San Diego (2015)
44. Doel, T., et al.: Pulmonary lobe segmentation from CT images using fission, airways, vessels and multilevel B-splines. In: 9th IEEE International Symposium on Biomedical Imaging, pp. 1491–1494. Barcelona (2012)

How to cite this article: Zheng S, Nie W, Pan L, et al. A dual-attention V-network for pulmonary lobe segmentation in CT scans. IET Image Process. 2021;15:1644–1654. https://doi.org/10.1049/ipr2.12133