Contextual Two-Stage U-Nets for Robust Pulmonary Lobe Segmentation in CT Scans of COVID-19 and COPD Patients

Weiyi Xie, Colin Jacobs, Jean-Paul Charbonnier, Bram van Ginneken

arXiv:2004.07443v1 [eess.IV] 16 Apr 2020

***Abstract*—Pulmonary lobe segmentation in computed tomography scans is essential for regional assessment of pulmonary diseases. Automated segmentation is still an open problem, especially for scans with substantial abnormalities, such as in COVID-19 infection. Recent works used Convolutional Neural Networks for automatic pulmonary lobe segmentation. Convolution kernels in these networks only respond to local information within the scope of their effective receptive field, and this may be insufficient to capture all necessary contextual information. We argue that contextual information is critically important for accurate delineation of pulmonary lobes, especially when the lungs are severely affected by diseases such as COVID-19 or COPD.**

In this paper, we propose a contextual two-stage U-net (CTSU-Net) that leverages global context by introducing a first stage in which the receptive field encompasses the entire scan and by using a novel non-local neural network module. The proposed module computes the filter response at one position as a weighted sum of feature responses at all positions, where geometric and visual correlations between features determine weights.

With a limited amount of training data available from COVID-19 subjects, we initially train and validate CTSU-Net on a cohort of 5000 subjects from the COPDGene study (4000 for training and 1000 for evaluation). Using models pretrained COPDGene, we apply transfer learning to retrain and evaluate CTSU-Net on 204 COVID-19 subjects (104 for retraining and 100 for evaluation). Experimental results show that CTSU-Net outperforms state-of-the-art baselines and performs robustly on cases with incomplete fissures and severe lung infection due to COVID-19.

***Index Terms*—Pulmonary Lobe, Segmentation, Computed Tomography, COVID-19, COPD, Convolution Neural Network, Global Context, Non-local Neural Networks.**

I. Introduction

THE human lungs consist of five disjoint pulmonary lobes. The right lung is composed of an upper, middle, and lower lobe, while the left lung only has an upper and a lower lobe. The lobes are separated by the pulmonary fissures, a double-fold of visceral pleura visible as a thin line on CT images. The lobes are functionally independent units because each has its own vascular and bronchial supply. As a result, the extent of disease often varies substantially across lobes, and lobe-wise assessment of pulmonary disorders is of great clinical importance. Computed Tomography (CT) is the best way to image the lungs in vivo.

In chronic obstructive pulmonary disease (COPD), one of the major causes of mortality and morbidity worldwide, measuring destruction due to emphysema at a lobar level, provides critical information for selecting target lobes for therapeutic intervention [1]. In COVID-19, the pandemic disease caused by the SARS-Cov2 virus that is straining healthcare systems worldwide, the severity of the disease are summarized in a CT severity score where each lobe is scored visually by radiologists on a scale from 0 to 5. These scores are then summed to quantify lung involvement on a scale from 0 to 25. The score provides a tool to assess disease severity and progression over time and is used for clinical decision making. To automate the CT severity score, lobe segmentation in COVID-19 scans is needed. CT scans of COVID-19 patients are affected by extensive patchy ground-glass region and consolidations and may even show lobes or complete lungs filled with pleural fluid. Automated lobe segmentation is highly challenging in scans with such extensive pathological changes.

Many automatic lobe segmentation approaches focused on finding visible fissures. Assuming that once the fissures are detected, the lobe segmentation can be derived by interpolation. Both early hand-crafted fissure enhancement filters [2]–[5] and more robust supervised learning methods [6], [7] relied heavily on local features, neglecting the global context. As a result, they often produced unsatisfactory results when pathological deformation, anatomical variation, or high noise levels are present. Moreover, because incomplete fissures are very common [8], interpolation of boundaries based on visible fissures may not suffice to find the lobe borders reliably. Therefore, contextual features were introduced [9]–[12] in the form of anatomical relations between lobes and nearby airways, vessels, and the lung borders. Most recently, LobeNet [13] proposed a fully convolutional neural network (FCN) for segmenting the lobes that uses segmentations of fissures and lung obtained from their previously proposed FCN methods. These approaches show that incorporating anatomical relations improves the robustness for the lobe segmentation. However, these approaches require prior segmentation of the lung or

This work was supported by the Dutch Lung Foundation. We acknowledge the COPDGene Study (ancillary study ANC-398) for providing the data used. COPDGene is funded by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors. It does not necessarily represent the official views of the National Heart, Lung, and Blood Institute of the National Institutes of Health. The COPD Foundation also supports the COPDGene project through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens, and Sunovion. (Corresponding author: Weiyi Xie. Contacting email: Weiyi.Xie@radboudumc.nl).

W. Xie, C. Jacobs, B. van Ginneken are with the Diagnostic Image Analysis Group, Radboudumc, Nijmegen, the Netherlands.

Jean-Paul Charbonnier is with Thirona, Nijmegen, the Netherlands.

other structures in the lung, thus requiring a multi-stage complex process. Each stage is then optimized separately, and therefore the final solution may be sub-optimal.

Based on fully convolution neural networks (FCNs), end-to-end lobe segmentation methods have been proposed [14], [15]. In these works, it was assumed that the local and contextual features for segmenting pulmonary lobes could be learned directly from the CT images without explicitly injecting contextual information. However, fully convolutional neural networks (FCN) may not be sufficient to harvest the global context. First, the effective receptive field may be much smaller than the theoretical receptive field due to the nature of the convolution [16]. Second, input CT images are mostly cropped before feeding into the networks to reduce GPU memory requirements, thus missing crucial scan-level semantics during the training. Third, in FCN-based methods, the context-dependency is often underused, especially long-range dependencies between objects and object parts, because intrinsically, the prediction for each voxel in an FCN is made independently of voxels outside its receptive field. However, contextual dependencies such as anatomical relations are essential in lobe segmentation.

Therefore, this paper focuses on capturing the global context in two ways. First, we introduce a two-stage U-net in which the first stage encapsulates the entire scan, and both stages are trained simultaneously in an end-to-end fashion. Second, we introduce a non-local neural network module that can be plugged into the two-stage U-Net. This module computes a feature response at one location using both appearance and geometric features from all other positions at the scan-level. We call this approach a contextual two-stage U-net, or CTSU-Net, for short.

The main contributions of this paper are as follows:

- We propose a lobe segmentation framework that uses a novel two-stage U-net architecture that can be trained end-to-end and contains a non-local neural network module for capturing the global context of a large 3D CT scan.
- The proposed CTSU-Net is robust and produces accurate lobe segmentations even for scans with severe pathology, and can be easily extended to other medical image analysis problems.
- CTSU-Net is fast and memory-efficient. It requires only a standard GPU with 12GB memory to train, and inference takes around 30 seconds to produce lung and lobe segmentations for a full thoracic CT scan.

A. Related Work

Fully convolutional network (FCN) based methods have become the standard in medical image semantic segmentation [17], [18]. The FCN approach can be thought of as sliding a classification network around a region in an input image, and each sliding region is then processed independently. Consequently, FCN are essentially local and do not take the full global context of an image into account. Atrous spatial pyramid pooling [19], [20] was later adopted to embed contextual information at different scales with dilated convolutions filters but suffering from the gridding artifacts [21]. Fully connected layers [22] were used to capture the global context in the segmentation of head and neck anatomy for radiotherapy. However, in this work, cropped CT scans were used as the input because of the large memory footprints of fully-connected layers. For aggregating contextual dependencies, [23] introduced recurrent neural networks to aggregate contextual information on the axial slices for cardiac segmentation in multi-slice MRI. A known issue with recurrent network networks is that they suffer from vanishing gradients [24] and therefore are hard to train. The global context could also be explicitly defined using Graph Models such as dense conditional random fields (CRF) [25]. However, due to their heavy computational demands, dense CRFs are often only used as the post-processing steps and optimized separately on a heuristic basis, making it hard for this approach to scale well.

Attention is widely used for various tasks such as machine translation, visual question answering, image and video classification, and semantic segmentation in natural images. Self-attention methods [26], [27] calculate the context encoding at one position by a weighted summation of embedding at all positions in sentences, which is a natural way to capture sentence context. As one of the self-attention applications, a non-local neural network was proposed for semantic segmentation [28] by computing a self-attention map for each feature based on all the other features in an input CNN feature map. The attention weights were determined by predefined similarity measurements between features in a linear-projected subspace. There are many recent extensions of this non-local method in semantic segmentation.

DANet [29] applies both spatial and channel attention to gather contextual information across channels of a feature map. Although non-local neural networks show superior performance on many benchmarks, their high computational costs limited their application to volumetric medical data. To reduce the computational intensity, the CCNet [30] was proposed. This network employed a simple criss-cross trick which reduces the space and time complexity of the non-local module from $\emptyset((H \times W) \times (H \times W))$ to $\emptyset((H \times W) \times (H + W - 1))$ in two-dimensional images. Our approach is motivated by the success of using self-attention in the above works for semantic segmentation. Our self-attention module uses the criss-cross trick, while also introducing geometric features.

II. Data

CT scans used in this study were obtained from two sources. We refer to the first set as the COPD set, and to the second set as the COVID-19 set.

A large set of scans from subjects with COPD, ranging from mild to very severe, was obtained from the COPDGene study [31]. This is a clinical trial with data from 21 imaging centers in the United States. In total, COPDGene enrolled 10,000 subjects. Each subject underwent both inspiration and expiration chest CT. Image reconstruction uses sub-millimeter slice thickness and in-plane resolution, with edge-enhancing and smooth algorithms. The CT protocols are detailed elsewhere [31]. Data from COPDGene is publicly available and

TABLE I: Characteristics of the two data sets used in this study. (a) lists the distribution of GOLD stages and other classes, see [31] in the COPD data set. (b) gives the distribution of CO-RADS scores across the training and test sets. CO-RADS score 1-6 indicates the level of suspicion for COVID-19 positive disease, ranging from very low, low, equivocal, high, very high, and confirmed PCR positive, respectively.

(a) COPD set GOLD stages

GOLD stages	# subjects for training	#subjects for testing
GOLD0	1709	433
GOLD1	319	80
GOLD2	734	184
GOLD3	441	110
GOLD4	226	57
Non Spirometry	30	2
Non Smoking	45	11
PRISm	496	123
Total	4000	1000

(b) COVID-19 set CO-RADS scores

CORADS	#subjects for training	#subjects for testing
1	24	23
2	10	9
3	20	20
4	17	16
5	24	24
6	9	8
Total	104	100

can be retrieved after submitting an ancillary study proposal (ANC-398 was used for this work).

We randomly selected 5000 subjects and used only Phase I inspiration CT scans (one scan per subject). Subjects were randomly grouped into a training set (n = 4000) and a test set (n = 1000). Slice thickness ranged from 0.625-0.9mm and pixel spacing from 0.478-1.0mm. Scans were performed using 200mAs and a tube voltage of 120kVP.

The other data set was obtained from Radboud University Medical Center, Nijmegen, the Netherlands. On March 18, 2020, this institution implemented a low-dose non-contrast CT protocol and all patients who arrived at the hospital with suspicion of COVID-19 disease and inpatients for whom COVID-19 was considered a possibility underwent CT. To date, 518 CT scans have been made. We only included scans from subjects who did not object to the use of their scans for research purposes. All scans were anonymized and permission for research use was obtained from our review board (file number CMO 2016-3045, Project 20027). It is unclear at this stage if we will be allowed to publicly share these scans. We will file a request for this later.

We randomly selected 204 subjects and use one scan per subject by selecting the CT scan of the smallest slice thickness in a study. Scans have a sub-millimeter pixel spacing and a slice thickness of 0.5 mm. 104 of these scans were used for training and the other 100 for testing.

See Table I for the distribution of GOLD stages in the training and the test set for the COPD set and the distribution of CO-RAD scores from the COVID-19 set. The CO-RAD scores defines the level of suspicion COVID-19 and was reported by in the radiology reports. Note that PCR status was not yet available for most subjects.

From the two training data sets, we selected 100 scans as the validation set for the COPD set, and we selected 24 scans for validation from the COVID-19 set. set for retraining all the models. Train, validation and test partitioning ensures the distribution of either GOLD stages or CO-RAD scores distributed similarly across partitions.

A. Reference Standard

Lobe segmentation references were obtained from Thirona, a company that specializes in chest CT analysis. First, automated segmentation of the left and right lung was generated using a commercialized software (LungQ, Thirona, Nijmegen, NL), followed by manual refinement if needed. Second, automatic algorithms [6], [10], [32] were used to extract the lobar boundary with possible interpolation for incomplete fissures using nearby airways and vasculature information. Next, the automatically found lobar boundaries were manually corrected separately for the left and the right lung, by trained analysts with at least one year experience in annotating the pulmonary structures on CT. Analysts are medical students who received training in lung anatomy and CT imaging and were instructed to draw a complete lobar boundary if needed. Once the manual modifications are made, the lobar boundaries are updated, possibly followed by final corrections.

III. Methods

We define the lobe segmentation problem as a voxel-wise classification problem. Given a scan I, the goal is to predict the voxel label $\hat{l_i}$ for every spatial location i, where $\hat{l_i} \in$ the label set $L = \{0, 1, 2, 3, 4, 5\}$ representing the background, left upper, left lower, right upper, right lower, and the right middle lobe, respectively.

This paper proposes a lobe segmentation framework consisting of two cascaded CNNs and is depicted in Fig. 1. Each CNN follows the design of the 3D U-Net [17], but we propose several novelties in this work that enable the CNN to capture more context effectively. We refer to the architecture of these CNNs as the contextualized U-Net (CU-Net), and the details are explained later in this section.

A. Cascading two CNNs

The first CU-Net reads an input scan at a down-sampled resolution to segment coarse lobes and lobe borders. The resolution of these coarse outputs is subsequently upsampled to a higher resolution. Then the high-resolution input scan and the output of the first CU-Net are concatenated and cropped into 3D patches to be fed to the second CU-Net. The purpose of this concatenation is to introduce extra context captured at the first CU-Net as the second CU-Net uses patch-based inputs at a higher resolution to segment lobes and lobe borders. The cascading of two U-Nets is trained end-to-end, allowing both local details in 3D patches and scan-level context to be learned in the same optimization process. Furthermore, we use the errors found in the predictions of the first CU-Net to sample 3D patches for training the second stage, which encourages

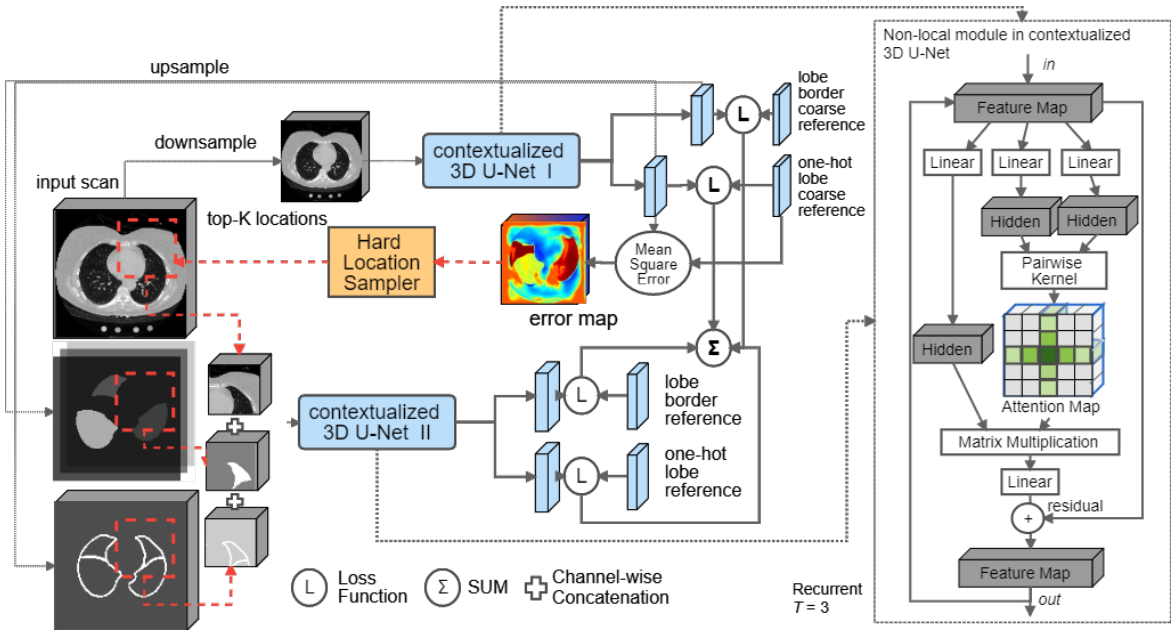


Fig. 1: The overview of our proposed framework with two cascaded Contextualized U-Nets. The output from the first stage is concatenated with the cropped 3D patches as the second stage input.

the second CU-Net to focus on the regions where the first CU-Net fails. This technique can be seen as a form of online hard example mining [33].

B. Contextualized U-Net

The contextualized U-Net architecture (CU-Net) is a 3D U-Net architecture [17] with less number of convolution filters and an additional non-local module. The CU-Net has three down-sampling layers in the encoding path, and each layer consists of two convolutions and a max-pooling operation. Following the down-sampling path, two more convolutions are used to double the number of convolution filters. We then place the non-local module before up-sampling. In the up-sampling path, three layers are used to reconstruct the resolution, and each contains one tri-linear interpolation, followed by two convolutions to reduce the interpolation artifacts. In the end, features are reshaped via a single 1x1x1 convolution in two parallel output branches, and each corresponds to a different learning objective; one produces 6-channel softmax probabilities for segmenting the background and the five lobes. The other provides a single channel probability map by sigmoid function for predicting the lobe border. Features from 3x3x3 convolutions are batch normalized and activated via a rectifier linear unit (ReLU). No dropout is used.

The first CU-Net uses padded convolutions, whereas the second uses valid convolutions. The details regarding CU-Net network architecture on both stages are provided in Table II, where the names of the down-sampling layers are prefixed with 'Down', and the name of up-sampling layers are prefixed with 'Up'. The numbers listed are based on the execution order.

C. The Non-Local module

The original non-local neural network [28] for semantic segmentation computes the feature response at a position as

TABLE II: Architectures for the first and the second stage of the contextualized U-Nets. The convolution filters are named by the kernel sizes K and number of filters N as $K \times K \times K, N$ (stride 1 for all). Non-local linear embedding parameters are defined in Eqs. (3) and (5). $\|$ denotes the operation performed in dual paths.

Layer	CU-Net I	CU-Net II
Down1	3x3x3,1-16 3x3x3,16-24	3x3x3,8-24 3x3x3,24-48
	2x2x2 max pool, stride 2	
Down2	3x3x3,24-24 3x3x3,24-48	3x3x3,48-48 3x3x3,48-96
	2x2x2 max pool, stride 2	
Down3	3x3x3,48-64 3x3x3,64-128	3x3x3,96-96 3x3x3,96-192
	2x2x2 max pool, stride 2	
Bridge	3x3x3,128-128 3x3x3,128-256	3x3x3,192-192 3x3x3,192-384
Non-local	W_θ,$W_\phi \in \mathbb{R}^{256\times32}$ W_ω,$W_\rho \in \mathbb{R}^{3\times32}$ $W_r \in \mathbb{R}^{32\times256}$	W_θ,$W_\phi \in \mathbb{R}^{384\times32}$ W_ω,$W_\rho \in \mathbb{R}^{3\times32}$ $W_r \in \mathbb{R}^{32\times384}$
Up1	3x3x3,384-128 3x3x3,128-128	3x3x3,576-192 3x3x3,192-192
	trilinear interpolation x2	
Up2	3x3x3,176-48 3x3x3,48-48	3x3x3,288-96 3x3x3,96-96
	trilinear interpolation x2	
Up3	3x3x3,72-24 3x3x3,24-24	3x3x3,144-48 3x3x3,48-48
	trilinear interpolation x2	
Output	1x1x1,6 $\|$ 1x1x1,1	1x1x1,6 $\|$ 1x1x1,1
MAC	5.71 G	8.79 G
#Parameter	3.85M	9.24 M

a weighted sum of the features at all locations in the input feature maps as

$$y_i = \frac{1}{\zeta(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{1}$$

where y_i at location i is computed as a weighted sum using the feature correspondence between the feature x_i at the location i and all features indexed by j in the input feature map x. The feature correspondence between feature x_i and x_j is also called the self-attention in this context, computed by the pairwise function f, which is used to weigh the feature embedding $g(x_j)$ before normalizing by $\zeta(x)$. For simplicity, g is set to a linear projection: $g(x_j) = W_g x_j$, and the pairwise function f can be the embedded Gaussian function using linear embeddings defined as $f(x_i, x_j) = e^{(W_\theta x_i)^T (W_\phi x_j)}$. We set the normalizing factor as $\zeta(x) = \sum_{\forall j} f(x_i, x_j)$. Then y becomes the softmax computation along the dimension j written, in matrix multiplication form, as $y = softmax(x^T W_\theta^T W_\phi x) g(x)$. To make the input and output of the non-local module the same size, the y_i is reshaped to have the same dimensions as the input x_i by applying the linear reconstruction function $r, r(y_i) = W_r y_i$. Therefore, the non-local response at location j can be written as $z_i = W_r y_i + x_i$.

The feature response z_i automatically achieves a global receptive field with respect to the input. The computed self-attention map $f(x_i, x_j)$ captures the contextual dependencies, as relevant features would have higher attention responses.

However, the original non-local module disregards the spatial locations of features which may provide critical information for the lobe segmentation. Hence, we propose to compute non-local responses with a geometric term considering the relative locations between features. Here, we denote μ_i, μ_j as geometric features for the position i and j. μ_i is extracted by computing the center of the receptive field of the feature at position i with respect to the input image and normalized by the size of the input image. Noting that if the feature map is produced from a cropped input, the center of receptive field is then shifted according to the offset to the original input. The normalized geometric features are then shifted by 0.5 to have zero mean. $\tau(\mu_i, \mu_j)$ is the pairwise function for measuring correlations. Then, the non-local response with geometric terms is defined as:

$$y_i = \frac{1}{\zeta(x, \mu)} \sum_{\forall j} (f(x_i, x_j) + \tau(\mu_i, \mu_j)) g(x_j), \tag{2}$$

A similar reparameterization can be applied using the softmax function row-wise under linear projections to reformulate Equation 2 into matrix multiplications:

$$y = softmax(x^T W_\theta^T W_\phi x + max(0, \mu^T W_\omega^T W_\rho \mu)) g(x), \tag{3}$$

where $f(x_i, x_j)$ is parameterized as a dot product in a subspace projected using the linear transformation matrix W_θ and W_ϕ. Similarly, W_ω and W_ρ are linear transformations that are used to project the geometric features μ into a subspace where their correspondence is measured by the pairwise kernel function τ, $\tau(\mu_i, \mu_j) = max(0, \mu^T W_\omega^T W_\rho \mu)$. Such correspondence is then trimmed at 0, to restrict geometric relations within a certain threshold.

The Equation 3 however, has high computational cost because the self-attention map requires computing $x^T W_\theta^T W_\phi x$ and $\mu^T W_\omega^T W_\rho \mu$ on all pairs of locations. Each term has a complexity in time and space of $O(C \times W^2 \times H^2 \times D^2)$ where C is the dimension of linear projected subspace and W, H, D denotes the width, height, and depth of a 3D feature map. To reduce computational complexity, we adopt the criss-cross trick [30], which has a time and space complexity of $O((C \times W \times H \times D) \times (H + W + D - 2))$. In CCNet the Equation 2 is modified to:

$$y_i = \frac{1}{\zeta(x, \mu)} \sum_{j \in \Omega j} (f(x_i, x_j) + \tau(\mu_i, \mu_j)) g(x_j), \tag{4}$$

where Ωj indicates the neighboring voxels with respect to j under criss-cross connectivity. Such sparse connectivity requires having three recurrent criss-cross modules to cover all spatial locations in computation.

Given the input feature x_i, the non-local response z_i^t for a feature location i at each t-th recurrent criss-cross module can be written as follows:

$$\begin{aligned} z_i^t &= \begin{cases} x_i & \text{if } t = 0 \\ W_r y_i^{t-1} + z_i^{t-1} & \text{if } t = 1, 2, \ldots, T \end{cases} \\ y_i^t &= \tfrac{1}{\zeta(z^t, \mu)} \textstyle\sum_{j \in \Omega j} (f(z_i^t, z_j^t) + \tau(\mu_i, \mu_j)) g(z_j^t) \end{aligned} \tag{5}$$

At each recurrent step, the non-local response z_i^t is used as the input feature for computing the non-local response for the next recurrent step. For the size of scans used in this work, full global context can be achieved with three recurrent steps for a 3D input feature map. Therefore, we set $T = 3$.

D. Online Hard Example Mining

As shown in Fig. 1 using the red dashed lines, we compute the mean square errors (MSE) between the lobe-wise softmax probabilities of the first CU-Net and the lobe reference standard. We then go through all sliding window 3D patches, and find K patches with the highest integral of MSE and use them to train the second CU-Net.

K is set to 1.0 such that all patches are used to train at the beginning and continuously reduced until it reaches a coverage of only approximately 20% of the scan volume at the end of the training process. The proposed online hard example mining does not introduce extra forward and backward passes on the network, therefore the additional computational cost is trivial.

E. Learning Objectives

There are two learning objectives for each CU-Net: lobe segmentation and lobe border segmentation. Therefore, the final loss function is a summation of four and each is the generalized Dice loss [34]. The lobe border reference is pre-computed from the lobe reference by detecting object boundaries.

Let r be the segmentation reference with n-th voxel values r_{ln} for the class label l and $\hat{r}_{ln}$ be the predicted probabilistic

map for the label l over n-th image voxel, then the generalized Dice loss is defined as:

$$GLD = 1 - 2\frac{\sum_l w_l \sum_n r_{ln}\hat{r}_{ln}}{\sum_l w_l \sum_n r_{ln} + \hat{r}_{ln}},$$

with $w_l = 1/(\sum_n^{N_l} r_{ln})^2$, where N_l the in total number of voxels for the class label l in the segmentation reference. w_l is to re-balance learning against the variance in object volumes.

IV. Experiments

As the COVID-19 pandemic emerged only recently, it was not possible to obtain a large amount of CT scans with annotations of COVID-19 patients. Therefore, we used a transfer learning approach in our experiments. For training of the models on the COVID-19 data, the models were initialized with the trained weights from our models developed on the COPD data set.

A. Training details

Training, validation, and testing of each experiment were carried out on a machine with a NVidia TitanX GPU with 12 GB memory. The methods were implemented using Python 3.6, Pytorch 1.1.0 library [35]. The trainable parameters of each method were initialized using Kaiming He initialization when training from scratch [36] and are optimized using stochastic gradient descent with a momentum of 0.9, and the initial learning rate is set to 10e-6. The initial models were trained using CT scans from the COPD data set. Therefore, the visual patterns in COVID-19 scans are not familiar to these models. For efficiently training on new visual patterns, all models were retrained using a combined loss between the generalized Dice loss (as we used to train initial models) and top-K cross-entropy loss where K is set to 30% of all voxels in the input. The top-K cross-entropy loss was implemented simply as the voxel-wise cross-entropy loss but selecting only K voxels with largest cross entropy to back-propagate.

B. Comparison with previous work

We compared our approach with two state-of-the-art baselines.

1) 3D U-Net: We implemented 3D U-Net following the original paper [17]. The input is a mini-batch of two $132 \times 132 \times 132$ 3D patches randomly cropped from the pre-processed scan (refer to IV-D). As a result of using valid convolutions, the output of this network is $44 \times 44 \times 44$ voxels. During test time, the softmax probabilities of all 3D patches are tiled together by sliding over the entire scan in the output without overlaps to build up a scan-level probability map.

2) FRV-Net: We compare the proposed method with an existing end-to-end lobe segmentation method called FRV-Net [14] which follows the design of the V-Net [18] and extensively uses the idea of deep supervision at almost all scales in the up-sampling pathways. Similar to our approach, FRV-Net introduced multi-tasking into the training process by using both lobe segmentation and prediction lobe borders simultaneously. We implemented the FRV-Net architecture following the paper at our best efforts. Note that this work uses specific pre-processing and post-processing steps where the input scan is resized into a fixed size of $128 \times 256 \times 256$ and intensities are clipped into the range $[-1000 \sim 400]$ HU.

C. Ablation studies

To assess the contribution of the different novel components that are introduced in CTSU-Net, we performed several ablation studies. During these experiments, the models are retrained from scratch using the COPD training set and performance is measured on the COPD test set of 1000 cases. The performance of our proposed model is assessed without the geometric features in the non-local module, and without the non-local module in the two 3D U-Net CNNs, thereby reducing the architecture to a two-stage cascaded 3D U-Net.

D. Pre-processing and post-processing

All training and test scans were standardized by clamping intensity values to the $[-1200 \sim 400]$ range before re-scaling into $[0 \sim 1]$. Then all scans were down-sampled to have a 256×256 in-plane resolution while z-spacing is adjusted to make the scan isotropic.

The input size of the second CNN for our proposed method consists of two $116 \times 116 \times 116$ sized 3D patches. The pre-processed scan is down-sampled by a factor of 2 as the input for the first stage. The predictions of all 3D patches at the second stage are tiled together by sliding over the entire scan to produce a scan-level probability map, which is used to generate the final prediction.

As a post-processing step, the predictions were then up-sampled by nearest neighbor interpolation to match the original resolution of the scans. All evaluations are performed by using predictions and reference segmentations at the original resolution.

E. Evaluation Metrics

The Intersection over Union (IOU), also known as Jaccard index, and 95% percentile in Hausdorff distance (HD95) between predictions and segmentation references were used for quantitative evaluation of segmentation performance. The IOU between two binary masks X, Y is defined as:

$$IOU(X, Y) = \frac{|X \cap Y|}{|X \cup Y|},$$

The Hausdorff distance (HD) measures the surface distance between two binary masks. Denote two surfaces as S_X,S_Y from the masks X, Y, and coordinate indices on the surface as x,y. HD is expressed as:

$$HD(X, Y) = \max\{\max_{x \in S_X} \min_{y \in S_Y} d(x, y), \max_{y \in S_Y} \min_{x \in S_X} d(y, x)\}$$

with $d(\cdot)$ being the Euclidean distance. We use the 95% percentile of the HD distance, and refer to this as HD95.

The overall performance of the method was evaluated by computing the average of the per-lobe metrics. A Wilcoxon

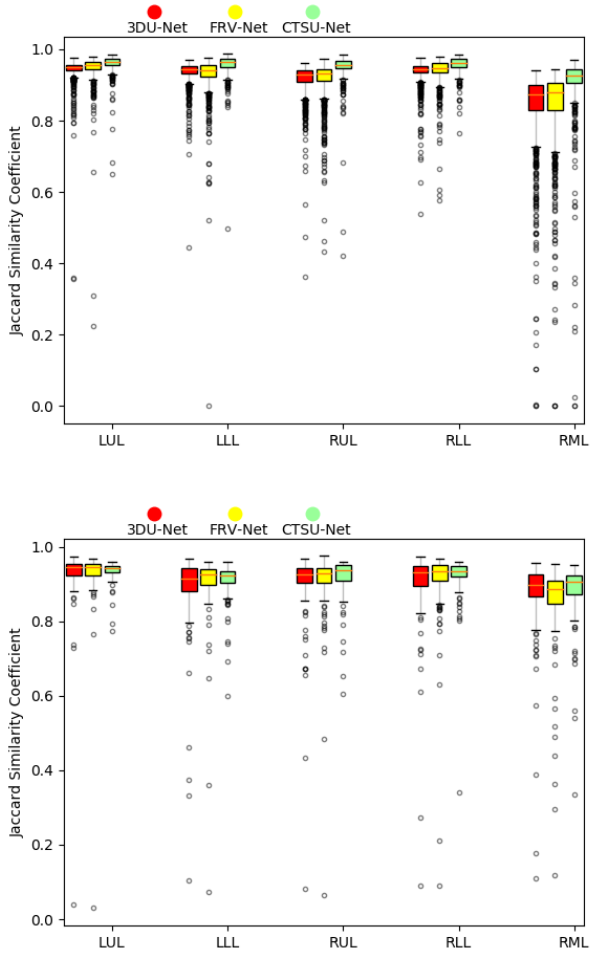


Fig. 2: Box and whisker plots of IOU per-lobe for different methods on the COPD data set (top) and the COVID-19 data set (bottom).

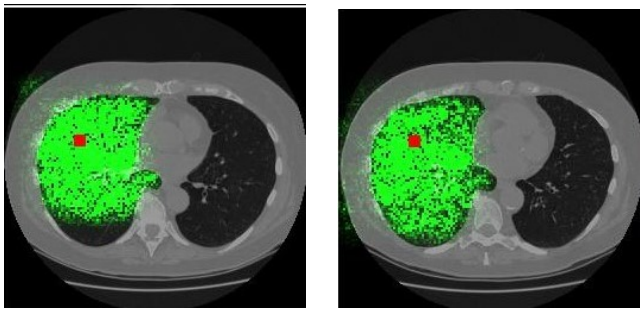
signed-rank test was employed to assess whether the performance difference was statistically significant (p <0.01 with Bonferroni correction).

In addition, we computed the number of Multi-Adds operations (MAC) and the number of parameters to assess computational efficiency. We also provide a comparison with independent human readers for a subset of the COPD data.

V. Results

A. Quantitative results

Table III reports the quantitative results on both data sets. The quantitative results show that our proposed method significantly outperformed the previously published methods in segmenting all lobes both when training from scratch on the COPD data set and retraining on the COVID-19 data set (p<0.01 with Bonferroni correction). We see that using the two cascaded networks boosts the IOU performance from 0.915 to 0.940 for the COPD set and from 0.899 to 0.912 for COVID-19 data. The non-local module helps to further improve the IOU to 0.949 and substantially reduces the HD95 from 8.2 to 2.8mm on the COPD data.



Before Non-Local After Non-Local

Fig. 3: Effective Receptive Field (ERF) before the non-local module (left) and after (right). The green area indicates non-zero gradients (with respect to the input scan) of a feature at a location in the input scan corresponding to the red square.

Box and whisker plots are provided in Fig. 2. These plots show that for both the COVID-19 and the COPD cases, the right middle lobe is the most difficult to segment, which is not surprising given that it can have incomplete fissures at both its boundaries with the upper and lower right lung lobes. Especially for the COPD data, CTSU-Net clearly outperforms the two baselines for the right middle lobe. For the COVID-19 cases, the differences are less pronounced, but it can be observed that there are less outliers with low IOU, indicating CTSU-Net exhibits more robust performance.

In terms of computational efficiency, the proposed method consumes even less memory than the baseline approach, with only a slight increase in the Multi-Adds operations (MAC). Hence, we conclude that the proposed method outperforms the other methods without introducing substantial computational overhead. The proposed method processes a single scan at test time in 20 seconds on average.

B. Ablation study

Table IV shows the results of the ablation study, in which two of our novel components, the non-local module and the geometric features, are removed from CTSU-Net. These models were trained on the COPD training set and reported metrics are the performance on the COPD test set. The results demonstrate the added value of the non-local module and show that the introduction of the geometric features increases the performance over the non-local module alone. This effect is most pronounced for the 95% Hausdorff Distance metric.

C. Effect of the Non-Local module

To measure the effective receptive field (ERF) size before and after the non-local operation, we computed the gradients $\frac{\partial F}{\partial I}\big|_i$ of the feature at the location i in the feature map F with respect to the input image I. For simplicity, only the feature maps at the first CU-Net were studied. The ERF of the features at the same corresponding location before and after non-local operation are visualized in Fig. 3 for one axial slice. The figure renders non-zero gradients in green and indicates the center of the ERF with a red square. The center is a mapped coordinate from the chosen feature in the feature map to the

TABLE III: Quantitative results on the COPD and COVID-19 test sets. IOU and HD95 (in mm) metrics are given in mean ± standard deviation. Boldface denotes the best result for each column.

(a) COPD results

Method	MAC	#Param	Metric	Overall	LUL	LLL	RUL	RLL	RML
3DU-Net [17]	10.5G	16.32M	IOU	0.915±0.037	0.944±0.033	0.937±0.007	0.918±0.043	0.937±0.032	0.840±0.032
			HD95	7.020±7.361	3.801±6.173	5.848±13.87	7.783±9.227	5.374±11.73	12.34±13.47
FRV-Net [14]	7.2G	15.5M	IOU	0.918±0.038	0.950±0.038	0.932±0.050	0.917±0.050	0.942±0.033	0.848±0.103
			HD95	8.672±9.239	4.586±12.73	5.847±9.048	10.44±20.02	5.570±13.65	16.90±21.51
CTSU-Net (ours)	14.5G	13.1M	IOU	**0.949±0.026**	**0.962±0.020**	**0.959±0.023**	**0.952±0.030**	**0.960±0.010**	**0.912±0.080**
			HD95	**2.772±3.342**	**2.016±4.955**	**2.016±3.347**	**3.102±6.388**	**2.489±6.729**	**4.234±5.981**

(b) COVID-19 results

Method	MAC	#Param	Metric	Overall	LUL	LLL	RUL	RLL	RML
3DU-Net [17]	10.5G	16.3M	IOU	0.899±0.060	0.934±0.038	0.886±0.104	0.902±0.081	0.905±0.091	0.868±0.109
			HD95	**6.338±8.894**	4.454±8.894	6.099±7.166	7.614±12.29	6.073±12.29	7.452±10.73
FRV-Net [14]	9.3G	15.5M	IOU	0.902±0.053	**0.936±0.029**	0.906±0.073	0.908±0.065	0.910±0.089	0.849±0.119
			HD95	6.586±6.883	**4.173±6.527**	**4.785±7.820**	9.064±12.49	5.876±12.53	9.031±11.61
CTSU-Net (ours)	13.3G	14.1M	IOU	**0.912±0.044**	0.935±0.028	**0.908±0.054**	**0.919±0.058**	**0.919±0.067**	**0.877±0.091**
			HD95	6.447±9.052	8.363±24.09	6.635±14.43	**6.351±11.25**	**3.762±8.019**	**7.127±9.723**

RUL, RML, RLL, LUL, LLL: Right upper, Right middle, Right lower, Left upper, Left lower lobes. Overall: per-lobe mean.

TABLE IV: Ablation study on the COPD data set for the non-local module (Non-local) and the geometric features (Geometric) into the two-stage cascading framework.

Method	Two-stage	Non-local	Geo-metric	HD95	IOU
only two-stage	✓			8.127±13.312	0.940±0.031
w/o geometric	✓	✓		5.523±10.72	0.942±0.031
CTSU-Net	✓	✓	✓	**2.772±3.342**	**0.949±0.026**

input image via up-sampling, thus a slight shift may occur. The left image shows the ERF before the non-local operation is contained in a square due to the nature of stacked convolutions. However, the ERF after non-local on the right side shows a non-square distribution, reaching the other side of the lung. We therefore conclude that the non-local module can enlarge effective receptive field dramatically and in theory can achieve a global receptive field.

D. Qualitative Results

Fig. 4 shows the qualitative performance of the proposed method and includes the 3D U-Net baseline, the proposed method without the non-local module, and the reference segmentations for comparison. We selected three COPD and three COVID-19 cases with various levels of pathological and anatomical variations. We observe that all methods usually do not produce oversegmentation of the lungs. By adding contextual dependencies using the non-local module with the appearance and geometric terms, we see that the proposed method generates generally smoother lobe borders and is able to infer about lobe shapes from the global context.

E. Comparison with human readers

To evaluate human performance, we asked two independent human readers to manually segment the lobes from scratch given a segmentation of the lung. Their results are evaluated on a random set of 100 scans from the COPD test set. The human readers achieved 0.953 ± 0.017 IOU and 1.790 ± 1.13 HD95 on average, while the proposed method achieved 0.955 ± 0.015 IOU and 2.322 ± 1.982 HD95. The human readers are both significantly better than the other methods, whereas there is no significant difference between the readers ($p<0.001$).

VI. Discussion and Conclusion

We have presented a novel method using contextualized two-stage neural networks for segmenting pulmonary lobes in CT images. The proposed method is capable of capturing global context by introducing a non-local module. This proposed non-local module takes geometric features into account when computing the self-attention. We show in our results that introducing global context improves the lobe segmentation performance significantly on the COPD and the COVID-19 data set. The Hausdorff distance metric in the ablation study shows that using geometric features is effective for generating finer object boundaries. This smoothing effect can also be observed from the qualitative results, where the lobe boundaries from the proposed method are more consistent with the reference lobe shapes. Our proposed architecture is a general framework and can be applied or extended for other medical segmentation problems. In terms of computational efficiency, our method has the same level of Multi-Adds operations (MAC). It requires even fewer trainable parameters compared to the standard 3D U-Net. Our method can be trained and tested on a consumer-level GPU with 12 GB memory, and inference time is around 30 seconds for a full resolution CT scan.

Segmentation of lobes in scans of patients with a severe pneumonia due to COVID-19 is not an easy task. In this work, we used only 104 COVID-19 CT scans for training. Thanks to pre-training on 4000 COPD scans, we still obtained good results with such a small training set and we were able to provide lobe segmentations robust to the presence of ground-glass, consolidations and crazy paving. Lobe segmentation is an important prerequisite for accurate quantification of lung damage in COVID-19 CT scans. Fig. 4 shows that the

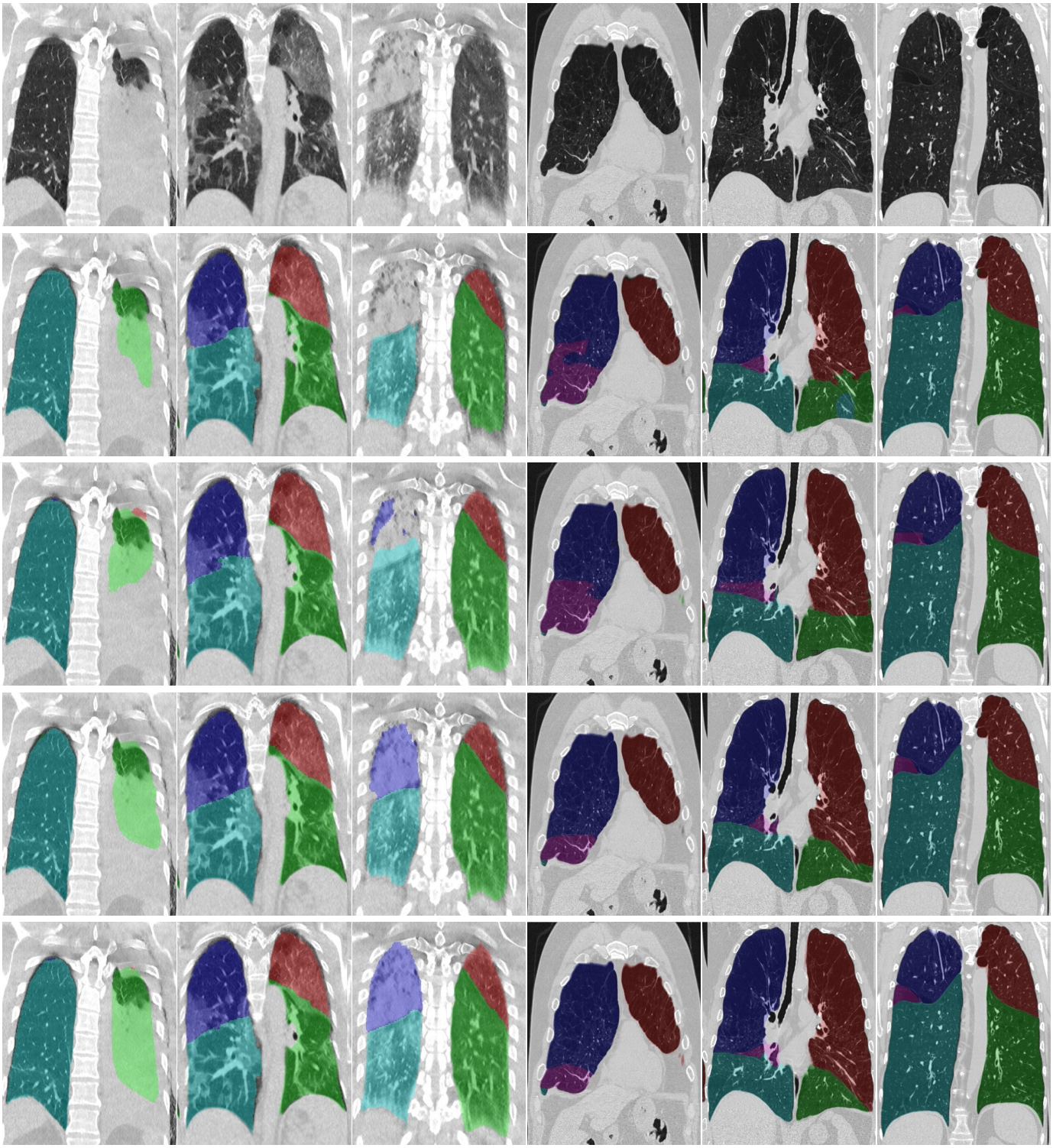


Fig. 4: Qualitative comparison of segmentation results for six representative test cases. The left three columns show COVID-19 cases, the right three columns show COPD cases. From top to bottom: input image, 3DU-Net baseline, CTSU-Net method w/o the non-local module, the proposed CTSU-Net, and the segmentation reference. ■ right upper, ■ right middle, ■ right lower, ■ left upper, ■ left lower lobes.

standard 3D U-Net may miss areas of consolidation (third row), but even CTSU-Net misses part of a lobe when this lobe is completely filled with pleural fluid (first column). We hypothesize that a larger training set would further improve performance on cases with gross pathological changes that are not yet well represented in the 104 scans we had available for training at the time of this research. Nevertheless, the results presented here are sufficient for further analysis and

we believe that they will prove useful in automated per-lobe severity scoring. This is a topic for future research.

We freely share our segmentation algorithm with the research community on `https://grand-challenge.org/algorithms/`.

References

[1] M. Grabenhorst, B. Schmidt, U. Liebers, and J.-W. Oestmann, "Radiologic manifestations of bronchoscopic lung volume reduction in severe chronic obstructive pulmonary disease," *AJR Am J Roentgenol*, vol. 204, no. 3, pp. 475–486, 2015.

[2] M. Kubo, N. Niki, S. Nakagawa, K. Eguchi, M. Kaneko, N. Moriyama, H. Omatsu, R. Kakinuma, and N. Yamaguchi, "Extraction algorithm of pulmonary fissures from thin-section CT images based on linear feature detector method," *IEEE Trans Nucl Sci*, vol. 46, no. 6, pp. 2128–2133, 1999.

[3] M. Kubo, N. Niki, K. Eguchi, M. Kaneko, M. Kusumoto, N. Moriyama, H. Omatsu, R. Kakinuma, H. Nishiyama, K. Mori *et al.*, "Extraction of pulmonary fissures from thin-section CT images using calculation of surface-curvatures and morphology filters," in *ICIP*, vol. 2, 2000, pp. 637–640.

[4] J. Wang, M. Betke, and J. P. Ko, "Pulmonary fissure segmentation on CT," *Med Image Anal*, vol. 10, no. 4, pp. 530–547, 2006.

[5] J. Pu, J. K. Leader, B. Zheng, F. Knollmann, C. Fuhrman, F. C. Sciurba, and D. Gur, "A computational geometry approach to automated pulmonary fissure segmentation in CT examinations," *IEEE Trans Med Imaging*, vol. 28, no. 5, pp. 710–719, 2008.

[6] E. M. van Rikxoort, B. van Ginneken, M. Klik, and M. Prokop, "Supervised enhancement filters: application to fissure detection in chest CT scans," *IEEE Trans Med Imaging*, vol. 27, no. 1, pp. 1–10, 2007.

[7] S. E. Gerard, T. J. Patton, G. E. Christensen, J. E. Bayouth, and J. M. Reinhardt, "Fissurenet: A deep learning approach for pulmonary fissure detection in CT images," *IEEE Trans Med Imaging*, vol. 38, no. 1, pp. 156–166, 2018.

[8] B. Raasch, E. Carsky, E. Lane, J. O'callaghan, and E. Heitzman, "Radiographic anatomy of the interlobar fissures: a study of 100 specimens," *AJR Am J Roentgenol*, vol. 138, no. 6, pp. 1043–1049, 1982.

[9] J.-M. Kuhnigk, V. Dicken, S. Zidowitz, L. Bornemann, B. Kuemmerlen, S. Krass, H.-O. Peitgen, S. Yuval, H.-H. Jend, W. S. Rau *et al.*, "New tools for computer assistance in thoracic CT. part 1. functional analysis of lungs, lung lobes, and bronchopulmonary segments," *Radiographics*, vol. 25, no. 2, pp. 525–536, 2005.

[10] E. M. Van Rikxoort, M. Prokop, B. de Hoop, M. A. Viergever, J. P. Pluim, and B. van Ginneken, "Automatic segmentation of pulmonary lobes robust against incomplete fissures," *IEEE Trans Med Imaging*, vol. 29, no. 6, pp. 1286–1296, 2010.

[11] B. Lassen, E. M. van Rikxoort, M. Schmidt, S. Kerkstra, B. van Ginneken, and J.-M. Kuhnigk, "Automatic segmentation of the pulmonary lobes from chest CT scans based on fissures, vessels, and bronchi," *IEEE Trans Med Imaging*, vol. 32, no. 2, pp. 210–222, 2012.

[12] F. J. Bragman, J. R. McClelland, J. Jacob, J. R. Hurst, and D. J. Hawkes, "Pulmonary lobe segmentation with probabilistic segmentation of the fissures and a groupwise fissure prior," *IEEE Trans Med Imaging*, vol. 36, no. 8, pp. 1650–1663, 2017.

[13] S. E. Gerard and J. M. Reinhardt, "Pulmonary lobe segmentation using a sequence of convolutional neural networks for marginal learning," in *ISBI*, 2019, pp. 1207–1211.

[14] F. T. Ferreira, P. Sousa, A. Galdran, M. R. Sousa, and A. Campilho, "End-to-end supervised lung lobe segmentation," in *IJCNN*, 2018, pp. 1–8.

[15] D. Terzopoulos, X. Ding, and N. Tajbakhsh, "Automatic, fast, reliable lung lobe segmentation using a 3D progressive dense v-network."

[16] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *NIPS*, 2016, pp. 4898–4906.

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *MICCAI*, 2016, pp. 424–432.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*. IEEE, 2016, pp. 565–571.

[19] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

[20] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, 2018, pp. 5410–5418.

[21] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *CVPR*, 2017, pp. 472–480.

[22] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu *et al.*, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv:1809.04430*, 2018.

[23] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *Reconstruction, segmentation, and analysis of medical images*, 2016, pp. 83–94.

[24] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *ICML*, 2013, pp. 1310–1318.

[25] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med Imag Anal*, vol. 36, pp. 61–78, 2017.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[29] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019, pp. 3146–3154.

[30] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," *arXiv:1811.11721*, 2018.

[31] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo, "Genetic epidemiology of COPD (COPDGene) study design," *COPD*, vol. 7, no. 1, pp. 32–43, 2011.

[32] E. M. van Rikxoort, M. Prokop, B. de Hoop, M. A. Viergever, J. P. Pluim, and B. van Ginneken, "Automatic segmentation of the pulmonary lobes from fissures, airways, and lung borders: evaluation of robustness against missing data," in *MICCAI*, 2009, pp. 263–271.

[33] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016, pp. 761–769.

[34] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in *DLMIA*, 2017, pp. 240–248.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.