Automatic, Fast, Reliable Lung Lobe Segmentation Using a 3D Progressive Dense V-Network

Abdullah-Al-Zubaer Imran1,2,∗, Ali Hatamizadeh1,2,∗, Shilpa P. Ananth2, Demetri Terzopoulos1,2, Xiaowei Ding1,2,†, Nima Tajbakhsh2†

1University of California, Los Angeles, Los Angeles, CA 90095
2VoxelCloud Inc, Los Angeles, CA 90024

Abstract

Reliable and automatic segmentation of lung lobes is important both clinically and technically for diagnosis, assessment, and quantification of pulmonary diseases. Current lung lobe segmentation techniques are prohibitively slow and undesirably reliant on prior airway/vessel segmentation and/or user interactions for optimal results. This work presents a reliable, fast, and fully automated lung lobe segmentation based on a progressive Dense V-network architecture. The proposed method can segment lung lobes in one forward pass of the network with an average runtime of 2 seconds, eliminating the need for any prior lung segmentation or any subsequent user intervention. For a dataset of 354 chest CT scans from LIDC data, we achieved a Dice score of 0.9497 ± 0.0176, significantly outperforming (p < 0.001) a 2D U-Net model with an average Dice score of 0.9201 ± 0.0431 and average run-time of 11 seconds. Our statistical analyses also demonstrated that our model is robust against the choice of reconstruction kernel, size of the reconstruction interval, and choice of CT scan vendors. We further measured the segmentation accuracy of our model in the presence of lung diseases by selecting 200 pathological cases from the LTRC database and 55 cases from the LOLA11 challenge, obtaining an average Dice score of 0.9580 and 0.9345, respectively—a performance level competitive to the state-of-the-art. Our results demonstrate that our model can reliably segment healthy and pathological lung lobes under varying configurations of CT scan reconstruction.

1 Introduction

Human lungs are divided into five lobes. The right lung has three lobes, namely, right upper lobe (RUL), right middle lobe (RML), and right lower lobe (RLL), which are separated by a minor and a major fissure, whereas the left lung has two lobes, namely, left upper lobe (LUL) and left lower lobe (LLL), separated by a major fissure. Fig. 1 shows the five lobes separated by major and minor fissures in a coronal CT slice. Each of the five lobes is functionally independent as they have separate bronchial and vascular systems.

Automatic segmentation of the lung lobes is important for both clinical and technical purposes. In clinical practice, doctors very often base their assessment of a disease severity and the corresponding treatment plan on the affected lung lobe. As such, upon encountering a disease or lesion in the lung, radiologists may navigate through the nearby slices to identify the affected lobe, especially when the fissure lines are not clearly visible in the target slice. An automatic lobe segmentation model can

∗Authors with equal contributions
†Corresponding authors: X. Ding: xding@voxelcloud.io, N. Tajbakhsh: ntajbakhsh@voxelcloud.io

Submitted to Medical Imaging with Deep Learning (MIDL) 2018, Amsterdam, Netherlands.
therefore shorten the CT reading session by continually informing the radiologists about their location in the lung anatomy. From the technical perspective, accurate lung lobe segmentation can improve several subsequent clinical tasks, including nodule malignancy prediction (cancers mostly occur in the left or right upper lobes), automatic lobe-aware report generation for each nodule, and assessment and quantification of chronic obstructive pulmonary diseases (COPD) and interstitial lung diseases (ILD), by narrowing down the search space to the lung lobes most-likely to be affected. However, identifying fissures poses a challenge for both human and machine perception. First, fissures are most often incomplete, not extending to the lobar boundaries. Several studies in the literature have confirmed the incompleteness of fissures as a very common phenomenon. After reviewing 100 fixed and inflated lung specimens, Raasch et al. [19] found incomplete right major fissures in 70% of the cases, left major fissures in 46% of the cases, and 94% across the minor fissures. Moreover, the studies of Gulsun et al. [12] and Aziz et al. [3] also showed more than 50% incompleteness in pulmonary fissures. Second, the visual characteristics of lobar boundaries can change in the presence of pathologies. Such morphological changes could also be related to the varying thicknesses, locations, and shapes of the fissures. Third, there also exist other fissures in the lungs that can be misinterpreted as the major and minor fissures that separate the lobes. Examples include accessory fissures and azygos fissures.

To address the need for accurate and robust lung lobe segmentation, we have proposed a fully automatic and reliable deep learning solution. The proposed technique generates accurate segmentation of the lung lobes in about 2 seconds in only a single forward pass of the network, eliminating the need for any user interactions or any prior segmentation of lungs, vessels, or airways, which are common assumptions in the design of existing models.

2 Related Work

Various automatic and semi-automatic approaches have been proposed for lung lobe segmentation. Depending on whether they rely on prior segmentation or anatomical information, these efforts can be grouped in two main categories: “reliant” and “non-reliant”.

Reliant approaches: Bragman et al. [5] proposed a method that relied on prior segmentation of airways and vessels, in which a population model of fissure priors was integrated to the patient-specific anatomical information for non-parametric surface fitting. Bauer et al. [4] employed prior lobe segmentation masks acquired in the inspiration phase for lung lobe segmentation in the expiration.

Figure 1: A coronal lung CT slice with visible fissures. Three lobes in the right lung, namely, right upper lobe (RUL), right middle lobe (RML), and right lower lobe (RLL) are defined by a major fissure (indicated by red arrows) and a minor fissure (directed by yellow arrows). Two lobes in the left lung, namely, left upper lobe (LUL) and left lower lobe (LLL) are defined by a major fissure (indicated by red arrows).
phase. Some efforts have integrated prior segmentation with anatomical information for final mask generation. Giuliani et al. [11] incorporated anatomical information with the segmentation of airways, vessels, and lungs in a multi-level graph cut algorithm, but their work is highly reliant on the quality of the prior airway and vessel segmentations as well as anatomical knowledge. In an effort to decrease such dependency, Doel et al. [7] used an initialization via fissure detection for segmentation. Moreover, Lassen and van Rikxoort [15] proposed a watershed-based lobe segmentation method that combined anatomical information and demonstrated improvement in case of incomplete fissures, but the overall segmentation performance was limited due to the failure of individual prior segmentations. Reliant segmentation can also be achieved by using a previously segmented atlas, but the final performance is highly dependent on the atlas generation algorithm. Ross et al. [21] employed thin-plate splines and maximum a posteriori estimation using a manually defined atlas as a reference. Pu et al. [18] performed lobe segmentation by fitting an implicit function to fissures without reliance on prior airway or vessel segmentation. In addition, van Rikxoort et al. [22] made use of multiple atlases for lobe segmentation. The shortcomings of such atlas-based methods include slow execution time, cumbersome process of generating the atlas, and relatively lower performance for pathological cases.

Non-reliant approaches: The deep convolutional neural network model has recently been deployed for lung lobe segmentation. George et al. [8] proposed a 2D fully convolutional neural network to segment lobar boundaries and then used a 3D random walker algorithm to segment the lobes. Despite the promising results, the method relied on the random walker algorithm whose optimal parameters could change from one dataset to another. It is most desirable to have an end-to-end solution that does not rely on any subsequent heuristic method.

In the present work, we mitigate the aforementioned limitations, namely reliance on prior masks, slow runtime, and lack of robustness by an end-to-end, single-pass, deep-learning-based framework that does not rely on any prior airway/vessel segmentation, anatomical knowledge, or atlases.

3 Method

3.1 3D Progressive Dense V-Net

By leveraging a multi-path supervision scheme, we propose a progressive convolutional architecture as an end-to-end solution for automatic lung lobe segmentation in CT images. Inspired by dense V-networks [10], the proposed method employs various building blocks, such as convolutional layers and dense feature blocks as well as convolutional downsampling and upsampling in the architecture, as illustrated in Fig. 2. In each convolutional layer, 3D convolutional operations are followed by batch normalization and rectified linear units (ReLU) while, in each dense feature block, every layer is connected with every other layer in a feed-forward manner, resulting in each layer receiving the

![Figure 2: The architecture of the progressive dense V-network.](image)
feature maps of all the preceding layers as input [13]. In addition, batch-wise spatial dropout [10] is incorporated for regularization purposes. The inputs to the network are first down-sampled using a strided convolutional layer, and then passed to the dense feature blocks. Consecutively, the outputs of the dense feature blocks are utilized in low and high resolution passes via convolutional down-sampling and skip connections. This enables the generation of feature maps at three different resolutions. The outputs of the skip connections of the second and third dense feature blocks are further up-sampled in order to be consistent with the size of the output in the first skip connection. These feature maps are concatenated and passed to a convolutional layer followed by a softmax layer, which outputs the probability maps. At each stage, we define three separate dice loss layers, as discussed in the subsequent section, with the aim of progressively improving the previous outputs.

All the convolutional down-sampling layers have a kernel size of $3 \times 3 \times 3$ with strides of 2, except for the initial convolutional down-sampling layer, which has a kernel size of $5 \times 5 \times 5$. In addition, each of the employed dense feature stacks has 2, 4, and 8 output channels for high, medium, and low resolution stacks, respectively, with a kernel size of $3 \times 3 \times 3$ and a stride of 1. All the Skip pathways have a kernel size of $3 \times 3 \times 3$ and a stride of 1.

### 3.2 Loss Layer

Volumetric predictions along with the corresponding ground truth are fed into each of the loss functions. Such predictions denote the probabilities by which each voxel belongs to the corresponding lobe. Since the background region often occupies the majority of the volume, we utilize the dice loss function [17], which directly maximizes the similarity between the predicted values and the ground truth over all the voxels. We employ this concept for a multi-class segmentation task:

$$D = \sum_{l=1}^{L} \frac{\sum_{i=1}^{N} p_{l}^i g_{l}^i}{\sum_{i=1}^{N} (p_{l}^i)^2 + \sum_{i=1}^{N} (g_{l}^i)^2},$$

where $N$ is the total number of voxels, $L$ is the number of classes, $p_{l}^i$ denotes the predicted probability for each class, and $g_{l}^i$ is the corresponding 1-hot encoded ground truth.

### 4 Experiments

#### 4.1 Datasets

We selected a subset of chest CT volumes (354 cases) from the publicly available LIDC dataset [2] for annotation. To ensure variability in the data, CT scans were selected such that both challenging and visible fissures are well-represented in the dataset. The ground truth masks were generated in a semi-automatic fashion. To mitigate bias in the ground truth, mask generation was performed by multiple observers, and the generated masks were later refined and validated by an expert radiologist.

The dataset was split into 270 training and 84 test cases. The CT scans range within 100–672 slices, each $512 \times 512$ pixels. The voxel dimensions are in the range of [0.49–0.98, 0.49–0.98, 0.45–3.00] mm in x-, y- and z-axes. Along with the 84 test cases in LIDC, we also validated our model against 200 cases from Lung Tissue Research Consortium (LTRC) database and 55 cases from the Lobe and Lung Analysis (LOLA11) challenge [16]. The LTRC dataset includes pathological cases with clear evidence of COPD or ILD diseases, including emphysema, fibrosis, etc.

#### 4.2 Base-Line Comparison: 2D U-Net

In order to construct a baseline model for comparison, we implemented a 2D U-net [20]. The implemented architecture is symmetric and consists of four contracting and expanding layers, starting with 16 features in the first layer and doubling the number of features in each step. Each contracting layer consists of two $3 \times 3$ convolutions and a ReLu activation followed by a $2 \times 2$ max-pooling layer. The expansion path consists of an up-convolution with feature concatenation from the respective contracting layer, and two $3 \times 3$ convolutions. In addition, all the ReLu layers are preceded by a batch-normalization layer. To improve the training process, we also use a generalized dice score as the loss function such that the contribution of each class in the image to the gradients is balanced.
Table 1: Performance comparison of 2D U-net and 3D progressive dense V-net models in segmenting 84 cases from LIDC and 200 cases from LTRC. Mean Dice score and standard deviation for each lobe have been shown.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>RUL</th>
<th>RML</th>
<th>RLL</th>
<th>LUL</th>
<th>LLL</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIDC(84)</td>
<td>2D U-net</td>
<td>0.9077±0.0492</td>
<td>0.8443±0.0756</td>
<td>0.9403±0.0541</td>
<td>0.9588±0.0415</td>
<td>0.9496±0.0560</td>
<td>0.9201±0.0431</td>
</tr>
<tr>
<td></td>
<td>3D PD V-net</td>
<td>0.9494±0.0259</td>
<td>0.9036±0.0439</td>
<td>0.9625±0.0178</td>
<td>0.9710±0.0131</td>
<td>0.9619±0.0349</td>
<td>0.9497±0.0176</td>
</tr>
<tr>
<td>LTRC(200)</td>
<td>2D U-net</td>
<td>0.9139±0.0388</td>
<td>0.8662±0.0539</td>
<td>0.9523±0.0234</td>
<td>0.9606±0.0225</td>
<td>0.9543±0.0209</td>
<td>0.9295±0.0250</td>
</tr>
<tr>
<td></td>
<td>3D PD V-net</td>
<td>0.9583±0.0102</td>
<td>0.9234±0.0198</td>
<td>0.9667±0.0096</td>
<td>0.9729±0.0065</td>
<td>0.9681±0.0079</td>
<td>0.9579±0.0061</td>
</tr>
</tbody>
</table>

Figure 3: Distribution of Dice scores for 2D U-net and 3D progressive dense V-net: overall and lobe-wise. The black boxes indicate the score distributions for the 2D model and the blue boxes indicate them for the 3D model. The red + symbols indicate outliers in a distribution.

4.3 Implementation Details

We implemented the proposed 3D progressive dense V-net using TensorFlow [1] and NiftyNet [9]. The training volumes were first normalized, followed by rescaling to $512 \times 512 \times 128$. Due to the large memory footprint of the model, we used a distributed training scheme wherein the model was split amongst 3 GPUs and the gradient check-pointing method [6] was used for memory-efficient back-propagation. We used the Adam optimizer [14] with a learning rate of 0.01 and a weight decay of $10^{-7}$.

For the 2-D U-net implementation, we trained the network with axial slices from all the training volumes, each sized $512 \times 512$ and normalized to have values between 0 and 1. To avoid over-fitting to the background class, we used only the axial slices, wherein at least one lung lobe is present. The network was implemented in Tensorflow and trained using the Adam optimizer with a learning rate of $5 \times 10^{-5}$ and batches of 10 images.

4.4 Results and Discussion

LIDC Results: We evaluated the proposed 3D progressive dense V-net and the baseline 2D U-net models by computing Dice scores. Table 1 shows the calculated overall and lobe-wise Dice scores for...
both models. The 3D model, with an overall score of $0.9497 \pm 0.0176$, outperformed the 2D model, with an overall score of $0.9201 \pm 0.0431$. As is evident in Table 1 and Fig. 3, the 3D progressive dense V-net yields consistently larger dice score for each of the lung lobes. Moreover, the lower standard deviation for each lobe indicates that the 3D model is more robust. The 2-sample t-test also confirmed that the 3D progressive dense V-net model achieves a significantly larger overall Dice score than the 2D U-net model.

The accuracy of the lung volume measurement is additional evidence that our 3D model (progressive dense V-net) outperforms the 2D U-net model in segmenting lung lobes. The 3D model generates a mean volume difference of $0.023 \pm 0.019$ whereas the 2D model exhibits a mean volume difference of $0.049 \pm 0.060$. A visual comparison can be seen in Fig. 4, where the original CT slices, slice-wise lobe segmentation results, and ground truth segmentations are visualized for both the 2D and 3D models. The patterns of the inherent fissures in the lungs seem to have been learned better by our progressive dense V-net model than by the baseline model. Fig. 5 shows 3D visualizations of a segmented CT scan from LIDC dataset.

**Robustness Analysis:** To investigate the robustness of our model, we further analyzed the Dice scores of the 84 cases by grouping them in three ways. For the first grouping, the scores were put in three different Z-spacing buckets: $Z$-spacing $\leq 1$, $1 < Z$-spacing $< 2$, and $Z$-spacing $\geq 2$. In the second grouping, the scores were put in four manufacturer buckets: GE, Philips, Siemens, and Toshiba. In the third grouping, the scores were grouped according to the reconstruction kernel into 2 buckets: smooth and sharp. Fig. 6 shows the Dice scores distributions for each grouping and its underlying buckets. The two-sample t-test confirmed that there were no significant differences between the average dice scores of the buckets within each grouping, suggesting that our model is robust against the choice of reconstruction kernel, the size of reconstruction interval, and the different CT scan vendors.
LTRC Results: The evaluation of the LTRC segmentation was conducted on 200 LTRC cases with the 3D progressive dense V-net resulting in an average Dice score of $0.958 \pm 0.006$. Once again, the 3D model outperformed the 2D U-net model with an average Dice score of $0.9295 \pm 0.025$. Individual lobes were segmented better in the 3D model than in the 2D model (Table 1). Note that the LTRC dataset includes many pathological cases where the fissure lines are either invisible, distorted, or absent in presence of pathologies such as emphysema, fibrosis, etc. As a result, lobe segmentation becomes more challenging. Nevertheless, our model performed well in segmenting lobes in pathological cases from the LTRC dataset. Moreover, our model outperformed the model of George et al. [8] in segmenting from LTRC cases both in Dice score ($0.9407 \pm 0.2548$) and inference speed (4-8 minutes per case).

LOLA11 Results: The segmentation results on the LOLA11 cases submitted online were evaluated as overlap (Jaccard) scores. To be consistent with our previous analyses, we converted the Jaccard scores to Dice scores. The results are shown in Table 2 and segmentation of some of the cases have been visualized (see A). Our method achieved an overall Dice score of 0.9345, which is competitive with the state-of-the-art method [5], while outperforming the methods of Giuliani et al. [11] and van Rikxoort et al. [22].

5 Conclusions

Automatic and reliable lung lobe segmentation is a challenging task in the presence of chest pathologies and the absence of visible, complete fissures. In this paper, we introduced a new 3D segmentation approach, namely, progressive dense V-networks for the automatic, fast, and reliable, segmentation of lung lobes from chest CT scans, without any prior airway/vessel segmentation. We evaluated
Table 2: Performance evaluation of 3D progressive dense V-net models on 55 LOLA cases: showing lobe-wise mean Dice scores, standard deviations, median scores, first quartiles, and third quartiles

<table>
<thead>
<tr>
<th>Lobe</th>
<th>Mean ± SD</th>
<th>Q₁</th>
<th>Median</th>
<th>Q₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUL</td>
<td>0.9518 ± 0.1750</td>
<td>0.9371</td>
<td>0.9688</td>
<td>0.9881</td>
</tr>
<tr>
<td>RML</td>
<td>0.8621 ± 0.4149</td>
<td>0.8107</td>
<td>0.9284</td>
<td>0.9663</td>
</tr>
<tr>
<td>RLL</td>
<td>0.9581 ± 0.1993</td>
<td>0.9621</td>
<td>0.9829</td>
<td>0.9881</td>
</tr>
<tr>
<td>LUL</td>
<td>0.9551 ± 0.2160</td>
<td>0.9644</td>
<td>0.9834</td>
<td>0.9924</td>
</tr>
<tr>
<td>LLL</td>
<td>0.9342 ± 0.3733</td>
<td>0.9546</td>
<td>0.9805</td>
<td>0.9902</td>
</tr>
<tr>
<td>Overall</td>
<td>0.9345</td>
<td>0.9282</td>
<td>0.9384</td>
<td>0.9195</td>
</tr>
</tbody>
</table>

* Jaccard score to Dice score conversion: Dice = 2 × Jaccard / (1 + Jaccard)

Figure 7: Sagittal plane visualization of LOLA11 segmentation by 3D progressive dense V-net. Upper row (from left): Case 8-slice (left lung), Case 8-segmentation, Case 6-slice (right lung), and Case 6-segmentation. Lower row (from left): Case 21-slice (left lung), Case 21-segmentation, Case 55-slice (right lung), and Case 55-segmentation.

our method using 3 test datasets: 84 cases from the public LIDC database, 200 cases from the public LTRC database, and 55 cases from the LOLA11 challenge. Our results demonstrated that the suggested model outperforms, or at worst performs comparably to, the state-of-the-art while running at an average speed of 2 seconds per case. Our analyses further demonstrated the robustness of the suggested method against varying reconstruction intervals and lung pathologies.

Acknowledgments

The authors are grateful to Gerard Nguyen, MD, Radiology Resident at the Washington University School of Medicine, for his extensive support and assistance in annotating lobe masks in the LIDC cases.
A Visualization of LOLA11 Results

Visualization of LOLA11 cases (Fig. 7) shows the left lung of case 8 and the right lung of case 6, where our method performed well with lobe-wise Dice scores greater than 0.99. For the left lung in case 8, the LUL and LLL Dice scores were 0.9940 and 0.9926, respectively. For the right lung in case 6, the scores are as follows: RUL: 0.9580, RML: 0.9480, and RLL: 0.9869. Again, the left lung of case 21 and right lung of case 55 are shown in Fig. 7, where the segmentation Dice scores were relatively low. For the left lung in case 21, the LUL score was 0.8170 and LLL score was 0.3035. For the right lung in case 55, although the right lower lobe was segmented with a high Dice score of 0.9818, because of the invisibility of the horizontal fissure, the RUL and RML had low segmentation Dice scores of 0.6827 and 0.7499, respectively.

References


