NaviAirway: a Bronchiole-sensitive Deep Learning-based Airway Segmentation Pipeline

Andong Wang, Terence Chi Chun Tam, Ho Ming Poon, Kun-Chang Yu, and Wei-Ning Lee

***Abstract*—Airway segmentation is essential for chest CT image analysis. However, it remains a challenging task because of the intrinsic complex tree-like structure and imbalanced sizes of airway branches. Current deep learning-based methods focus on model structure design while the potential of training strategy and loss function have not been fully explored. Therefore, we present a simple yet effective airway segmentation pipeline, denoted *NaviAirway*, which finds finer bronchioles with a bronchiole-sensitive loss function and a human-vision-inspired iterative training strategy. Experimental results show that NaviAirway outperforms existing methods, particularly in identification of higher generation bronchioles and robustness to new CT scans. Besides, NaviAirway is general. It can be combined with different backbone models and significantly improve their performance. Moreover, we propose two new metrics (Branch Detected and Tree-length Detected) for a more comprehensive and fairer evaluation of deep learning-based airway segmentation approaches. NaviAirway can generate airway roadmap for Navigation Bronchoscopy and can also be applied to other scenarios when segmenting fine and long tubular structures in biomedical images. The code is publicly available on https://github.com/AntonotnaWang/NaviAirway.**

***Index Terms*—Airway segmentation, Computed Tomography (CT), semi-supervised learning, deep learning training strategy**

I. Introduction

COMPUTED Tomography (CT) is a predominant medical imaging modality for the assessment of lung diseases, such as lung cancer and chronic obstructive pulmonary disease (COPD). Airway segmentation plays a vital role in the CT image analysis procedure. For example, Navigation Bronchoscopy (NB) is the safest and superior for accessing peripheral pulmonary lesions [1]. For better procedural efficiency and patient care, NB requires a pre-planned 3D airway roadmap segmented and reconstructed from CT images which navigates the bronchoscope down into the bronchioles for target nodule sampling [2]–[4]. In the case of COPD, airway segmentation from CT images is the key to accurate measurement of airway lumen size and wall thickness [5].

However, accurate and automated airway segmentation from CT images remains to be challenging primarily because of morphological complexity of airways and imbalanced sizes between airways and background as well as between low- and high-generations airways in CT images. **1)** The first challenge comes from the most prominent characteristic of airways – the complex tree structure, which begins from the trachea and ends at the alveoli. The trachea beginning at the larynx is denoted as Generation 0, while the divided left and right main-stem bronchi as Generation 1. After that, the airways become finer until the 23rd generation—alveolar sacs [6]. In our paper, *low generation* stands for large airways closer to the trachea, while *high generation* refers to fine bronchioles closer to the alveolar sacs. **2)** The second challenge comes from the fact that the airways only take up a small fraction in CT images. It is difficult even for human experts to exclude all background information and target the airways accurately. Moreover, unlike humans, who can recognize airway branches as geometric shapes, existing computer algorithms can only identify them pixel-wisely and thus often overlook fine bronchioles of few-pixel thickness.

Over the years, many airway segmentation methods have been developed. There are two main categories — traditional methods which rely on manually selected features [7]–[20] and deep learning-based methods which combine Convolutional Neural Networks (CNNs) with traditional methods or focus on new model architecture design [21]–[37]. While the proposed model architectures become increasingly sophisticated [32]–[37], the potential of training strategy and loss function have not been fully explored. We found that with special design of training strategy and loss function considering the morphological characteristics of airways, the segmentation pipeline trained on a simple backbone model could finds more finer bronchioles than existing methods.

Therefore, we present our method, coined as NaviAirway, which consists of a simple backbone model plus an airway-specific loss function and a human-vision-inspired training strategy. Based on 3D U-Net [21], the backbone model only adds dilated convolution and self-attention to extract richer airway features from larger areas. The novelty of NaviAirway lies on the training framework. During model training, the airway-specific loss function pushes the model to recognize more bronchioles of higher generations. At the same time, the human-vision-inspired iterative training strategy guides the model to specifically extract features of both low and high generation airways and preserve those learned features. The

This project was in part supported by COVID-19 Action Seed Funding of Faculty of Engineering, The University of Hong Kong. Wei-Ning Lee is the corresponding author.

Andong Wang is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China (e-mail:wangad@connect.hku.hk).

Terence Chi Chun Tam is with Respiratory Division, Department of Medicine, The University of Hong Kong, Hong Kong, China, and also with Queen Mary Hospital, Hong Kong, China (e-mail:tcctam@netvigator.com).

Ho Ming Poon is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China (e-mail:hmpoon6@connect.hku.hk).

Kun-Chang Yu is with Broncus Medical, Inc., San Jose, CA, 95134 USA (e-mail:jyu@broncus.com).

Wei-Ning Lee is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China, and also with the Biomedical Engineering Programme, The University of Hong Kong, Hong Kong, China (e-mail:wnlee@eee.hku.hk).

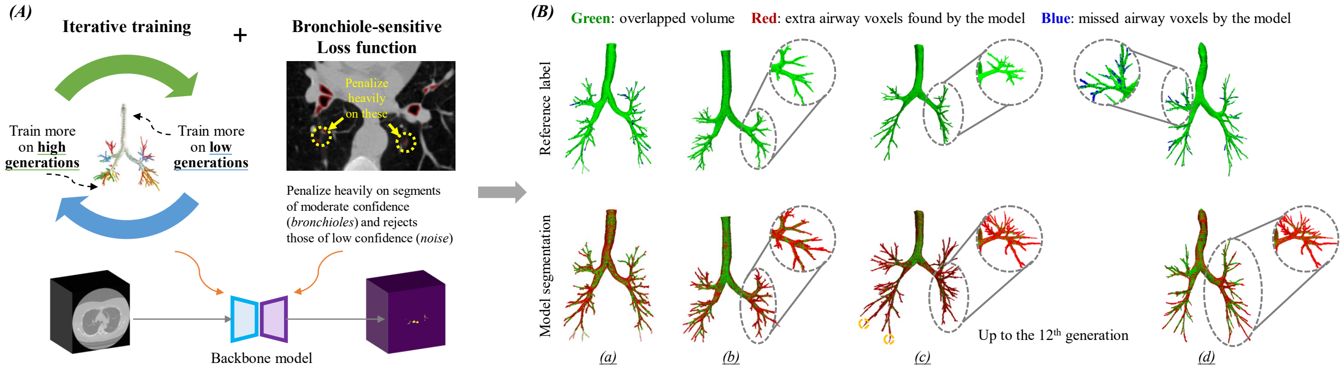
(A)
Iterative training
+
Bronchiole-sensitive
Loss function
Penalize heavily on these
Train more on high generations
Train more on low generations
Penalize heavily on segments of moderate confidence (bronchioles) and rejects those of low confidence (noise)
Backbone model
(B)
Green: overlapped volume
Red: extra airway voxels found by the model
Blue: missed airway voxels by the model
Reference label
Model segmentation
Up to the 12th generation
(a)
(b)
(c)
(d)

Fig. 1. **(A)** NaviAirway detects more finer bronchioles by the proposed iterative training strategy and bronchiole-sensitive loss function. **(B)** Exemplary test results on EXACT'09-TR-LIDC-IDRI showing that NaviAirway finds more finer bronchioles that reference labels (see Table III for the accuracy values of the four cases). (a) is a randomly chosen case. (b) and (c) are good cases where the model detected almost all the airways in the reference model. (d) is a case with relativity inferior performance and more undetected branches (the blue voxels) than others; nevertheless, the model only missed few airway branches and detected extra airways (the red voxels) overlooked by the reference.

proposed loss function and training strategy are general and can improve the performance of different model architectures. Furthermore, they can be combined with semi-supervised learning approach (i.e., teacher-student training built based on Noisy Student [38]) to further increase model accuracy and robustness.

To evaluate model performance, we first test NaviAirway on two public datasets. The comparison results show that our method is more accurate and detects longer airway trees than existing methods. Moreover, the computational cost of NaviAirway is deemed acceptable for navigation bronchoscopy as our method on average takes five minutes to process the CT scans of one patient. We subsequently test NaviAirway on a private dataset, demonstrating its robustness to a previously unexposed dataset.

In addition to the method itself, evaluation metrics also play a vital role in the development process. Many studies followed the metrics defined in the EXACT'09 Challenge [9]. However, the metrics in [9] are essentially evaluation standards for traditional methods and may not be suitable for deep learning-based methods. It is because, in the airway segmentation task, the airway labels are not 100% correct and may miss some or even a substantial number of bronchioles. In particular, we only have *"reference"* labels, instead of *"ground truth"* labels (which can only be obtained when the bronchoscopy procedure is eventually performed). Hence, in cases of airway segmentation, the goal of a deep learning model is not to provide airway segments that perfectly match the reference; instead, the deep learning model is to learn from the reference and apply the knowledge of airway features back to the raw CT scans to find all recognizable bronchioles. It is expected that the deep learning model can do even better than the reference. Therefore, we additionally standardize existing metrics and propose new metrics to have a fairer and more comprehensive evaluation of deep learning-based airway segmentation methods.

Our contributions are summarized as follows:

- A simple yet effective airway segmentation pipeline which finds more finer bronchioles by exploring the potential of training strategy and loss function.
- A new loss function which drives the model to recognize finer bronchioles (fine and long tubular shapes).
- A new human-vision-inspired iterative training strategy that guides the model to learn both the features of fine and coarse airways, while preventing knowledge loss of airway features.
- Two new metrics (Branch Detected and Tree-length Detected) for a fairer and more complete evaluation of deep learning-based airway segmentation methods.

II. Related Works

A. Traditional Methods

Traditional methods mainly include **1)** region growing and thresholding, **2)** morphologic and geometric model-based methods, and **3)** hybrid approaches combining the above two methods [7], [8]. For example, EXACT'09 Challenge [9] presented 15 airway tree extraction algorithms submitted to the competition. Ten out of the 15 methods used region growing and thresholding techniques which utilize brightness of different tissues. Similar techniques were also developed, including pixel value filtering and thresholding [10], thresholding and rectangular region mask [11], GVF snakes [12], fuzzy connectivity [13], [14], and two-pass region growing [15]. Alternatively, airways were mathematically defined according to their morphologic and geometric features of airway for extraction [16], [17]. Hybrid methods combined the strengths of the two to provide better segmentation [18]–[20].

B. Deep Learning-based Methods

Compared with traditional methods, deep learning-based models, on average, detect twice longer airways [22], [23], [26]–[28], [31], [34], [39]–[41]. Abundant studies combined Convolutional Neural Networks (CNNs) and traditional methods based on the idea that CNN provided preliminary results, and the traditional method was responsible for refinement. One mainstream of studies used 3D U-Net [21] as the backbone model and built different post-processing approaches, which included fuzzy connectedness region growing and skeletonization guided leakage removal [22], image boundary post-processing to minimize the boundary effect in airway reconstruction [23], and freeze-and-grow propagation [24]. Besides, other works focused on designing the backbone network. A simple and low-memory 3D U-Net was proposed in [25]. Graph Neural Networks (GNNs) were adopted to segment airways [26], [27]. In [28], a GNN module was incorporated into a 3D U-Net, while [29] developed a graph refinement-based airway extraction method by combining GNN and mean-field networks. Beyond utilizing existing deep learning models which were built for general tasks, in more recent studies, new network architectures (usually based on 3D U-Net) considering special features of airways were designed to achieve a higher segmentation accuracy. They included patch classification by 2.5 CNN [30], AirwayNet [31], Airway-SE [32], 2D plus 3D CNN [33], attention distillation modules plus feature recalibration modules [34], [35], group supervision plus union loss function [36], and attention on weak feature regions [37].

III. Method

Consider a labeled set $\mathcal{D} : \{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a 3D CT image, $\boldsymbol{a}_i$ is the corresponding airway annotation map (which denotes each voxel on $\boldsymbol{x}_i$ as either background or airway), and N is the number of labeled images. Also, we have an unlabeled set $\mathcal{U} : \{\boldsymbol{x}_i\}_{i=1}^{M}$, where M is the number of unlabeled images and $M \gg N$.

Our pipeline consists of a backbone model which is simply adapted from 3D U-Net (Section III-A) , a new bronchiole-sensitive loss function (Section III-B), and a new human-vision-inspired iterative training strategy (Section III-C). We also introduce our simple post-processing (Section III-D) and how the proposed loss function and training strategy combine with semi-supervised learning to increase model robustness (Section III-E).

A. Backbone Model: Feature Extraction from a Larger Area

We built our model (denoted as Φ_θ) based on the structure of 3D U-Net. We simply adapted the down-sampling and up-sampling operations by introducing a new feature extraction module which consists of one dilated convolution, one self-attention block, and two typical convolutional kernels (Figure 1(a)). Compared with the conventional convolution kernels, the proposed feature extraction module helps to extract features from a larger surrounding area to avoid the interference from other tubular shapes, such as the esophagus and the vessels. Please see Appendix A for more explanation of our model adaptation. It is noted that the backbone model architecture can be designed with flexibility. We present that the proposed loss function and training strategy can also improve segmentation accuracy with other backbone models (See Table VI).

B. Loss Function: Detecting More Finer Bronchioles

Based on dice loss, a new bronchiole-sensitive loss function L has been formulated to let the model recognize more bronchioles of higher generations, which are of small diameters but essential for navigational bronchoscopy. The proposed model outputs two prediction maps of airway and background, respectively. Therefore, L consists of two components (Equ. (1)) to deal with airway and background voxels differently (Equ. (2) and (3)):

$$L = L_{aw} + L_{bg}, \tag{1}$$

$$L_{aw} = 1 - \frac{2\sum_{k=1}^{K} {p_k}^2 a_k w_k}{\sum_{k=1}^{K} {p_k}^4 + \sum_{k=1}^{K} {a_k}^2}, \tag{2}$$

$$L_{bg} = 1 - \frac{2\sum_{k=1}^{K} p_k a_k w_k}{\sum_{k=1}^{K} {p_k}^2 + \sum_{k=1}^{K} {a_k}^2}, \tag{3}$$

where subscripts aw and bg denote airway and background, respectively; k denotes voxel position; K is the total number of voxels; $p_k \in [0, 1]$ is the model confidence of a voxel being background or airway, while $a_k \in 0, 1$ is the computer-assisted manual annotation of the k-th voxel, where the voxel being background is assigned as 0 and the voxel being the airway is given a 1. w_k is a weight value. (In this paper, we follow a simple strategy to assign the weights: We assign same weights for all voxels in L_{aw} and set the weights to be 1 in L_{bg}. Therefore, Equ. (1) becomes $L = w_{aw} L_{aw} + L_{bg}$. w_{aw} is the ratio of the number of annotated airway voxels to the number of annotated background voxels.)

Different from conventionally used dice loss function, L_{aw} replaces model prediction confidence p_k with pseudo-confidence ${p_k}^2$. The reason of using pseudo-confidence is analogous to the scheme where athletes bear extra weights during training and release those weights in competitions. When p_k is 0.5, the pseudo-confidence is only 0.25. Hence, the model must learn harder to push p_k to be larger than 0.7, which means the pseudo-confidence of being airway can be larger than 0.5. However, during model inference, we only used p_k to decide airway segments. With the above training approach, our model could recognize more fine bronchioles than the one with dice loss function solely.

The derivative of L (Equ. (4) and (5)) helps elaborate the mechanism of the proposed loss function. For p_k on the airway prediction map, $\frac{\partial\ L}{\partial\ p_k} = \frac{\partial\ L_{aw}}{\partial\ p_k}$, while on the background prediction map, $\frac{\partial\ L}{\partial\ p_k} = \frac{\partial\ L_{bg}}{\partial\ p_k}$. $\frac{\partial\ L_{bg}}{\partial\ p_k}$ follows a linear pattern, whereas $\frac{\partial\ L_{aw}}{\partial\ p_k}$ indicates that L_{aw} penalizes heavily on airway segments of moderate confidence (which could be bronchioles) and rejects those of too low confidence (which tend to be noises). Please see Section VI-2 for more discussion.

$$\frac{\partial L_{aw}}{\partial p_k} = \frac{-4p_k a_k \left(\sum_{l=1}^{K} {p_l}^4 + \sum_{l=1}^{K} {a_l}^2\right) + 8{p_k}^3 (\sum_{l=1}^{K} {p_l}^2 a_l)}{\left(\sum_{l=1}^{K} {p_l}^4 + \sum_{l=1}^{K} {a_l}^2\right)^2}, \tag{4}$$

$$\frac{\partial L_{bg}}{\partial p_k} = \frac{-2a_k \left(\sum_{l=1}^{K} {p_l}^2 + \sum_{l=1}^{K} {a_l}^2\right) + 4p_k (\sum_{l=1}^{K} p_l a_l)}{\left(\sum_{l=1}^{K} {p_l}^2 + \sum_{l=1}^{K} {a_l}^2\right)^2}. \tag{5}$$

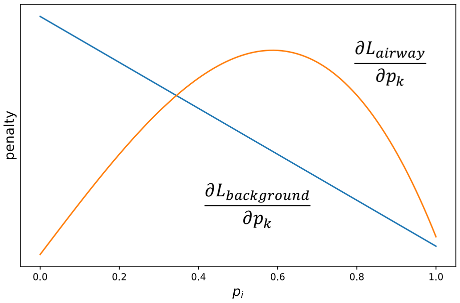

Fig. 2. The change of $\frac{\partial\ L_{aw}}{\partial\ p_k}$ and $\frac{\partial\ L_{bg}}{\partial\ p_k}$ with different p_k values. The curves indicate L_{aw} penalizes heavily on airway segments of moderate confidence and rejects those of too low confidence

C. Iterative Training Strategy: Solution to Two Imbalanced Distributions in Annotation

There are two imbalanced distributions in airway annotation. One is that airway segments only take a small fraction of space compared with the background. The other is that airway segments of high generations have a larger number but take up fewer voxels compared with airways of low generations.

As the images are first cropped to cuboids and the cuboids are fed to the model for training, experiments show that if all cuboids are chosen with the same frequency, the model tends to memorise the airways of low generation while ignoring those of high generation. To address such issue, existing methods following a first-low-generation-then-high-generation training scheme to push the model to detect the fine bronchioles in the second stage. However, this scheme causes another problem that the model loses the knowledge of low generation airways after long training on airways of high generation.

To make the model perform well on both high and low generations, inspired by that humans observe an image in a repetitive zoom-in-and-zoom-out manner, we propose a iterative training strategy. For each iteration, the model is trained on both generations but with different frequency. Specifically, for cuboid pair $(\boldsymbol{x}_i^j,\ \boldsymbol{a}_i^j)$ (i.e., the j-th cuboid pairs of the i-th CT image), we first define t_i^j which is the ratio of the number of outermost voxels of the airway segment to the total number of airway voxels. Larger t_i^j means more airways in $\boldsymbol{a}_i^j$ are fine bronchioles and vice versa. When training more frequently on high generations, the probability of the pair being chosen in a batch is proportional to t_i^j. In contrast, when training more frequently on low generations, the probability value is inversely proportional to t_i^j. When a cuboid $\boldsymbol{x}_i^j$ contains no airway voxel, its probability is a small constant value δ. The conditions for probability values are listed in Equ. (6):

$$p[(\boldsymbol{x}_i^j,\ \boldsymbol{a}_i^j)] = \begin{cases} \beta_h t_i^j & (t_i^j \neq 0,\ \text{more on high generations}) \\ \beta_l \frac{1}{t_i^j} & (t_i^j \neq 0,\ \text{more on low generations}) \\ \beta_0 & (t_i^j = 0) \end{cases}, \tag{6}$$

where β_h, β_l, and β_0 are manually selected parameters (in this paper, $\beta_h = 1$, $\beta_l = 10$, and $\beta_0 = 1$) to control the scale of the probability values.

Additionally, experiments have shown that the proposed iterative training strategy can improve segmentation accuracy while it is not sensitive to different iteration interval and the manually selected parameters.

D. Post Processing

Our post processing is to identify and delete unconnected noise shapes. After model training, an airway confidence map $\Phi_\theta(\boldsymbol{x})$ is thereafter obtained. By setting a threshold th, we obtain an airway mask $\boldsymbol{A} = \Phi_\theta(\boldsymbol{x}) > th$. Next, the common sense that airways are connected leads us to find the largest connected shape in the model output. Suppose $\boldsymbol{A}$ consists of several separated shapes $\{\boldsymbol{S}_h\}$, we find $\boldsymbol{S}_{h^*} = max\,|\boldsymbol{S}_h|$ as the airway segmentation. In addition, there exist some broken airway branches in $\{\boldsymbol{S}_h\}\backslash\boldsymbol{S}_{h^*}$. We connect those broken branches to $\boldsymbol{S}_{h^*}$ if they are close enough. To connect, we first define a search range R. For every end point e of airway branches in $\boldsymbol{S}_{h^*}$, if any shape $\boldsymbol{S}_{\hat{h}}$ in $\{\boldsymbol{S}_h\}\backslash\boldsymbol{S}_{h^*}$ is within $B_R(e)$ (i.e., the neighborhood of e), update $\boldsymbol{S}_{h^*}$ by $\boldsymbol{S}_{h^*} := \boldsymbol{S}_{h^*} + \boldsymbol{S}_{\hat{h}}$, and make $\boldsymbol{S}_{h^*}$ a connected shape by lowering threshold th within $B_R(e)$ to fill the gap. In our experiments, we found that $\boldsymbol{S}_{h^*}$ was already good, so connecting broken branches to it was just for refinement, and R was set to be small.

E. Combination with Semi-supervised Learning

As we have much more unlabeled images $\mathcal{U}$ than the labeled ones $\mathcal{D}$ (i.e., $M \gg N$), the proposed loss function and training strategy can be further combined with semi-supervised learning to increase model accuracy and robustness by learning a more complete picture of the CT image distribution [42]. Therefore, we built our semi-supervised learning approach—teacher-student training—based on Noisy Student [38]. Please see Appendix B for the details.

IV. Experiments

A. Datasets

As shown in Table I, two public datasets (EXACT'09 [9] and LIDC-IDRI [43]) and one private dataset (QMH) were used to evaluate NaviAirway. Each case has multiple 2D lung CT images of one patient for 3D reconstruction, and the size of each 2D slice is 512x512 (pixels). Note that both EXACT'09 and LIDC-IDRI did not provide official airway annotations. Twenty cases (CASE1 to CASE20) from EXACT'09 and 40 cases (chosen by slice thickness and pixel spacing) from LIDC-IDRI were annotated using ITK-SNAP and manually corrected by Shanghai Jiao Tong University [35]. After that, the 60 labeled cases (denoted as **EXACT'09-TR-LIDC-IDRI**) were randomly split into a training set (50 cases) and a testing set (10 cases). Cross-validation was conducted by repeating the splitting multiple times, performing training, testing for each splitting, and taking the average of testing accuracy values. The remaining 20 cases (CASE21 to CASE40) from EXACT'09 were labeled with the same method by us (denoted as **EXACT'09-TE**). For the private dataset QMH, nine cases (denoted as **QMH-L**) of different slice thicknesses and pixel spacings (Table I) were labeled using commercial software named LungPoint with experts' manual correction.

TABLE I
DATASETS USED FOR TRAINING AND EVALUATION.

Name of dataset	Number of cases	Labeled cases	Labeling method	Slice thickness (mm)	Pixel spacing (mm)	Remark
EXACT'09 [9]	40	40	By ITK-SNAP with experts' manual correction [35]	0.6 - 1.25	0.5 - 0.8	Public
LIDC-IDRI [43]	1018	40	By ITK-SNAP with experts' manual correction [35]	<0.625	0.5 - 0.7	Public
QMH	21	9	By LungPoint with experts' manual correction	1.0 - 5.0	0.5 - 0.9	Private

TABLE II
THE PROPOSED METRICS FOR MODEL PERFORMANCE EVALUATION.

Metric category	Metric name	Formula	Evaluation aspect
Overall accuracy	Dice Similarity Coefficient (DSC)	$\frac{2TP}{2TP+FP+FN}$	Similarity between model segmentation and reference segmentation.
	Sensitivity	$\frac{TP}{TP+FN}$	Ratio of the overlapped segmentation to the reference segmentation, indicating how well the model grasps the existing knowledge given by the reference.
	Precision	$\frac{TP}{TP+FP}$	Ratio of the overlapped segmentation to the model segmentation, indicating the percentage of segmented volume inferred by the model
Structural accuracy	Branch Detected (BD)	$\frac{N_{TP}}{N_{ref}} \times 100\%$	The number of branches in the overlapped segmentation over the number of branches in the reference segmentation, indicating how well the model identifies the total number of existing airways in the reference.
	Tree-length Detected (TD)	$\frac{L_{TP}}{L_{ref}} \times 100\%$	The length of airway tree in the overlapped segmentation over the length of airway tree in the reference segmentation, indicating how well the model identifies the total length of existing airways in the reference.
	Branch Ratio (BR)*	$\frac{N_{P}}{N_{ref}} \times 100\%$	The number of branches in the model segmentation over the number of branches in the reference segmentation, indicating how well the model learns the airway features hidden in the reference segmentation and uses these features to find more bronchioles.
	Tree-length Ratio (TR)*	$\frac{L_{P}}{L_{ref}} \times 100\%$	The length of airway tree in the model segmentation over the length of airway tree in the reference segmentation, indicating how well the model learns the airway features hidden in the reference segmentation and uses these features to find more bronchioles.

* Note that the extra airways which do not exist in reference should be examined visually upon deletion of noises and exclusion of the influence from pulmonary vessels.

The training set of EXACT'09-TR-LIDC-IDRI was used for supervised learning, while the remaining 978 unlabeled cases from LIDC-IDRI and 12 unlabeled cases from QMH were used for semi-supervised learning (denoted as **LIDC-IDRI-QMH-U**). After training, the testing set of EXACT'09-TR-LIDC-IDRI was used for internal testing, and QMH-L was used for external testing because the data distribution was unseen by the model.

B. Implementation

1) Data preparation: First, we stacked up the CT slices (in DICOM format) to form 3D image data. Then, thresholding was done to keep the Hounsfield Unit (HU) values within [-1000, 600]. Both the training sets of EXACT'09-TR-LIDC-IDRI and LIDC-IDRI-QMH-U were further cropped into 32x128x128 cuboids owing to available GPU memory.

2) Model development: We used PyTorch [44] to implement our model. Besides the techniques built by ourselves described in the Method section, data augmentation, including random flip, random affine, random blur, random noise, random motion, and random spike [45], was performed to further increase model robustness. We trained the model on an NVIDIA Tesla M60 with 8 GB memory for approximately three days (including semi-supervised learning). Adam optimizer and a learning rate of 10^{-5} were used during training.

C. Metrics

Ideally, the accuracy of model-based airway segmentation is best evaluated through actual bronchoscopy by experts. However, it is labor-intensive, and any additional examination that may prolong the clinical procedure should be avoided. For this pilot study on retrospective data, we adopted five existing metrics and devised two new metrics (Branch Ratio (BR) and Tree-length Ratio (TR)) as listed in Table II for quantitative evaluation. These metrics can be divided into two categories—overall accuracy and structural accuracy. The former quantifies volumetric information voxel-wise, while the latter counts the number of branches and the length of detected tree structure. Three metrics are pertinent to overall accuracy and calculated from TP, FP, and FN, which denote true positive, false positive, and false negative, respectively. TP indicates percentage of model-based airway segmentation that overlaps with the reference airway segmentation; FP represents the rate of airway segments which only exist in model segmentation; FN signifies the proportion of reference airway segments that are not detected by model segmentation. These seven metrics constitute a more complete evaluation of the model.

In structural accuracy, the two new metrics, Branch Ratio (BR) and Tree-length Ratio (TR), are to supplement Branch Detected (BD) and Tree-length Detected (TD) which were proposed by the EXACT'09 challenge and widely used in previous studies [9]. For BD and TD, the underlying assumption that FP segments are all noise is not necessarily valid. For deep learning models, FP segments may be airway branches missed by the reference annotation. Therefore, we introduced BR and TR. When calculating BR and TR, visual inspection was a preliminary step to check whether noises or pulmonary vessels were present in FP segmentation to verify that bronchioles to be included were real but overlooked by the reference.

In this study, we also calculated another widely used metric, False Positive Rate (FPR) but did not include it in Table II as FPR indicated the percentage of leakage volume. We argue that FPR may not be a good indicator for deep learning-based methods because it assumes all extra segments that do not exist in the reference are noise. Another concern is that FPR is defined as FP/(FP+TN), in which FP+TN stands for the background area, namely non-airways, in the reference segmentation. Different sizes of background areas were given in different published works [9], [30], [32], [34], [35], [46], making FPR an unstandardized metric for fair comparison. In our calculation, we only counted the background voxels within the cuboid, which was just fit by the airway reference mask in FP+TN. FPR was thus only used for our reference and found to be less reliable for model performance assessment than the other adopted metrics.

Additionally, we show the results of all the seven metrics of NaviAirway in Table VI in Results. However, when comparing NaviAirway with existing methods, not all the seven metrics were used because some accuracy results of other models were directly cited from literature. We did not reimplemented them because the source codes were not provided while they were trained and tested on the same datasets as ours. Moreover, we believe that they provided their best results, and our reimplementation would perform worse. Therefore, it is a fairer comparison by using their reported accuracy results.

V. Results

A. Performance comparison

First, our pipeline was compared with state-of-the-art airway segmentation methods on the testing sets of both EXACT'09-TR-LIDC-IDRI and EXACT'09-TE. Table IV shows the comparison on the testing set of EXACT'09-TR-LIDC-IDRI. Our method outperformed others in both overall and structural accuracy. Note that we tested our method using two threshold values: 0.5 and 0.7. Setting the threshold value to be 0.5 was to identify more bronchioles of higher generations, some of which were not included in the reference but were real airway segments by experts' visual inspection. Setting the threshold value to be 0.7 was to make a fair comparison in terms of overall accuracy. From Table IV, our method has the highest overall accuracy with its DSC and sensitivity being 94.2%±1.1% and 96.6%±2.3% at the threshold of 0.7. In the aspect of structural accuracy, our method (threshold = 0.5) delineated much longer airway trees (TD: 94.2%±5.7%). Moreover, our method could find more bronchioles of higher generations than the reference (Table IV, BR: 115.0%±19.0% and TR: 115.5%±18.5% when threshold = 0.5). Our method detects finer bronchioles up to the 12th generation, whereas the mean and median values of the detected generation number are 7.9 and 7.5, respectively. Four examplary cases in Figure 1 (B) also indicate that NaviAirway could detect mode finer bronchioles than the reference annotations.

TABLE III
METRICS VALUES OF (A)-(D) SHOWN IN FIGURE 1 (B).

	DSC	Sensitivity	Precision	BD	TD	BR	TR	FPR
(a)	92.0%	98.5%	86.3%	95.7%	95.2%	104.3%	102.5%	0.17%
(b)	92.3%	99.7%	86.0%	98.8%	100.0%	148.0%	133.9%	0.19%
(c)	87.6%	99.9%	77.9%	98.5%	99.7%	175.9%	181.6%	0.29%
(d)	91.6%	98.3%	85.8%	89.4%	95.8%	113.9%	112.0%	0.15%

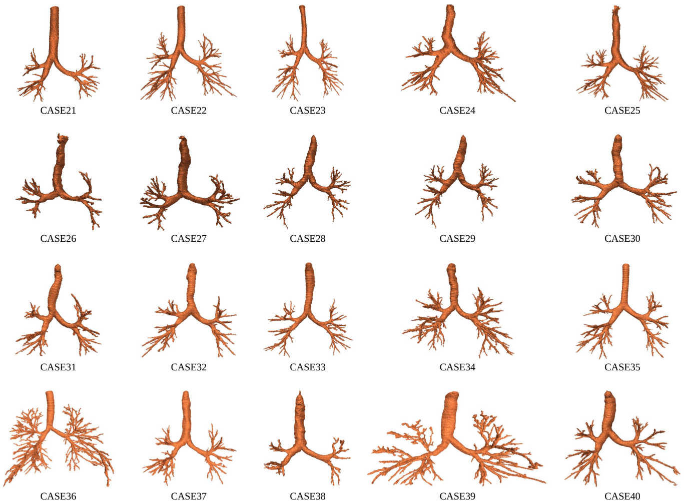

Fig. 3. Exemplary segmentation results obtained by our method on EXACT'09-TE.

Table V summarizes the performance of our (the 1st row), published (rows 2-5), and newly submitted methods in the EXACT'09 challenge (rows 6-10) on the EXACT'09-TE dataset. Results show that our method detects two times more branches and two times longer airway trees than benchmarks (Branch count: 272.9±142.4 and Tree length (cm): 286.6±158.4). The 20 cases of EXACT'09-TE segmented by NaviAirway are illustrated in Figure 3. And the details of each case are summarized in Table VII. Results show that our method could detect more bronchioles (BR: 117.7% and TR: 132.2%) of higher generations, up to the 13th generation on average. For CASE36, our model detects twice more branches and twice longer airway trees, and the detected bronchioles are up to the 18th generation.

B. Ablation study

We tested how each of the components in our model contributed to performance improvement. Table VI shows that the proposed loss function and training strategy significantly improve segmentation performance (around 20% to 30% improvement in terms of TR), while other components, feature extraction module and semi-supervised learning, have lower improvements (around 10% improvement in terms of TR). Data augmentation only slightly improves model performance.

TABLE IV
PERFORMANCE COMPARISON OF AIRWAY SEGMENTATION METHODS ON EXACT'09-TR-LIDC-IDRI.
THE MEAN VALUE AND STANDARD DEVIATION ARE SHOWN FOR EACH METRIC.

	DSC	Sensitivity	BD	TD	FPR
AG U-Net [26]**	82.7%±22.2%	72.5%±28.9%	70.1%±33.3%	63.5%±30.8%	0.014%±0.012%*
Wang et al. [27]**	93.5%±2.2%	88.6%±8.8%	93.4%±8.0%	85.6%±9.9%	0.018%±0.012%*
Juarez et al. [28]**	87.5%±13.2%	77.5%±15.5%	77.5%±20.9%	66.0%±20.4%	0.009%±0.009%*
AirwayNet [31]**	93.7%±1.9%	87.2%±8.9%	91.6%±8.3%	82.1%±10.9%	0.014%±0.009%*
Juarez et al. [23]**	93.6%±2.2%	86.7%±9.1%	91.9%±9.2%	80.7%±11.3%	0.014%±0.009%*
Jin et al. [22]**	93.6%±2.0%	88.1%±8.5%	93.1%±7.9%	84.8%±9.9%	0.017%±0.010%*
Qin et al. [34]**	92.5%±2.0%	93.6%±5.0%	**96.2%±5.8%**	90.7%±6.9%	0.035%±0.014%*
3D U-Net [21]***	92.9%±1.7%	95.8%±2.3%	66.5%±18.8%	72.3%±18.8%	0.048%±0.038%
V-Net [47]***	85.9%±3.4%	81.8%±7.0%	34.2%±9.1%	35.0%±9.8%	0.177%±0.074%
VoxResNet [48]***	85.8%±6.3%	78.3%±9.8%	29.8%±9.9%	33.1%±10.2%	0.044%±0.023%
Ours (th=0.5)****	90.7%±1.8%	**98.4%±1.4%**	88.4%±10.7%	**94.2%±5.7%**	0.224%±0.128%
Ours (th=0.7)****	**94.2%±1.1%**	96.6%±2.3%	83.3%±11.4%	90.4±8.4%	0.117%±0.079%

* The definition FPR of other methods may be different from ours as the exact computation of TN was not clarified.
** Rows 1-7 are state-of-the-art methods with performance values quoted from published articles (tested on same datasets as our method).
*** Rows 8-10 are three well-received medical image processing models reimplemented by us to do airway segmentation using same-frequency sampling strategy and ordinary dice loss function plus weights.
**** Rows 11 and 12 show the performance of our method when the thresholds are 0.5 and 0.7, respectively.

TABLE V
PERFORMANCE COMPARISON OF AIRWAY SEGMENTATION METHODS ON EXACT'09-TE.
THE MEAN VALUE AND STANDARD DEVIATION ARE SHOWN FOR EACH METRIC.

	Branch count	Tree length (cm)	BD/BR	TD/TR	FPR
Ours (th=0.5)	**272.9±142.4**	**286.6±158.4**	(BR) **117.7%±38.0%***	(TR) **132.2%±39.8%***	0.094%±0.025%
Xu et al. [46]	128.7±60.3	94.8±44.7	51.7%±10.8%	44.5%±9.4%	0.85%±1.6%**
Yun et al. [30]	163.4±79.4	129.3±66.0	65.7%±13.1%	60.1%±11.9%	4.5%±3.7%**
Qin et al. [34]	190.4***	166.5***	76.7%±11.5%	72.7%±11.6%	3.7%±2.9%**
Zheng et al. [36]	199.9***	180.9***	80.5%±12.5%	79.0%±11.1%	5.8%±4.3%**
Neko [39]	84.5±40.5	61.9±30.9	35.5%±8.2%	30.4%±7.4%	0.89%±1.8%**
UCCTeam [40]	99.0±50.3	75.1±39.4	41.6%±9.0%	36.5%±7.6%	0.71%±1.67%**
FF_ITC [41]	198.3±98.6	177.1±97.0	79.6%±13.5%	79.9%±12.1%	11.92%±13.16%**
MISLAB [9]	104.7±55.2	78.7±41.7	42.9%±9.6%	37.5%±7.1%	0.89%±1.64%**
NTNU [49]	72.4±37.8	54.3±33.9	31.3%±10.4%	27.4%±9.6%	3.60%±3.37%**

* We report the BR and TR values of our method and the BD and TD values of the compared methods.
** The calculation approach of these benchmark methods may be different from ours because their used TN was not specified.
*** Branch count and tree length values were not reported in [25] and [52]. We calculated them based on the reported BD and TD values.

1) Effect of the proposed bronchiole-sensitive loss function: Comparing the use of our new loss function with the original dice loss function plus weights ("w/ org dice + weights" row in Table VI), we note a significant increase in sensitivity (93.5% to 98.4%), BD (71.3% to 88.4%), TD (74.0% to 94.2%), BR (73.1% to 115.0%), and TR (76.9% to 115.5%), indicating the new loss function pushes the model to detect more finer bronchioles.

2) Effect of human-vision-inspired iterative training: In the ablation experiment, we replaced the iterative training strategy with same-frequency training, where all pre-cropped cuboids were selected with the same probability in each training batch. When using iterative training, the model significantly improves on most metrics, particularly structural accuracy.

3) The proposed loss function and training strategy can applied to different model architectures: To demonstrate the generalizability of our training framework, we replaced our backbone model with 3D U-Net, V-Net, and VoxResNet ("w/ 3D U-Net", "w/ V-Net", and "w/ VoxResNet" rows in Table VI). Compared to the situation (Rows 8-10 in Table IV) where no training tricks applied, our loss function and training strategy significantly improve segmentation performance by detecting almost twice longer airway branches.

C. Test results on unseen images

The bottom two rows in Table VI examined the performance of our model in the unseen dataset (QMH) (see Section). Although model performance drops when compared with that on the EXACT'09-TR-LIDC-IDRI dataset, our model still achieves high accuracy (DSC: 84.9% and 85.5%, sensitivity: 88.7% and 95.4%, where the accuracy values show the model performance without and with semi-supervised learning, respectively). In terms of structural accuracy, our method detects more bronchioles than the reference labels (BR: 139.7% and TR: 106.0%).

VI. DISCUSSION

1) Training strategy matters: From Table VI, the effects of the proposed loss function and iterative training strategy are much more significant than other components. Unlike most earlier methods that followed a first-low-generation-then-high-generation training strategy, this study investigated the influence of different training strategies. In our settings, the first-coarse-then-fine training strategy performs a little bit poorer than training on all pre-cropped cuboids with the same frequency (Table VIII). This might be because training with

TABLE VI

ABLATION STUDY AND RESULTS ON THE PRIVATE DATASET (QMH). THE MEAN VALUE AND STANDARD DEVIATION ARE SHOWN FOR EACH METRIC.

Th		DSC	Sensitivity	Precision	BD	TD	BR	TR	FPR
0.5	Ours	90.7%±1.8%	98.4%±1.4%	84.2%±3.4%	88.4%±10.7%	94.2%±5.7%	115.0%±19.0%	115.5%±18.5%	0.224%±0.128%
	w/ org dice + weights	93.9%±1.3%	93.5%±3.1%	94.4%±1.9%	71.3%±14.1%	74.0%±14.1%	73.1%±19.3%	76.9%±19.4%	0.065%±0.046%
	w/o iterative training	90.7%±1.3%	95.7%±3.0%	86.4%±2.9%	76.1%±14.8%	80.4%±13.4%	90.6%±23.7%	90.5%±23.0%	0.180%±0.109%
	w/o feature extraction module	90.5%±1.5%	97.3%±2.2%	84.8%±3.1%	81.0%±12.2%	85.5%±10.5%	103.3%±28.7%	99.4%±24.3%	0.201%±0.115%
	w/o semi	91.6%±1.5%	97.7%±1.6%	86.4%±3.1%	83.7%±11.2%	89.8%±7.8%	104.2%±26.8%	106.4%±25.4%	0.180%±0.116%
	w/o data aug	93.2%±1.4%	97.7%±1.7%	89.1%±3.2%	86.9%±9.8%	89.3%±8.7%	112.9%±32.8%	107.8%±29.2%	0.135%±0.071%
	w/ 3D U-Net [21]	90.5%±1.5%	97.3%±2.2%	84.8%±3.1%	81.0%±12.2%	85.5%±10.5%	103.3%±28.7%	99.4%±24.3%	0.201%±0.115%
	w/ V-Net [47]	87.6%±2.1%	85.4%±5.5%	83.8%±3.7%	77.1%±10.4%	82.5%±9.4%	90.8%±22.8%	89.3%±18.7%	0.143%±0.052%
	w/ VoxResNet [48]	88.0%±3.4%	82.9%±6.4%	81.6%±3.0%	78.1%±12.5%	80.2%±11.9%	95.4%±21.6%	88.9%±19.8%	0.059%±0.031%
0.7	Ours	94.2%±1.1%	96.6%±2.3%	91.2%±2.6%	83.3%±11.4%	90.4±8.4%	93.6%±22.6%	100.7%±23.6%	0.117%±0.079%
	w/ org dice + weights	93.4%±1.8%	90.2%±3.9%	96.9%±1.3%	61.8%±14.8%	70.0%±17.7%	62.2%±14.1%	69.0%±14.9%	0.035%±0.025%
	w/o iterative training	92.6%±1.5%	93.2%±3.9%	92.1%±2.1%	70.8%±16.2%	76.3%±14.2%	78.6%±21.4%	81.5%±19.4%	0.094%±0.064%
	w/o feature extraction module	93.1%±1.2%	95.3%±3.1%	91.0%±2.4%	74.8%±13.7%	79.9%±12.6%	80.8%±19.2%	86.3%±20.1%	0.110%±0.071%
	w/o semi	93.9%±2.4%	95.3%±2.8%	92.7%±2.3%	76.6%±11.9%	82.7%±11.0%	86.9%±21.9%	91.8%±22.2%	0.090%±0.068%
	w/o data aug	94.8%±0.9%	95.5%±2.4%	94.4%±2.5%	80.1%±10.6%	84.8%±9.9%	92.9%±23.2%	95.0%±23.0%	0.066%±0.046%
0.5	QMH w/o semi	84.9%±6.1%	88.7%±8.3%	81.9%±7.0%	75.6%±14.1%	72.9%±11.9%	105.0%±17.0%	88.2%±9.7%	0.039%±0.027%
	QMH w/ semi	85.5%±4.7%	95.4%±1.5%	77.8%±8.1%	90.2%±7.8%	83.7%±6.2%	139.7%±23.7%	106.0%±14.5%	0.058%±0.045%

TABLE VII

BRANCH COUNT, TREE LENGTH, BR, TR, AND DETECTED AIRWAY GENERATIONS FOR THE 20 CASES IN EXACT'09-TE. OUR METHOD WAS DEVELOPED WITH AND WITHOUT SEMI-SUPERVISED LEARNING.

	Branch count			BR	BR	Tree length (cm)			TR	TR	Airway generation (w/ semi)		
	Ref[*]	w/o semi	w/ semi	(w/o semi)	(w/ semi)	Ref[*]	w/o semi	w/ semi	(w/o semi)	(w/ semi)	Avg	Median	Max
CASE21	200	154	174	77.0%	87.0%	120.3	135.1	155.3	112.3%	129.2%	5.7	6	12
CASE22	388	271	328	69.8%	84.5%	341.5	296	357.6	86.7%	104.7%	8.3	9	17
CASE23	285	123	275	43.2%	96.5%	271.3	154.5	361.3	57.0%	133.2%	7.2	7	15
CASE24	187	267	319	142.8%	170.6%	185.8	339.4	377	182.7%	202.9%	5.8	7	15
CASE25	235	365	371	155.3%	157.9%	270.2	451.4	437.7	167.1%	162.0%	4.5	7	15
CASE26	81	107	111	132.1%	137.0%	83.2	142.6	128.9	171.5%	155.0%	4.7	5	9
CASE27	102	115	151	112.7%	148.0%	94.4	139.7	171.7	148.0%	182.0%	4.9	5	10
CASE28	124	167	187	134.7%	150.8%	122	201.1	227.5	164.7%	186.4%	7	7	13
CASE29	185	191	211	103.2%	114.1%	150.3	227.5	237.9	151.4%	158.3%	7.4	8	13
CASE30	196	177	211	90.3%	107.7%	164.4	178.4	187.5	108.6%	114.1%	5.7	6	12
CASE31	215	175	191	81.4%	88.8%	188.4	187.1	195.3	99.3%	103.7%	5.9	6	12
CASE32	234	156	174	66.7%	74.4%	231.4	199.7	212.9	86.3%	92.0%	6.3	7	11
CASE33	169	187	220	110.7%	130.2%	159.6	203.5	230.9	127.5%	144.7%	6.1	6	12
CASE34	459	324	442	70.6%	96.3%	368.8	361.5	440.7	98.0%	119.5%	7.9	8	17
CASE35	345	302	335	87.5%	97.1%	320.4	286.4	305.5	89.4%	95.3%	6.6	7	12
CASE36	365	525	741	143.8%	203.0%	422.6	615.6	792.9	145.7%	187.6%	8.7	9	18
CASE37	186	237	291	127.4%	156.5%	190.3	115.9	135.1	60.9%	71.0%	6.8	7	15
CASE38	99	78	121	78.8%	122.2%	78.2	81.6	108.9	104.4%	139.4%	4.2	5	10
CASE39	521	279	369	53.6%	70.8%	420.1	306.5	395.6	73.0%	94.2%	8.6	9	15
CASE40	390	191	236	49.0%	60.5%	399.3	239.6	271.7	60.0%	68.0%	6.7	7	13
Avg	248.3	219.6	272.9	96.5%	117.7%	229.1	243.2	286.6	114.7%	132.2%	6.5	6.9	13.3
Std	124.5	105.5	142.4	34.3%	38.0%	115.3	127.6	158.4	40.1%	39.8%	1.3	1.3	2.5
Max	521	525	741	155.3%	203.0%	422.6	615.6	792.9	182.7%	202.9%	8.7	9	18
1st quartile	181	156	184	70.4%	88.4%	143.2	151.6	183.6	86.6%	101.6%	5.7	6	12
Median	208	189	228	88.9%	110.9%	189.4	202.3	234.4	106.5%	131.2%	6.5	7	13
3rd quartile	350	273	330	128.6%	148.7%	325.6	298.6	365.2	148.9%	159.2%	7.2	7.3	15
MIN	81	78	111	43.2%	60.5%	78.2	81.6	108.9	57.0%	68.0%	4.2	5	9

[*] The "Ref" columns list the corresponding values of the reference labels quoted from the website of the EXACT'09 challenge.

a more focus on finer bronchioles in the second stage results in the loss of learned features of low-generation airways. On the other hand, first-fine-then-low-coarse training does not have satisfactory performance, either, because the extracted features of finer bronchioles might be lost when training on low generations. In contrast, our proposed iterative training could teach the model to learn unique features of both low and high airway generations while preserving the knowledge of airway features of both high and low generations.

TABLE VIII

DSC VALUES UNDER DIFFERENT TRAINING STRATEGIES INCORPORATED INTO OUR PIPELINE (THRESHOLD = 0.7).

Strategy	DSC	Strategy	DSC
More on coarse ($\beta = 1$)	89.60%	First coarse then fine	90.00%
More on coarse ($\beta = 10$)	90.40%	First fine then coarse	88.20%
More on fine ($\beta = 1$)	87.30%	Same frequency	92.60%
More on fine ($\beta = 10$)	∼0%	Iterative	94.20%

[*] "coarse" represents airways of low generations and "fine" represents airways of high generations.

2) The proposed bronchiole-sensitive loss function is more effective than the original dice loss function: As shown in Table VI, in terms of sensitivity and structural accuracy, the proposed loss function performs significantly better (around 5% to 30% improvement) than the original dice loss function. Results show that the original dice loss is more "conservative" than the proposed one and tends to circumscribe the airway segmentation within the airway mask in reference labels.

Therefore, dice loss is deemed suboptimal because the labels used in this study serve as a reference, not the ground truth.

3) Review metrics for airway segmentation: Most existing works followed the metrics built in the EXACT‘09 challenge in 2009 [9]. However, because we only had the “reference” labels, instead of the “ground truth” segmentation, we proposed to devise new metrics. First, quantitative evaluation was insufficient. We proposed to include visual inspection because the model segmentation might make mistakes in some image regions while finding more bronchioles than the reference in some others. Then, for quantitative metrics, we need to reconsider the implication of False Positive FP defined in Table II (also “leakage” in [9]). For traditional methods using techniques, such as region growing, FP may usually be meaningless leaking volumes. However, for deep learning approaches, after removing unconnected shapes, the FP may include the undetected airway segments in the reference labels. Therefore, both Branch Detected (BD) and Tree-length Detected (TD) metrics are not deemed comprehensive measures. As a result, we proposed Branch Ratio (BR) and Tree-length Ratio (TR) to compare the total number of branches and length of airway trees between model segmentation and reference labels. Moreover, False Positive Rate (FPR) may not be a good measure because it assumes all FP are meaningless leakage volumes and its definition is ambiguous.

4) Future directions: To translate our method into clinical practice, further investigations are needed.

- Design of improved model robustness to tackle the problem of domain shift, i.e., training on some CT scans but application on another set of CT scans: It is not feasible to ask the clinicians to label the airways and retrain the model on the CT scans acquired at their affiliated hospitals. We need to build techniques to increase model robustness by making the model learn the features of airways in general CT scans, not a specific set of CT scans.
- Model interpretability to increase its trustworthiness to humans: Clinical applications are high-stakes scenarios. Humans naturally need explanations on a machine’s decisions to use the machine wisely and give timely feedback on parameter settings and machine reconfiguration.

VII. Conclusions

In this paper, we present a novel airway segmentation pipeline which provide extensive airway roadmaps with more detailed bronchioles by the proposed bronchiole-sensitive loss function and human-vision-inspired iterative training strategy. Our method is general. The proposed loss function and iterative training strategy can be applied to any backbone model and can be combined with semi-supervised learning to further improve model accuracy. The results demonstrated that our pipeline was robust on new CT scans and outperformed existing methods. Additionally, except airway segmentation, our method can also be applied to other scenarios when segmenting fine and long tubular structures in biomedical images.

Appendix A
Backbone Model

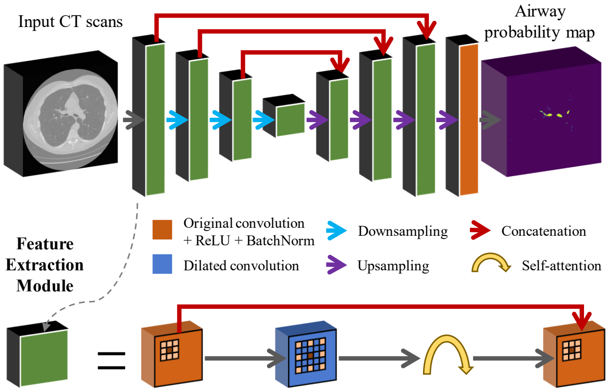

Fig. 4. Structure of the backbone model.

Please see Figure 4 for the structure of the backbone model. We adopted dilated convolutions as they enlarged the feature extraction area without increasing the size of the model. For our model, the receptive field size r_0 of the feature map after l-th layer would be [50]

$$r_0 = \sum_{p=1}^{l} \left(d_p \left(k_p - 1 \right) \prod_{q=1}^{p-1} s_q \right) + 1, \tag{7}$$

where k_p is the size of the p-th kernel, d_p is the dilation spacing between kernel elements, and s_q is the stride of the q-th kernel.

Therefore, compared with the original 3D U-Net where only typical convolution kernels are used, our feature extraction module detects approximately $\prod_{p \in D} d_p$ times larger areas.

The self-attention we incorporated into the model was inspired by Squeeze-and-Excitation Network [51]. It could further improve the model performance by emphasizing the areas where the model should focus on while deemphasizing the minor regions.

Appendix B
Teacher-Student Training

In brief, the proposed teacher-student training uses the model that is trained on labeled dataset $\mathcal{D}$ as a teacher model Φ_θ^T to generate pseudo labels for unlabeled images $\mathcal{U}$, and a student model Φ_θ^S is then trained with both labeled and unlabeled images. We used data augmentation to help the student model Φ_θ^S learn by analogy. General data augmentation approaches for medical images, including random flip and affine, random blur, random motion [45], were applied on chest CT scans, while some airway morphology specific data augmentation approaches were applied on airway annotations. Details of the student-teacher training are provided in Algorithm 1.

References

[1] T. Ishiwata, A. Gregor, T. Inage, and K. Yasufuku, “Bronchoscopic navigation and tissue diagnosis,” *General thoracic and cardiovascular surgery*, vol. 68, no. 7, pp. 672–678, 2020.

Algorithm 1: Teacher-student training

```
Init: Given the optimized model $\Phi_{\theta*}$ trained on a labeled dataset $\mathcal{D}$, initialize a teacher model $\Phi_\theta^T = \Phi_{\theta*}$ and a student model $\Phi_\theta^S = \Phi_{\theta*}$. As the unlabeled dataset $\mathcal{U}$ has large size, divide $\mathcal{U}$ into $V$ batches $\{\mathcal{U}_b\}_{b=1}^V$. Set two probability values, $q^t$ and $q^c$, used for data augmentation. Set the number of iterations $I$.
repeat
    for $\mathcal{U}_b$ in $\{\mathcal{U}_b\}_{b=1}^V$ do
        /* Generate pseudo labels */
        for $\boldsymbol{x}_i$ in $\mathcal{U}_b$ do
            /* Airway morphology specific data augmentation */
            $\boldsymbol{A}_i = \Phi_\theta^T(\boldsymbol{x}_i) > t$ ($t = 0.5$ with probability of $q^t$, otherwise $t = 0.7$) ;
            Set $\boldsymbol{A}_i$ to be $\boldsymbol{S}_{h*}$ with probability of $q^c$ (see Sec. III-D for the definition of $\boldsymbol{S}_{h*}$) ;
        Integrate $\mathcal{D}$ and $\mathcal{U}_b$ to be a new dataset $\mathcal{W}:\ (\boldsymbol{x}_i,\ \widehat{\boldsymbol{a}_i})$, where $\widehat{\boldsymbol{a}_i}$ is $\boldsymbol{a}_i$ or $\boldsymbol{A}_i$ ;
        /* Train the student model */
        $\theta_b^* = \underset{\{(\boldsymbol{x}_i^j, \widehat{\boldsymbol{a}_i^j})\} \in \mathcal{W}}{\arg\min} L\left(\Phi_\theta^S\left(\left\{g\left(\boldsymbol{x}_i^j\right)\right\}\right), \widehat{\left\{\boldsymbol{a}_i^j\right\}}\right)$ (train $\Phi_\theta^S$ on $\mathcal{W}$ with augmentation $g$) ;
        Set $\Phi_\theta^S = \Phi_{\theta_b^*}^S$ ;
    /* Update the teacher model */
    Update $\Phi_\theta^T = \Phi_\theta^S$ ;
until Finishing $I$ iterations;
```

[2] E. Edell and D. Krier-Morrow, "Navigational bronchoscopy: Overview of technology and practical considerations—new current procedural terminology codes effective 2010," *Chest*, vol. 137, no. 2, pp. 450–454, 2010.

[3] F. Asano, R. Eberhardt, and F. J. Herth, "Virtual bronchoscopic navigation for peripheral pulmonary lesions," *Respiration*, vol. 88, no. 5, pp. 430–440, 2014.

[4] S. V. Kemp, "Navigation bronchoscopy," *Respiration*, vol. 99, no. 4, pp. 277–286, 2020.

[5] P. Berger, V. Perot, P. Desbarats, J. M. Tunon-de Lara, R. Marthan, and F. Laurent, "Airway wall thickness in cigarette smokers: quantitative thin-section ct assessment," *Radiology*, vol. 235, no. 3, pp. 1055–1064, 2005.

[6] W. O. Reece, "Overview of the respiratory system," *Dukes' physiology of domestic animals*, vol. 203, 2015.

[7] E. M. Van Rikxoort and B. Van Ginneken, "Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review," *Physics in Medicine & Biology*, vol. 58, no. 17, p. R187, 2013.

[8] J. Pu, S. Gu, S. Liu, S. Zhu, D. Wilson, J. M. Siegfried, and D. Gur, "Ct based computerized identification and analysis of human airways: a review," *Medical physics*, vol. 39, no. 5, pp. 2603–2616, 2012.

[9] P. Lo, B. Van Ginneken, J. M. Reinhardt, T. Yavarna, P. A. De Jong, B. Irving, C. Fetita, M. Ortner, R. Pinho, J. Sijbers *et al.*, "Extraction of airways from ct (exact'09)," *IEEE Transactions on Medical Imaging*, vol. 31, no. 11, pp. 2093–2107, 2012.

[10] D. Aykac, E. A. Hoffman, G. McLennan, and J. M. Reinhardt, "Segmentation and analysis of the human airway tree from three-dimensional x-ray ct images," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 940–950, 2003.

[11] H. Shi, W. C. Scarfe, and A. G. Farman, "Upper airway segmentation and dimensions estimation from cone-beam ct image datasets," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, no. 3, pp. 177–186, 2006.

[12] I. Cheng, S. Nilufar, C. Flores-Mir, and A. Basu, "Airway segmentation and measurement in ct images," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 795–799.

[13] J. Tschirren, E. A. Hoffman, G. McLennan, and M. Sonka, "Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose ct scans," *IEEE transactions on medical imaging*, vol. 24, no. 12, pp. 1529–1539, 2005.

[14] ——, "Segmentation and quantitative analysis of intrathoracic airway trees from computed tomography images," *Proceedings of the American Thoracic Society*, vol. 2, no. 6, pp. 484–487, 2005.

[15] A. Fabijańska, "Two-pass region growing algorithm for segmenting airway tree from mdct chest scans," *Computerized Medical Imaging and Graphics*, vol. 33, no. 7, pp. 537–546, 2009.

[16] M. W. Graham, J. D. Gibbs, D. C. Cornish, and W. E. Higgins, "Robust 3-d airway tree segmentation for image-guided peripheral bronchoscopy," *IEEE transactions on medical imaging*, vol. 29, no. 4, pp. 982–997, 2010.

[17] C. Fetita, M. Ortner, P.-Y. Brillet, F. Prêteux, P. Grenier *et al.*, "A morphological-aggregative approach for 3d segmentation of pulmonary airways from generic msct acquisitions," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 215–226.

[18] A. P. Kiraly, W. E. Higgins, G. McLennan, E. A. Hoffman, and J. M. Reinhardt, "Three-dimensional human airway segmentation methods for clinical virtual bronchoscopy," *Academic radiology*, vol. 9, no. 10, pp. 1153–1168, 2002.

[19] Q. Meng, T. Kitasaka, Y. Nimura, M. Oda, J. Ueno, and K. Mori, "Automatic segmentation of airway tree based on local intensity filter and machine learning technique in 3d chest ct volume," *International journal of computer assisted radiology and surgery*, vol. 12, no. 2, pp. 245–261, 2017.

[20] B. van Ginneken, W. Baggerman, and E. M. van Rikxoort, "Robust segmentation and anatomical labeling of the airway tree from thoracic ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 219–226.

[21] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[22] D. Jin, Z. Xu, A. P. Harrison, K. George, and D. J. Mollura, "3d convolutional neural networks with graph refinement for airway segmentation using incomplete data labels," in *International workshop on machine learning in medical imaging*. Springer, 2017, pp. 141–149.

[23] A. G.-U. Juarez, H. A. Tiddens, and M. de Bruijne, "Automatic airway segmentation in chest ct using convolutional neural networks," in *Image analysis for moving organ, breast, and thoracic images*. Springer, 2018, pp. 238–250.

[24] S. A. Nadeem, E. A. Hoffman, J. C. Sieren, A. P. Comellas, S. P. Bhatt, I. Z. Barjaktarevic, F. Abtin, and P. K. Saha, "A ct-based automated algorithm for airway segmentation using freeze-and-grow propagation and deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 405–418, 2020.

[25] A. Garcia-Uceda, R. Selvan, Z. Saghir, H. Tiddens, and M. de Bruijne, "Automatic airway segmentation from computed tomography using robust and efficient 3-d convolutional neural networks," *arXiv preprint arXiv:2103.16328*, 2021.

[26] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[27] C. Wang, Y. Hayashi, M. Oda, H. Itoh, T. Kitasaka, A. F. Frangi, and K. Mori, "Tubular structure segmentation using spatial fully connected network with radial distance loss for 3d medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 348–356.

[28] A. G.-U. Juarez, R. Selvan, Z. Saghir, and M. de Bruijne, "A joint 3d unet-graph neural network-based method for airway segmentation from chest cts," in *International workshop on machine learning in medical imaging*. Springer, 2019, pp. 583–591.

[29] R. Selvan, T. Kipf, M. Welling, A. G.-U. Juarez, J. H. Pedersen, J. Petersen, and M. de Bruijne, "Graph refinement based airway extraction using mean-field networks and graph neural networks," *Medical Image Analysis*, vol. 64, p. 101751, 2020.

[30] J. Yun, J. Park, D. Yu, J. Yi, M. Lee, H. J. Park, J.-G. Lee, J. B. Seo, and N. Kim, "Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5 dimensional convolutional neural net," *Medical image analysis*, vol. 51, pp. 13–20, 2019.

[31] Y. Qin, M. Chen, H. Zheng, Y. Gu, M. Shen, J. Yang, X. Huang, Y.-M. Zhu, and G.-Z. Yang, "Airwaynet: a voxel-connectivity aware approach for accurate airway segmentation using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 212–220.

[32] Y. Qin, Y. Gu, H. Zheng, M. Chen, J. Yang, and Y.-M. Zhu, "Airwaynet-se: A simple-yet-effective approach to improve airway segmentation using context scale fusion," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 809–813.

[33] H. Zhang, M. Shen, P. L. Shah, and G.-Z. Yang, "Pathological airway segmentation with cascaded neural networks for bronchoscopic navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9974–9980.

[34] Y. Qin, H. Zheng, Y. Gu, X. Huang, J. Yang, L. Wang, F. Yao, Y.-M. Zhu, and G.-Z. Yang, "Learning tubule-sensitive cnns for pulmonary airway and artery-vein segmentation in ct," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1603–1617, 2021.

[35] Y. Qin, H. Zheng, Y. Gu, X. Huang, J. Yang, L. Wang, and Y.-M. Zhu, "Learning bronchiole-sensitive airway segmentation cnns by feature recalibration and attention distillation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 221–231.

[36] H. Zheng, Y. Qin, Y. Gu, F. Xie, J. Yang, J. Sun, and G.-Z. Yanga, "Alleviating class-wise gradient imbalance for pulmonary airway segmentation," *IEEE Transactions on Medical Imaging*, 2021.

[37] W. Wu, Y. Yu, Q. Wang, D. Liu, and X. Yuan, "Upper airway segmentation based on the attention mechanism of weak feature regions," *IEEE Access*, vol. 9, pp. 95 372–95 381, 2021.

[38] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[39] H. Balacey, "Mise en place d'une chaîne complète d'analyse de l'arbre trachéo-bronchique à partir d'examen (s) issus d'un scanner-ct: de la 3d vers la 4d," Ph.D. dissertation, Bordeaux 1, 2013.

[40] P. Nardelli, K. A. Khan, A. Corvò, N. Moore, M. J. Murphy, M. Twomey, O. J. O'Connor, M. P. Kennedy, R. S. J. Estépar, M. M. Maher *et al.*, "Optimizing parameters of an open-source airway segmentation algorithm using different ct images," *Biomedical engineering online*, vol. 14, no. 1, pp. 1–24, 2015.

[41] T. Inoue, Y. Kitamura, Y. Li, and W. Ito, "Robust airway extraction based on machine learning and minimum spanning tree," in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670. International Society for Optics and Photonics, 2013, p. 86700L.

[42] C. Wei, K. Shen, Y. Chen, and T. Ma, "Theoretical analysis of self-training with deep networks on unlabeled data," *arXiv preprint arXiv:2010.03622*, 2020.

[43] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[45] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106236, 2021.

[46] Z. Xu, U. Bagci, B. Foster, A. Mansoor, J. K. Udupa, and D. J. Mollura, "A hybrid method for airway segmentation and automated measurement of bronchial wall thickness on ct," *Medical image analysis*, vol. 24, no. 1, pp. 1–17, 2015.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[48] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[49] E. Smistad, A. C. Elster, and F. Lindseth, "Gpu accelerated segmentation and centerline extraction of tubular structures from medical images," *International journal of computer assisted radiology and surgery*, vol. 9, no. 4, pp. 561–575, 2014.

[50] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, vol. 4, no. 11, p. e21, 2019.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.