# Adversarial Transformer for Repairing Human Airway Segmentation

Zeyu Tang, Yang Nan, Simon Walsh, and Guang Yang, *Senior Member, IEEE*

***Abstract*—Automated airway segmentation models often suffer from discontinuities in peripheral bronchioles, which limits their clinical applicability. Furthermore, data heterogeneity across different centres and pathological abnormalities pose significant challenges to achieving accurate and robust segmentation in distal small airways. Accurate segmentation of airway structures is essential for the diagnosis and prognosis of lung diseases. To address these issues, we propose a patch-scale adversarial-based refinement network that takes in preliminary segmentation and original CT images and outputs a refined mask of the airway structure. Our method is validated on three datasets, including healthy cases, pulmonary fibrosis, and COVID-19 cases, and quantitatively evaluated using seven metrics. Our method achieves more than a 15% increase in the detected length ratio and detected branch ratio compared to previously proposed models, demonstrating its promising performance. The visual results show that our refinement approach, guided by a patch-scale discriminator and centreline objective functions, effectively detects discontinuities and missing bronchioles. We also demonstrate the generalizability of our refinement pipeline on three previous models, significantly improving their segmentation completeness. Our method provides a robust and accurate airway segmentation tool that can help improve diagnosis and treatment planning for lung diseases.**

***Index Terms*—Segmentation, airway, GAN, transformer, refinement, explainable.**

Manuscript received 4 October 2022; revised 13 May 2023 and 9 June 2023; accepted 24 June 2023. Date of publication 28 June 2023; date of current version 5 October 2023. This work was supported in part by ERC IMI under Grant 101005122, in part by H2020 under Grant 952172, in part by MRC under Grant MC/PC/21013, in part by Royal Society under Grant IEC/NSFC/211235, in part by Imperial College UROP, in part by NVIDIA Academic Hardware Grant Program, in part by SABER Project through Boehringer Ingelheim Ltd., NIHR Imperial Biomedical Research Centre under Grant RDA01, and in part by UKRI Future Leaders Fellowship under Grant MR/V023799/1. *(Zeyu Tang and Yang Nan are the co-first authors.) (Simon Walsh and Guang Yang are the co-last senior authors.) (Corresponding author: Guang Yang.)*

Zeyu Tang is with the National Heart and Lung Institute, Imperial College London, SW7 2BX London, U.K., and also with the Department of Bioengineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: zeyu.tang19@imperial.ac.uk).

Yang Nan is with the National Heart and Lung Institute, Imperial College London, SW7 2BX London, U.K. (e-mail: y.nan20@imperial.ac.uk).

Simon Walsh and Guang Yang are with the National Heart and Lung Institute, Imperial College London, SW7 2BX London, U.K., and also with the Royal Brompton Hospital, SW3 6NP London, U.K. (e-mail: s.walsh@imperial.ac.uk; g.yang@imperial.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3290136

## I. Introduction

AIRWAY segmentation is a crucial step in the diagnosis and prognosis of pulmonary diseases using chest computerised tomography (CT) scans. Manual segmentation by radiologists is highly time-consuming and error-prone due to the large volume of CT data and complex airway tree structure. To relieve the radiology experts from these tedious manual labelling processes, many automated and semi-automated algorithms are being developed. The performance of traditional segmentation methods, such as thresholding [1] and region-growing [2], have been benchmarked in the EXACT'09 challenge [3], and no algorithm could extract more than an average of 74% of the total tree length. In recent years, convolutional neural networks have shown promising results in airway segmentation, with models like U-Net [4], 3-D U-Net [5], V-Net [6], and their derivatives [7], [8], [9] being used for this task. Despite the significant improvements in the Dice score achieved by CNN-based models for airway segmentation, local discontinuity remains a challenge, as shown in Fig. 1. This challenge is particularly prevalent in the peripheral region of the lung, where there exists an imbalance between the target distal small airways and the background. Another limitation is that these deep neural networks are trained on image patches due to the high computation complexity, which makes it challenging to detect small errors until patches are stitched into the full image. Recent works by [10], [11], and [12] have attempted to address the issue of discontinuity by designing complex architectures and loss functions. However, these methods do not account for the variability in airway structure caused by pathological conditions such as honeycombing, bronchial wall thickening, and bronchiectasis, which can make the model less robust and lead to missing bronchi or bronchioles in the segmentation. Moreover, the heterogeneity of data across different medical centres and institutes and variable airway branching patterns between individuals further compound the challenge.

To tackle these problems, we present in this article a simple yet effective adversarial-based patch-scale refinement network to improve the connectivity of preliminary segmentation on both normal and pathological cases. We reason that since it is difficult to derive a single model to accomplish the task, refinement can be done to the existing segmentation. Compared to other similar GAN-based medical segmentation models, we have several novel improvements. Specifically, *tanh* is used as the final activation function in the generator to remove or add pixels to the preliminary labels guided by two centreline objective functions. The refined labels are then dilated using a fixed-size

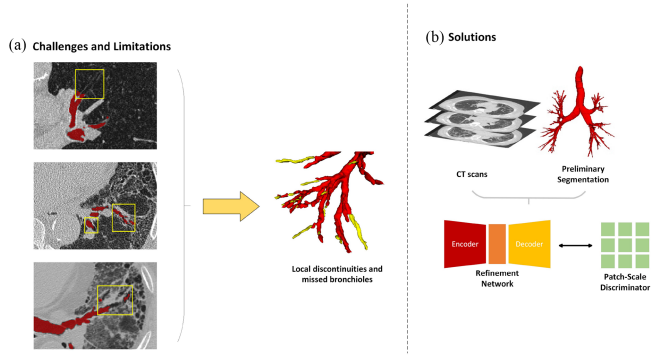


Fig. 1. Current challenges and solutions of airway segmentation. (a) Examples of blurry boundaries, pathological changes and low resolution (indicated in yellow boxes) leading to discontinued segmentation and missing branches highlighted in yellow. (b) This article proposes an adversarial-based refinement network to help achieve complete and robust airway segmentation. Patch-scale discriminators are incorporated to penalize local structural errors.

kernel and then times with the CT scan image before feeding into the discriminator. The dilation helps expand the visual field of the discriminator, which we assume could better help it distinguish real and fake pairs. We let the discriminator take image patches instead of the complete image because this allows penalizing false segmentation at the local scale and better capture the high-frequency details (i.e., tiny distal bronchioles). Our model's performance was assessed on three datasets: the BAS dataset and two in-house datasets that comprised 25 cases of both pulmonary fibrosis and COVID-19. The evaluation metrics utilized were the detected length ratio, detected branch ratio, and false negative rate. Our model yielded a notable improvement of over 15% in all three metrics. The refinement pipeline also works on preliminary segmentation of other models, suggested by rises of more than 10% in the metrics mentioned above. We also did an ablation study to analyze the contribution of each component in our refinement scheme.

The main contributions of this article can be summarised as follows:

- Unlike traditional methods that aim to delineate the airway structure from scratch, our refinement model focuses on modifying the preliminary segmentation by adding or removing pixels to improve connectivity. This approach yields better results compared to traditional methods, especially in the peripheral regions of the lungs where discontinuity is more prevalent. Furthermore, the proposed refinement pipeline can be easily integrated with existing segmentation models, offering a flexible and versatile solution for improving the segmentation performance.
- Our method utilizes patch-scale discriminators to help the generator better focus on small airway structures.
- Two novel centreline-based loss functions are implemented to help the generator model maintain the continuity of refined results.
- The refinement pipeline is also tested on pathological cases, which were neglected by previous research, and achieves state-of-the-art performance.

## II. Related Works

### A. Airway Segmentation

The development of the U-Net family [4], [5], [6], [13] has revolutionized medical image segmentation, and encoding-decoding-based architectures have become the mainstream in this field. Numerous deep-learning models have been proposed for human airway segmentation, and some of them have achieved impressive results in terms of IoU, detected length ratio, and detected branch ratio. Qin et al. [11] proposed AirwayNet, a two-step approach to preserve the CNN's high performance in segmenting peripheral bronchioles. Zheng et al. [12] developed WingsNet with group supervision to improve the CNN's learning of small structures. They also introduced a new objective function called the general union loss to address the imbalance between large and small airways. Charbonnier et al. [7] proposed a ConvNet to detect and remove airway segmentation leakage. They used a method that combined segmentations generated by varying parameters to increase the detected length and reduce leakage. Jin et al. [8] used a 3-D FCN for high-quality labeling from incomplete ones. They also employed a graph-based refinement approach that included fuzzy connectivity and skeletonization. Wang et al. [10] proposed a novel radial distance loss function based on the distance transform to help the network recover tiny tubular structures. However, discontinuous predictions and missed tiny branches remain persistent issues.

### B. Transformer-Based Medical Segmentation

Transformers were originally proposed for NLP-related tasks due to their ability to encode long-range dependencies, but a recent study [14] showed that the architecture can also work with images. This has led to a surge of interest in the use of transformers in medical imaging. For 3-D segmentation tasks, Wang et al. [15] proposed TransBTS which positioned a transformer module in between an encoder and a decoder as the bottleneck layer. Hatamizadeh et al. [16] introduced a U-Net Transformer (UNETR) that uses a transformer as the encoding part of the U-Net. Li et al. [17] proposed a Squeeze-and-Expansion Transformer where a squeezed attention block regularizes the self-attention module, and an expansion block learns diversified representations. Liu et al. [18] proposed a hierarchical transformer which adds a shifted window scheme to the multi-head attention (MSA) module. Unlike our methods, all previous work used the transformer module as a generator, not as a discriminator.

### C. Adversarial-Based Medical Segmentation

GAN-based segmentation has gained significant momentum in the field of medical imaging in recent years. Zhang et al. [19] used a modified 2-D dense U-Net to generate initial predictions of the airway and then employed a series of morphological operations to extract the small airways. A 3-D dense U-Net is then used to learn from the sample of small airways using adversarial training with a cGAN. Zhao et al. [20] proposed a generative adversarial learning model with large receptive fields using the dilated residual block as the generator and a multi-layer

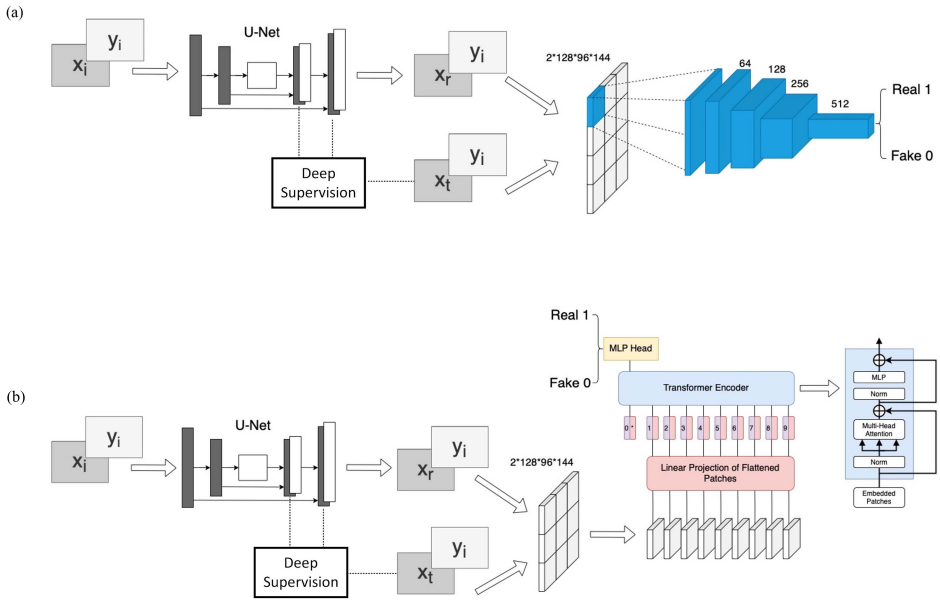


Fig. 2. Proposed refinement model for airway segmentation utilizes U-Net with multi-scale supervision as the generator and a patch-scale discriminator. The discriminator is either a PatchGAN with five convolution layers that extracts features in (a) or a ViT that splits the image into patches and encodes them into fixed-dimensional vectors with embedded positions in (b).

TABLE I
PRELIMINARY SEGMENTATION GENERATED FOR TRAINING

| | Model | IoU | Dice | DLR | DBR | Precision | Leakages | AMR |
|---|---|---|---|---|---|---|---|---|
| Valid | small U-Net | 0.8372±0.0752 | 0.9093±0.0505 | 0.7346±0.1419 | 0.6425±0.1555 | 0.9344±0.0314 | 0.0621±0.0269 | 0.1117±0.0762 |
| Fibrosis | small U-Net | 0.7911±0.0532 | 0.8823±0.0349 | 0.5616±0.1076 | 0.4829±0.1218 | 0.9296±0.0294 | 0.0657±0.0362 | 0.1568±0.0637 |
| COVID | small U-Net | 0.8637±0.1017 | 0.9228±0.0756 | 0.7173±0.1507 | 0.6302±0.1507 | 0.9566±0.0134 | 0.0407±0.0137 | 0.1008±0.1079 |

CNN with residual blocks as the discriminator. Guo et al. [21] used a dense U-Net with an inception module as the generator and a multi-layer CNN as the discriminator. Park et al. [22] designed an M-GAN which also consists of a multi-layer CNN with residual blocks as the discriminator. However, unlike previous works that let the discriminator take the entire image, our patch-scale approach is novel. Additionally, some of the previous studies utilized complex multi-stage pipelines, which did not achieve competitive results compared to others.

## III. METHODOLOGY

This section details our novel refinement method, including a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ shown in Fig. 2. The generator takes in the preliminary segmentation $x_i$ along with its corresponding CT image $y_i$ and outputs a refined label $x_r = \mathcal{G}(y_i, x_i)$. The discriminator $\mathcal{D}$ then tries to distinguish the $y_i \cdot (x_r \oplus k)$ from the $y_i \cdot (x_t \oplus k)$, where $k$ is a kernel of size $5 \times 5 \times 5$, and $x_t$ is the ground truth.

### A. Generator

We employed a 3-D U-Net [5] with multi-scale deep supervision [23] as the generator to segment the airway structure. To feed the generator, we combined CT images and preliminary labelling into a 2-channel input, which was then cut into patches with a dimension of $128 \times 96 \times 144$. The preliminary labelling was generated using a small and efficient version of U-Net [24]. The preliminary results were evaluated using various metrics, which are presented in Table I. To refine the segmentation output, we replaced the sigmoid activation layer with *tanh*, as it produces an output that lies within the range of $[-1, 1]$. We set the threshold to 0, where outputs less than zero are classified as -1 to remove false positive regions in the preliminary labels, while outputs greater than zero are classified as 1 to eliminate false negative regions. We also used specific loss functions to ensure the connectivity of the refinement, which will be discussed in detail in a subsequent section.

### B. Discriminator

Since previous work in medical segmentation has never tried to use discriminators on the scale of patches, we would like to test two types of them. One is the PatchGAN [25], which assumes the pixels separated by a patch distance are independent, and it divides the image into smaller patches and applies convolutional operations on these patches individually. The output of the convolutional layers is then used to determine whether the input image is real or fake. The other one is the vision transformer [14], which models an image as a sequence of patches and focuses on the long-range dependencies between them. It encodes these

patches into fixed-dimensional vectors while preserving positional information using embedded positions. We hypothesized that if the discriminator can better distinguish the ground truth from the refined labels, then the generator can be better trained to focus on small-scale structures and therefore less breakage in the peripheral bronchioles. Both refined labels and true labels are dilated using a kernel of size $5 \times 5 \times 5$ and then timed with the corresponding CT images prior to feeding into the 5-layer Markovian discriminator, which then tries to classify each $70 \times 70 \times 70$ patch as a synthesized label or ground truth. We also adopted a ViT-Base(number of layers = 12, Hidden Size = 768, MLP size = 3072, heads = 12) as our discriminator and change the *tanh* in the MLP head to LeakyReLU with negative slope set to 0.2, and specified the patch size to be $32 \times 38 \times 36$.

### C. Loss Functions

In order to train the refinement model, we adopt a hybrid loss function that combines several loss terms. Specifically, we incorporate the GAN adversarial loss with the Dice loss [6], the cl-Dice loss [26], and the Continuity and Completeness F-score (CCF) [27]. The mathematical definitions of these loss terms are as follows:

$$\text{Dice} = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{1}$$

where $p_i \in P$ is the predicted binary segmentation and $g_i \in G$ is the ground truth binary volume. The summation is taken over $N$ voxels. To maintain the topological integrity of the predicted airway labels, we used clDice defined as:

$$\text{clDice} = \frac{2 \times T_{prec} \times T_{sens}}{T_{prec} + T_{sens}} \tag{2}$$

where *Topology precision* $T_{prec} = \frac{|S_P \cap V_L|}{|S_P|}$ and *Topology sensitivity* $T_{sens} = \frac{|S_L \cap V_P|}{|S_L|}$. $V_L$ represents the ground truth mask and $V_P$ represents the predicted mask. $S$ is the skeletonised version of $V$. Dice and clDice are together in the following manner:

$$L_D = (1-\alpha)(1-Dice) + \alpha(1-clDice) \tag{3}$$

where $\alpha \in [0, 0.5]$ is a weight parameter. To further enhance the centreline detection, we employed another objective function focused on continuity and completeness:

$$\begin{aligned} L_{CCF} &= 1 - C \\ &= 1 - \frac{\sum X \cdot Y_{CL}}{\sum Y_{CL}} \end{aligned} \tag{4}$$

where $X$ is the predicted airway labels and $Y_{CL}$ is the centreline of the ground truth. L1 loss is used to maintain the overall segmentation accuracy.

$$L_1 = |\sigma(x_i) - y_i| \tag{5}$$

The three loss functions described above are combined in the following manner across each layer $j$ of supervision:

$$L_j = \psi_1 L_1 + \psi_2 L_{CCF} + \psi_3 L_D + \psi_4 L_{cGAN} \tag{6}$$

where $\psi_1$ to $\psi_4$ are weight parameters with range [0,1.0], and $L_{cGAN}$ is the GAN adversarial loss:

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[logD(x, y)] + \mathbb{E}_{x,z}[logG(x, z)] \tag{7}$$

The final loss is computed across all supervision layers in the following fashion:

$$L = \phi_1 L_{layer1} + \phi_2 L_{layer2} + \phi_3 L_{layer3} + \phi_4 L_{final} \tag{8}$$

where $\phi_1$ to $\phi_4$ are weight parameters for each layer.

## IV. Experiments and Results

### A. Dataset

We trained and evaluated our model on the Binary Airway Segmentation (BAS) dataset which contains 90 CT scans with 20 of them from the training set of EXACT'09 [3][1] and 70 of them are from LIDC-IDRI [28]. The original LIDC-IDRI dataset includes 1018 cases but with no airway annotations. [11] and [12] selected 70 cases whose slice thickness is less than or equal to 0.625 mm and carefully annotated them.[2] We split the 90 scans into 72 for training and 18 for testing. In addition, we also tested our model on the in-house fibrosis datasets (25 cases) and COVID-19 datasets (25 cases) respectively.

### B. Implementation Details

We implemented the model using PyTorch and trained it for 150 epochs on an NVIDIA GeForce RTX 3090. We utilized an *Adam* solver with a learning rate of 0.001 and set the learning rate to decay by half at the 50th, 80th, 100th, and 120th epochs for the generator. For the discriminator, we used an *Adam* solver with a learning rate of 0.001. We performed backpropagation on the discriminator first and then on the generator. To augment the data, we randomly applied horizontal and vertical flips with a probability of 0.5 during training. The value of hyperparameter $\alpha$ is set to 0.5, as suggested by [26]. We set the values of $\psi_1$ to $\psi_4$ to be 0.1, 0.6, 0.3, and 0.01, respectively, and the values of $\phi_1$ to $\phi_4$ to be 0.1, 0.1, 0.3, and 0.5, respectively.

During the inference stage on the test dataset, we extracted input patches using a sliding window approach (window size = $128 \times 96 \times 144$) with 50% overlap in the x, y, and z directions. We stitched the patches of binary predictions back into the full image and preserved the largest connected component to reduce noise.

### C. Evaluation Metrics

Given the binary voxel-wise prediction $X$ and the ground truth $Y$, we adopt IoU and Dice coefficient to measure the overall segmentation accuracy as well as other metrics specific for tree-like structures, defined as follows:

*IoU and Dice Coefficient* both measure the proportion of overlapping between the prediction and the ground truth:

$$IoU = \frac{XY}{X + Y - XY} \tag{9}$$

[1]Accessible at http://image.diku.dk/exact/

[2]Accessible at https://geronsushi.github.io/lung.html

TABLE II
COMPARISON EXPERIMENTS ON DIFFERENT DATASETS

| | Model | IoU↓ | Dice↓ | DLR↑ | DBR↑ | Precision↓ | Leakage↑ | AMR↓ |
|---|---|---|---|---|---|---|---|---|
| Valid | Wang et al. [10]† | 0.7330±0.0786‡ | 0.8434±0.0554‡ | 0.8505±0.1227‡ | 0.7858±0.1420‡ | 0.7636±0.0737‡ | 0.3070±0.1412‡ | 0.0507±0.0619 |
| | WingsNet [12]* | **0.8544±0.0673**‡ | **0.9120±0.0422**‡ | 0.8698±0.1175‡ | 0.8166±0.1305‡ | **0.9458±0.0271**‡ | **0.0529±0.0300**‡ | 0.1002±0.0769‡ |
| | NaviAirway [29]* | 0.8348±0.0335 | 0.9096±0.0200 | 0.8734±0.0715 | 0.8099±0.0950‡ | 0.8672±0.0406 | 0.1500±0.0536 | **0.0413±0.0304** |
| | **U-Net+PatchGAN** | 0.8150±0.0519 | 0.8971±0.0330 | 0.8828±0.1012 | 0.8337±0.1263 | 0.8556±0.0393 | 0.1620±0.0522 | 0.0544±0.0525 |
| | **U-Net+ViT** | 0.8132±0.0518 | 0.8961±0.0334 | **0.8902±0.0967** | **0.8439±0.1261** | 0.8600±0.0401 | 0.1549±0.0539 | 0.0623±0.0504 |
| Fibrosis | Wang et al. [10]† | 0.6979±0.0647‡ | 0.8203±0.0462‡ | 0.6961±0.0924‡ | 0.6261±0.1117‡ | 0.7468±0.0773‡ | 0.3272±0.1445‡ | 0.0823±0.0388 |
| | WingsNet [12]* | 0.8052±0.0539‡ | 0.8910±0.0440‡ | 0.6951±0.0977‡ | 0.6198±0.1175‡ | **0.9505±0.0116**‡ | **0.0438±0.0112**‡ | 0.1595±0.0576‡ |
| | NaviAirway [29]* | **0.8074±0.0533**‡ | **0.8924±0.0440**‡ | 0.5994±0.1440‡ | 0.5148±0.1490‡ | 0.9247±0.0165‡ | 0.0714±0.0188‡ | 0.1345±0.0645‡ |
| | **U-Net+PatchGAN** | 0.7481±0.0678 | 0.8541±0.0464 | 0.7157±0.1067 | 0.6437±0.1195 | 0.8035±0.0775 | 0.2374±0.1286 | **0.0823±0.0332** |
| | **U-Net+ViT** | 0.7272±0.0631 | 0.8405±0.0436 | **0.7242±0.1096** | **0.6550±0.1266** | 0.7879±0.0816 | 0.2606±0.1401 | 0.0916±0.0325 |
| COVID | Wang et al. [10]† | 0.7433±0.1010‡ | 0.8481±0.0798‡ | 0.8487±0.1320‡ | 0.7993±0.1409‡ | 0.7749±0.0736‡ | 0.2843±0.1500‡ | 0.0541±0.1035‡ |
| | WingsNet [12]* | **0.9176±0.0363**‡ | **0.9566±0.0209**‡ | 0.9073±0.0647 | 0.8682±0.0831 | **0.9683±0.0204**‡ | **0.0311±0.0203**‡ | 0.0544±0.0283‡ |
| | NaviAirway [29]* | 0.8861±0.0298‡ | 0.9394±0.0171‡ | 0.8729±0.0792‡ | 0.8060±0.1034‡ | 0.9100±0.0245‡ | 0.0970±0.0316‡ | 0.0286±0.0239 |
| | **U-Net+PatchGAN** | 0.8292±0.0335 | 0.9063±0.0203 | 0.8911±0.0827 | 0.8443±0.1044 | 0.8506±0.0413 | 0.1741±0.0575 | **0.0281±0.0258** |
| | **U-Net+ViT** | 0.8090±0.0483 | 0.8936±0.0307 | **0.9104±0.0751** | **0.8772±0.0987** | 0.8354±0.0458 | 0.1933±0.0705 | 0.0378±0.0231 |

* refers to results obtained from open-source implementations with model weights provided.
† refers to reproduced results.
‡ represents statistical significance (with Wilcoxon signed-rank test $p < 0.01$) compared with the proposed method.
The bold values represent statistical significance (with Wilcoxon signed-rank test p < 0.01).

$$Dice = \frac{2XY}{X+Y} = \frac{2IoU}{IoU+1} \tag{10}$$

*Detected Length Ratio (DLR)* measures the proportion of detected branch length with respect to that of the ground truth:

$$DLR = \frac{L_X}{L_Y} \tag{11}$$

where $L_X$ is the total length of the correctly detected airway, $L_Y$ is the total length of the airway in the ground truth.

*Detected Branch Ratio (DBR)* measures the proportion of detected branch number with respect to that of the ground truth:

$$DBR = \frac{N_X}{N_Y} \tag{12}$$

where $N_X$ is the total number of correctly detected airway branches, $N_Y$ is the number of branches in the ground truth. In this study, branches with the intersection over union (IoU) score greater than 0.8 are referred to be correctly identified.

*Precision* refers to the fraction of correctly identified airway voxels among the predicted airway:

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

where $TP$ and $FP$ are the numbers of true positive voxels and false positive voxels.

*Leakage* measures the proportion of total false positives with respect to the ground truth annotations:

$$Leakage = \frac{V_X}{V_Y} \tag{14}$$

where $V_X$ is the volume of false-positive predictions, $V_Y$ is the volume of ground truth annotations.

*Airway Missing Ratio (AMR)* measures the proportion of total undetected airways (false negatives) with respect to the ground truth annotations:

$$AMR = \frac{FN}{V_Y} \tag{15}$$

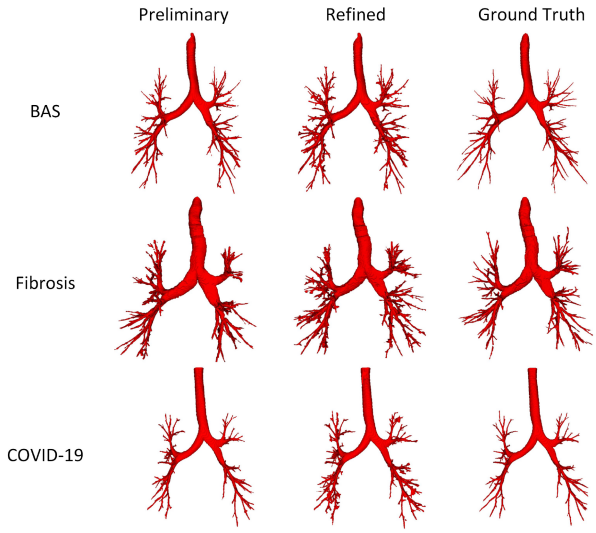


Fig. 3. Visualization of the preliminary segmentation, refined segmentation and the ground truth on the BAS dataset (CASE 02) as well as our in-house dataset (CASE 162_01 in fibrosis and CASE RM451 in COVID-19). While the overall structure appears similar, differences are mainly observed at a fine scale. Therefore, zoomed-in views are provided in Figs. 4 to 7.

where $FN$ is the volume of false-negative predictions, $V_Y$ is the volume of ground truth annotations.

### D. Segmentation Results

To comprehensively evaluate the performance of our model, we used seven evaluation metrics including intersection over union (IoU), dice coefficient, detected length ratio (DLR), detected branch ratio (DBR), precision, leakage and false negative rates (AMR). We also performed a Wilcoxon signed-rank test ($\alpha = 0.01$) for statistical analysis.

*Comparison on BAS dataset:* Our proposed method achieved state-of-the-art performance on BAS (Table II and Fig. 3 and 4) with 0.8902 DLR, 0.8439 DBR and 0.05441 AMR. Among all

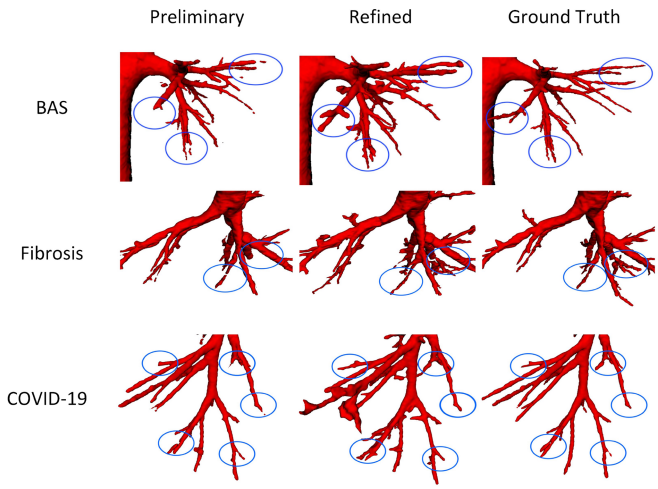


Fig. 4. Visualization of a zoomed-in region in the preliminary segmentation, refined segmentation, and ground truth on the BAS dataset (CASE 02) and our in-house datasets (CASE 162_01 in fibrosis and CASE RM451 in COVID-19).

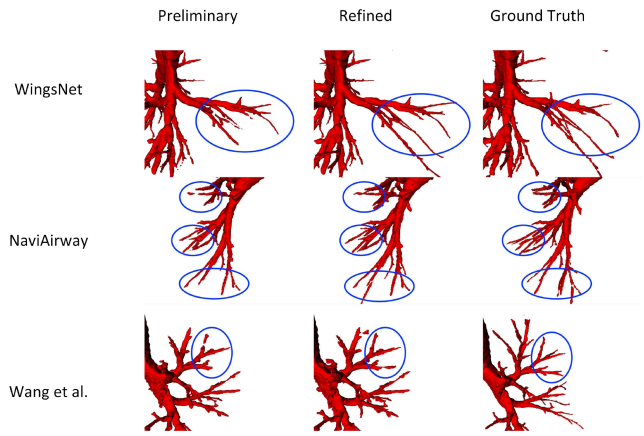


Fig. 6. Zoomed-in images show bronchioles of CASE 162_01 in the fibrosis test set, with the results from WingsNet, NaviAirway, and Wang et al. presented in rows from top to bottom, respectively. Columns from left to right display the initial segmentation of each model, the refined segmentation, and the ground truth. Blue circles are used to indicate any discontinuities in the preliminary segmentation, which were later resolved by our refinement model.

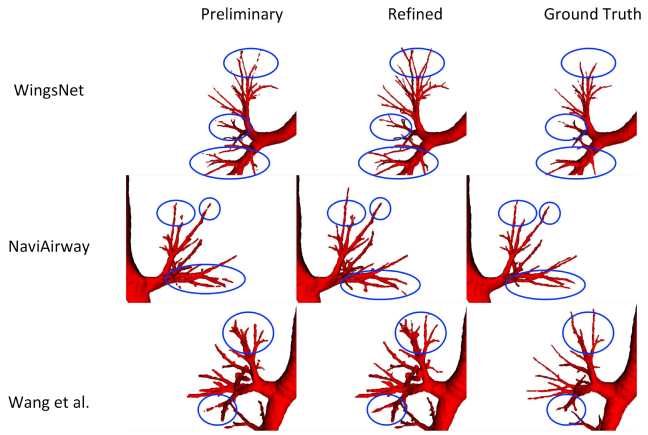


Fig. 5. Zoomed-in view of the bronchioles in CASE02 from the EXACT'09 test set. The top to bottom rows correspond to the results obtained by WingsNet, NaviAirway, and Wang et al. respectively. The left to right columns represent the initial segmentation of their models, the refined segmentation, and the ground truth. Blue circles highlight the breaks in the preliminary segmentation, which are later resolved by the refinement model.

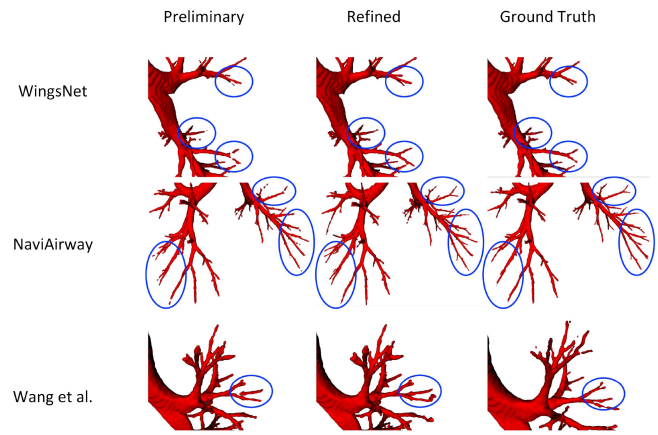


Fig. 7. Zoomed-in view of bronchioles in CASE RM451 from the COVID test set, with results from WingsNet, NaviAirway, and Wang et al. presented from top to bottom, respectively. The left-to-right columns show the initial segmentation of their models, the refined segmentation, and the ground truth. The blue circles highlight the areas of breakage in the preliminary segmentation, which were later corrected by the refinement model.

models, WingsNet achieved the highest intersection over union (IoU) of 0.8544 and voxel-wise precision of 0.9458. NaviAirway had the lowest ASD of 0.0413, and the model proposed by Wang et al. suffered from a high leakage rate of 0.3070. NaviAirway and the model proposed by Wang et al. also showed performance in DLR and DBR.

*Comparison on in-house fibrosis dataset:* All models performance decreases drastically in the fibrosis cases (Table II and Figs. 3 and 4). Notwithstanding, our model still achieves the highest DLR (0.7242), DBR (0.6550) and lowest AMR (0.08232). WingsNet and the model proposed by Wang et al. achieve similar results on DLR and DBR, but WingsNet has a better IoU score. NaviAirway obtains the highest IoU while unexpectedly achieving a poor performance on maintaining continuity with 0.5994 DLR and 0.5148 DBR.

*Comparison on in-house COVID-19 dataset:* All models perform slightly better compared to them on the BAS dataset (Table II and Figs. 3 and 4). The proposed method achieves the best DLR (0.9104), DBR (0.8772) and AMR (0.0281) in the COVID-19 cases. WingsNet achieves the second highest DLR (0.9073) and DBR (0.8682) followed by NaviAirway with 0.9394 DLR and 0.8729 DBR. WingsNet also achieves highest IoU (0.9176), Dice (0.9566) and precision (0.9683)

*Refined segmentation on other three models:* To demonstrate our refinement pipeline can be extended to other models, we refined all three other aforementioned models (WingsNet, NaviAirway and the model proposed by Wang et al.). Overall, all three models' performance on all three datasets (BAS, fibrosis and COVID-19) improved significantly with DLR and DBR increasing more than 10% on average (Table III). The IoU and

TABLE III
REFINED SEGMENTATION RESULTS

| | Model | IoU↓ | Dice↓ | DLR↑ | DBR↑ | Precision↓ | Leakages↑ | AMR↓ |
|---|---|---|---|---|---|---|---|---|
| Valid | Wang et al. [10]† | 0.7330±0.0786‡ | 0.84348±0.0554‡ | 0.8505±0.1227‡ | 0.7858±0.1420‡ | 0.7636±0.0737‡ | 0.3070±0.1412‡ | 0.0507±0.0619‡ |
| | Refined Wang et al. | 0.7136±0.1000 | 0.8284±0.0763 | 0.9093±0.1263 | 0.8700±0.1380 | 0.7341±0.0793 | 0.3552±0.1281 | 0.0423±0.0960 |
| | WingsNet [12]* | 0.8544±0.0673 | 0.9200±0.0422 | 0.8698±0.1175‡ | 0.8166±0.1305‡ | 0.9458±0.0271‡ | 0.0529±0.0300‡ | 0.1002±0.0769‡ |
| | Refined WingsNet | 0.8504±0.0651 | 0.9177±0.0409 | 0.9050±0.1032 | 0.8621±0.1230 | 0.9331±0.0303 | 0.0668±0.0347 | 0.0927±0.0752 |
| | NaviAirway [29]* | 0.8348±0.0335‡ | 0.9096±0.0200‡ | 0.8734±0.0715‡ | 0.8099±0.0950‡ | 0.8672±0.0406‡ | 0.1500±0.0536‡ | 0.0413±0.0304‡ |
| | Refined NaviAirway | 0.8246±0.0354 | 0.9034±0.0213 | 0.9064±0.0595 | 0.8590±0.0857 | 0.8516±0.0432 | 0.1716±0.0584 | 0.0357±0.0261 |
| Fibrosis | Wang et al. [10]† | 0.6979±0.0647‡ | 0.8203±0.0462‡ | 0.6961±0.0924‡ | 0.6261±0.1117‡ | 0.7468±0.0773‡ | 0.32727±0.1445‡ | 0.0822±0.0388 |
| | Refined Wang et al. | 0.6528±0.0925 | 0.7859±0.0722 | 0.8007±0.0847 | 0.7399±0.1042 | 0.6736±0.0951 | 0.4953±0.2456 | 0.0456±0.0277 |
| | WingsNet [12]* | 0.8052±0.0539‡ | 0.8910±0.0340‡ | 0.6951±0.0977‡ | 0.6198±0.1175‡ | 0.9505±0.0116‡ | 0.0438±0.0113‡ | 0.1595±0.0576‡ |
| | Refined WingsNet | 0.8035±0.0515 | 0.8901±0.0326 | 0.7318±0.1012‡ | 0.6638±0.1221‡ | 0.9341±0.0159 | 0.0604±0.0163 | 0.1480±0.0565 |
| | NaviAirway [29]* | 0.8074±0.0533‡ | 0.8924±0.0340‡ | 0.5994±0.1440‡ | 0.5148±0.1490‡ | 0.9247±0.0165‡ | 0.0714±0.0188‡ | 0.1345±0.0645‡ |
| | Refined NaviAirway | 0.8061±0.0514 | 0.8917±0.0329 | 0.6534±0.1513 | 0.5725±0.1617 | 0.9090±0.0189 | 0.0891±0.0226 | 0.1215±0.0657 |
| COVID | Wang et al. [10]† | 0.7433±0.1010 | 0.8481±0.0798 | 0.8487±0.1320‡ | 0.7993±0.1409‡ | 0.7749±0.0736‡ | 0.2843±0.1500‡ | 0.0541±0.1035‡ |
| | Refined Wang et al. | 0.7408±0.0566 | 0.8498±0.0392 | 0.9538±0.0435 | 0.9316±0.0603 | 0.7462±0.0562 | 0.3453±0.1152 | 0.0098±0.0106 |
| | WingsNet [12]* | 0.9176±0.0363‡ | 0.9566±0.0209‡ | 0.9073±0.0647‡ | 0.8682±0.0831‡ | 0.9683±0.0204‡ | 0.0311±0.0203‡ | 0.0544±0.0283‡ |
| | Refined WingsNet | 0.8971±0.0345 | 0.9454±0.0200 | 0.9571±0.0575 | 0.9409±0.0778 | 0.9424±0.0266 | 0.0588±0.0297 | 0.0510±0.0259 |
| | NaviAirway [29]* | 0.8861±0.0298‡ | 0.9394±0.0171‡ | 0.8729±0.0792‡ | 0.8060±0.1034‡ | 0.9100±0.0245‡ | 0.0970±0.0316‡ | 0.0286±0.0239‡ |
| | Refined NaviAirway | 0.8725±0.0315 | 0.9316±0.0186 | 0.9153±0.0724 | 0.8732±0.0997 | 0.8904±0.0295 | 0.1218±0.0407 | 0.0223±0.0214 |

* refers to results obtained from open-source implementations with model weights provided.
† refers to reproduced results.
‡ represents statistical significance (with Wilcoxon signed-rank test $p < 0.01$) compared with the refined results.

TABLE IV
ABLATION STUDY OF THE PROPOSED MODEL

| Model | IoU | Dice | DLR | DBR | Precision | Leakages | AMR |
|---|---|---|---|---|---|---|---|
| U-Net+PatchGAN+dilation(BL) | 0.8274±0.0563 | 0.9044±0.0355 | 0.7936±0.1303 | 0.7088±0.1557 | 0.8866±0.0405 | 0.1215±0.0504 | 0.0730±0.0622 |
| BL+clDice | 0.8325±0.0465 | 0.9079±0.0294 | 0.8059±0.1285 | 0.7324±0.1525 | 0.8895±0.0313 | 0.1178±0.0400 | 0.0693±0.0608 |
| BL+ccf | 0.7133±0.0854 | 0.8295±0.0631 | 0.8526±0.1148 | 0.7947±0.1474 | 0.7441±0.0880 | 0.3519±0.2151 | **0.0516±0.0541** |
| BL+ccf+multi-scale | 0.7701±0.0522 | 0.8691±0.0343 | 0.8420±0.1277 | 0.7788±0.1542 | 0.8101±0.0484 | 0.2266±0.0761 | 0.0573±0.0600 |
| BL+clDice+multi-scale | **0.8342±0.0670** | **0.9080±0.0427** | 0.8352±0.1156 | 0.7671±0.1421 | **0.8969±0.0478** | **0.1082±0.0578** | 0.0781±0.0565 |
| BL+ccf+ clDice + multi-scale | 0.8150±0.0519 | 0.8971±0.0330 | 0.8828±0.1012 | 0.8337±0.1263 | 0.8556±0.0393 | 0.1620±0.0522 | 0.0544±0.0525 |
| BL(ViT) + ccf + cl_dice + multi-scale | 0.8133±0.0518 | 0.8961±0.0334 | **0.8902±0.0967** | **0.8439±0.1261** | 0.8600±0.0401 | 0.1549±0.0539 | 0.0623±0.0505 |

The bold values represent statistical significance (with Wilcoxon signed-rank test p < 0.01).

Dice decrease a little as a trade-off in all scenarios with some even not considered to be statistically significant. The false negative ratio is also reduced significantly in all cases. Notably, the improvement on these three models is not as much as the refinement of our own preliminary results. Figs. 5 to 7 provide a zoom-in illustration of the effect of our refinement visually.

### E. Ablation Study

We conducted ablation studies to analyze the individual impact of each component in our refinement model. Table IV summarizes the results obtained by comparing 3-D U-Net + PatchGAN with dilation (BL) against six variants: BL + clDIce, BL + CCF, BL + CCF with multi-scale supervision, BL+clDice with multi-scale supervision, BL+clDice+CCF with multi-scale supervision, and BL (PatchGAN replaced with ViT)+clDice+CCF with multi-scale supervision. The adoption of cl-Dice led to a slight improvement of roughly 1.0% across all seven metrics. However, replacing cl-Dice with CCF resulted in a decline in IoU, Dice, and leakage metrics while the remaining four metrics improved significantly by over 5.0% on average. The incorporation of multi-scale supervision enabled the model guided by CCF and cl-Dice to perform better in all seven metrics. Ultimately, we achieved the best performance in DLR (0.8902) and DBR (0.8439) by combining CCF, cl-Dice, and multi-scale supervision with ViT in our model.

## V. DISCUSSION AND CONCLUSION

In this article, we proposed a novel adversarial-based refinement model using *tanh* as the final activation function and trained the network using objective functions clDice and CCF that focus on the continuity of the airway. The refinement model corrects breakage and adds missing branches in the preliminary segmentation generated by other networks such as U-Net.

Our model was evaluated on three different datasets: the BAS dataset, as well as two in-house datasets comprising 25 cases each of pulmonary fibrosis and COVID-19. Our refined method outperforms the preliminary approach, as shown in Tables I and II. Specifically, our method significantly reduces false negatives and increases the length and number of detected branches across all three test datasets. In addition, we use blue circles in Figs. 3 to 7 to highlight instances where our refinement successfully detects breakage and bronchioles missed in the preliminary segmentation. Ensuring airway segmentation completeness is crucial for clinical implementation, as it provides biomarkers for evaluating the severity of lung diseases, such as traction bronchiectasis [30] and airway tapering [31]. Our model maximizes completeness by maintaining local connectivity, making

it a desirable tool for clinical use. This is because disconnected branches are typically removed when retrieving the largest connected component for airway evaluation.

There are also limitations in our current approach. Firstly, our model is trained on a specific set of preliminary segmentation generated using a small U-Net. Hence, the improvement in the preliminary segmentation produced by the same network (small U-Net) is higher than in the other three models (Wang et al., WingsNet, and NaviAirway). Moreover, the quality of the preliminary results affects the model's ability to refine, as shown in Table III, where better initial results generally lead to better refinement. To overcome this limitation in the future, we plan to train the model using preliminary segmentation generated by randomly masking out ground truth. Secondly, while our approach maintains airway continuity, it may over-segment some peripheral bronchioles, which is reflected in the higher leakage. This issue is more noticeable in the fibrosis dataset than in the BAS and COVID-19 datasets. We also observed that in some cases, our predictions detected airway branches that were not shown in the ground truth, which could explain why IoU and Dice decrease slightly while leakage rises after refinement.

In conclusion, we have demonstrated that patch-scale discriminators can help improve airway segmentation in terms of better connectivity and lower false negative rates. This refinement pipeline can also be extended to other models and segmentation tasks in the future.

## REFERENCES


[1] 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 Trans. Med. Imag.*, vol. 22, no. 8, pp. 940–950, Aug. 2003.

[2] J. Tschirren, E. A. Hoffman, G. McLennan, and M. Sonka, "Intrathoracic airway trees: Segmentation and airway morphology analysis from low-dose CT scans," *IEEE Trans. Med. Imag.*, vol. 24, no. 12, pp. 1529–1539, Dec. 2005.

[3] P. Lo et al., "Extraction of airways from CT (EXACT'09)," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 2093–2107, Nov. 2012.

[4] O. Ronneberger, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[5] Özgün Çiçek et al., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.

[6] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[7] J.-P. Charbonnier et al., "Improving airway segmentation in computed tomography using leak detection with convolutional networks," *Med. Image Anal.*, vol. 36, pp. 52–60, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184151630202X

[8] D. Jin et al., "3D convolutional neural networks with graph refinement for airway segmentation using incomplete data labels," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2017, pp. 141–149.

[9] Q. Meng et al., "Tracking and segmentation of the airways in chest CT using a fully convolutional network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2017, pp. 198–207.

[10] C. Wang et al., "Tubular structure segmentation using spatial fully connected network with radial distance loss for 3D medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 348–356.

[11] 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 *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 809–813.

[12] H. Zheng et al., "Alleviating class-wisegradientimbalance for pulmonary airway segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2452–2462, Sep. 2021.

[13] F. Isensee et al., "nnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[14] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[15] W. Wang et al., "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 109–119.

[16] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[17] S. Li et al., "Medical image segmentation using squeeze-and-expansion transformers," in *Proc. 30th Int. Joint Conf. Artif. Intell.. Int. Joint Conf. Artif. Intell. Org.*, 8 2021, pp. 807–815.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[19] H. Zhang, M. Shen, P. L. Shah, and G. -Z. Yang, "Pathological airway segmentation with cascaded neural networks for bronchoscopic navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 9974–9980.

[20] H. Zhao et al., "High-quality retinal vessel segmentation using generative adversarial network with a large receptive field," *Int. J. Imag. Syst. Technol.*, vol. 30, pp. 828–842, 2020.

[21] X. Guo et al., "Retinal vessel segmentation combined with generative adversarial networks and dense U-net," *IEEE Access*, vol. 8, pp. 194551–194560, 2020.

[22] K.-B. Park, S. H. Choi, and J. Y. Lee, "M-GAN: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks," *IEEE Access*, vol. 8, pp. 146308–146322, 2020.

[23] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised Nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, 2015, pp. 562–570. [Online]. Available: https://proceedings.mlr.press/v38/lee15a.html

[24] A. Garcia-Uceda, H. A. W. M. Juarez Tiddens, and M. de Bruijne, "Automatic airway segmentation in chest CT using convolutional neural networks," in *Proc. Image Anal. Moving Organ, Breast, Thoracic Images*, 2018, pp. 238–250.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[26] S. Shit et al., "clDice - A novel topology-preserving loss function for tubular structure segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16560–16569.

[27] Y. Nan et al., "Fuzzy attention neural network to tackle discontinuity in airway segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 19, 2022. doi: 10.1109/TNNLS.2023.3269223.

[28] S. G. Armato et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[29] A. Wang et al., "NaviAirway: A bronchiole-sensitive deep learning-based airway segmentation pipeline," 2022. [Online]. Available: https://arxiv.org/abs/2203.04294

[30] S. L. Walsh et al., "Connective tissue disease related fibrotic lung disease: High resolution computed tomographic and pulmonary function indices as prognostic determinants," *Thorax*, vol. 69, no. 3, pp. 216–222, 2014. [Online]. Available: https://thorax.bmj.com/content/69/3/216

[31] W. Kuo et al., "Airway tapering: An objective image biomarker for bronchiectasis," *Eur. Radiol.*, vol. 30, pp. 2703–2711, 2020.