Medical Image Analysis (2025)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Skeleton2Mask: Skeleton-Supervised Airway Segmentation

Mingyue Zhao[a,b], Han Li[a,b], Di Zhang[c], Jin Zhang[c], Xiuxiu Zhou[c], Li Fan[c,*], Xiaolan Qiu[d,e], Shiyuan Liu[c], S. Kevin Zhou[a,b,f,*]

[a] *School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei Anhui 230026, China*
[b] *Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, University of Science and Technology of China (USTC), Suzhou Jiangsu 215123, China*
[c] *Department of Radiology, Second Affiliated Hospital of Navy Medical University, Shanghai 200003, China*
[d] *Suzhou Key Laboratory of Microwave Imaging, Processing and Application Technology, Suzhou Aerospace Information Research Institute, Suzhou 215123, China.*
[e] *Aerospace Information Research Institute, CAS, 100190, China.*
[f] *Key Laboratory of Precision and Intelligent Chemistry, USTC, Hefei Anhui 230026, China*

ARTICLE INFO

Article history:
Received 2025

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords:
Airway segmentation
Skeleton annotation
Sparsely-supervised learning

ABSTRACT

Airway segmentation has achieved considerable success. However, it still hinges on precise voxel-wise annotations, which are not only labor-intensive and time-consuming but also subject to challenges like missing branches, discontinuous branch labeling, and erroneous edge delineation. To tackle this, this paper introduces two novel contributions: a skeleton annotation (SKA) strategy for airway tree structures, and a sparse supervision learning approach- Skeleton2Mask, built upon SKA for dense airway prediction. The SKA strategy replaces traditional slice-by-slice, voxel-wise labeling with a branch-by-branch, control-point-based skeleton delineation. This approach not only enhances the preservation of topological integrity but also reduces annotation time by approximately 80%. Its effectiveness and reliability have been validated through **clinical experiments**, demonstrating its potential to streamline airway segmentation tasks. Nevertheless, the absolute sparsity of this annotation, along with the typical tree structure, can easily cause the failure of sparse supervision learning. To tackle this, we further propose Skeleton2Mask, a two-stage label propagation learning method, involving dual-stream buffer propagation and hierarchical geometry-aware learning, to ensure reliable and structure-friendly dense prediction. Experiments reveal that 1) Skeleton2Mask outperforms other sparsely supervised approaches on two public datasets by a large margin, achieving comparable results to full supervision with no more than 3% of airway annotations. 2) With the same annotation cost, our algorithm demonstrated significantly superior performance in both topological and voxel-wise metrics.

1. Introduction

Airway segmentation is fundamentally important for early diagnosis, treatment, and assessment of longitudinal lung diseases (De Jong et al., 2020; Xiao et al., 2020). Recently, deep learning-based algorithms (Qin et al., 2020, 2021; Yu et al., 2022; Zheng et al., 2021b; Wang et al., 2022a; Zheng et al., 2021a; Zhang and Gu, 2023; Zhang et al., 2021; Zhao et al., 2019; Nadeem et al., 2020; Guo et al., 2022; Selvan et al., 2020; Yun et al., 2019) have emerged as highly effective tools for extracting airways. Despite this, existing algorithms still confront significant challenges in obtaining high-quality voxel-level annotations for the airway, mainly attributed to the **complex inherent characteristics of the airway** (as shown in Fig. 1(a)): *1) extensive blurred surface/edge area.* Compared to typical convex organs with well-defined boundaries, the airway tree

*Corresponding authors: fanli0930@163.com, skevinzhou@ustc.edu.cn

contains a substantially higher proportion of ambiguous edge voxels. This increases the annotation burden, as more effort is required to delineate unclear and gradually fading boundaries. *2) numerous thin and dispersed peripheral branches.* These branches occupy a substantial proportion of the total airway branches, with the 0-4mm diameter branches accounting for 86% of the total branches, as reported in (Zheng et al., 2021b). Due to the large number and diverse orientations of these branches, annotators are required to constantly adjust their observation perspectives. These two factors make voxel-wise airway annotation a time-consuming and labor-intensive task. Furthermore, compared to the labeling of near-convex structures such as organs or tumors, it inevitably leads to more inconsistent and noisy annotations, such as missing branches, branch fragmentation, or erroneous edge annotations (as shown in Fig. 1(b)). Hence, there is an urgent need to propose a label-efficient solution for airway segmentation.

Learning from sparse annotation (Zhou et al., 2023a; Han et al., 2023; Zhang and Zhuang, 2022b; Liang et al., 2022; Zhang and Zhuang, 2022a; Chen and Hong, 2022; Lee et al., 2021; Wei and Ji, 2021; Zhou et al., 2023b; Xu et al., 2021; Zhang et al., 2023b; Cai et al., 2023; Wang et al., 2023a; Li et al., 2023a, 2022), which utilizes only a subset of annotated pixels, has demonstrated comparable performance to fully supervised learning both in natural and medical images. It shows great potential in difficult-to-annotate segmentation tasks, yet its application in tubular structures remains unexplored. Although some research (Xu et al., 2021; Zhang et al., 2023b) attempts to reduce the number of voxel-wise annotations to lessen the annotation burden, e.g., slice-level or branch-level partial annotation, they inherently struggle to ensure topological completeness due to their limited focus on individual slices or branches. Besides, the challenge of elaborately delineating fine yet dispersed branches of the airway tree still remain. To this end, we introduce a novel **SK**eleton **A**nnotation (**SKA**) strategy tailored to the airway which depicts the skeleton by marking control points along its topological structure. Note that although a skeleton typically refers to the centerline or medial axis of a shape, that is, a thin, connected set of lines or curves that run through the center of the object (Blum, 1967), we do not necessarily enforce such a criterion in practice. As shown later, it is empirically sufficient as long as our skeleton annotation is within the lumen.

As illustrated in Fig.1(c), the SKA shows three main advantages over the voxel-wise annotations. 1) *Reduced annotation workload.* SKA, focusing only on the skeleton and implemented in a point-wise manner, is simpler and more efficient, reducing numerous blurred edge annotations. Our experiments also confirm this via a user study. 2) *Enhanced annotation consistency and precision*. Voxels annotated in SKA tend to be closer to the lumen center, exhibiting more prominent airway characteristics, effectively enhancing annotation accuracy and consistency. 3) *Improved topology preservation*. It is more conducive to maintaining the complete topological structure of the airway, which is vital for accurate modeling and analysis. Its effectiveness and reliability have been validated through **clinical experiments**, demonstrating its potential to streamline airway segmentation tasks.

Nevertheless, relying solely on SKA to achieve reliable airway segmentation still poses significant challenges: 1) *Extremely sparse supervision*: Since only the skeleton is annotated, the supervision signal is too sparse to directly train a deep neural network. 2) *Limited diversity in supervision signal*: This issue can be observed in two main aspects. The first is the absence of edge annotations, which is crucial for defining precise boundaries in segmentation tasks. Secondly, due to the inherent nature of SKA, this labeling method exhibits a stronger preference (Wang et al., 2023b) for annotation positions (*i.e.*, closer to the lumen center) compared to other sparse annotations (*e.g.*, scribbles) that can cover a broader area with an accumulation of extensive annotated samples. This annotation bias implies less effective information and makes it challenging for airway segmentation training.

To address the above challenges, we propose **Skeleton2Mask**, a skeleton-supervised learning method tailored for SKA, aiming to achieve progressive label propagation for airway segmentation. The process begins with a **dual-stream buffer propagation strategy (DBP)** that integrates spatial and intensity information to initialize label propagation from SKAs, mitigating the risk of learning collapse. Subsequently, we introduce a **hierarchical geometry-aware learning (HGL)** framework that leverages the tree-like airway structure via three complementary components: topology-aware learning, hard geometry-aware propagation, and soft geometry-guided supervision. By combining SKA and Skeleton2Mask, our method reduces annotation costs by approximately 80% while achieving 98–99% of the topological accuracy compared to fully supervised learning, demonstrating performance on par with fully supervised learning.

To summarize, the contributions of our work are threefold: **1)** We propose a skeleton annotation strategy for tubular structures which is implemented with a control-point-based branch-by-branch annotation manner. Compared with full annotation, it saves about **80%** annotation time while preserving a more complete topology in clinical practice. **2)** A skeleton-supervised learning method, termed Skeleton2Mask, is proposed to progressively propagate the semantic knowledge from SKA to unlabeled regions, thus realizing airway extraction. **3)** Extensive experiments indicate that our method is not only superior to other sparse supervision learning methods but also achieves comparable performance to the fully-supervised method yet with **less than 3%** pixel-annotated information. Besides, with comparable annotation costs, our algorithm exhibits superior performance.

2. Related Literature

In this section, we briefly review the related literature, mainly focusing on label-efficient tubular structure segmentation and sparsely-supervised segmentation.

2.1. Label-Efficient Elongated Branching Structure Segmentation

Recently, an array of weakly supervised segmentation models have emerged, which operate with different label levels,

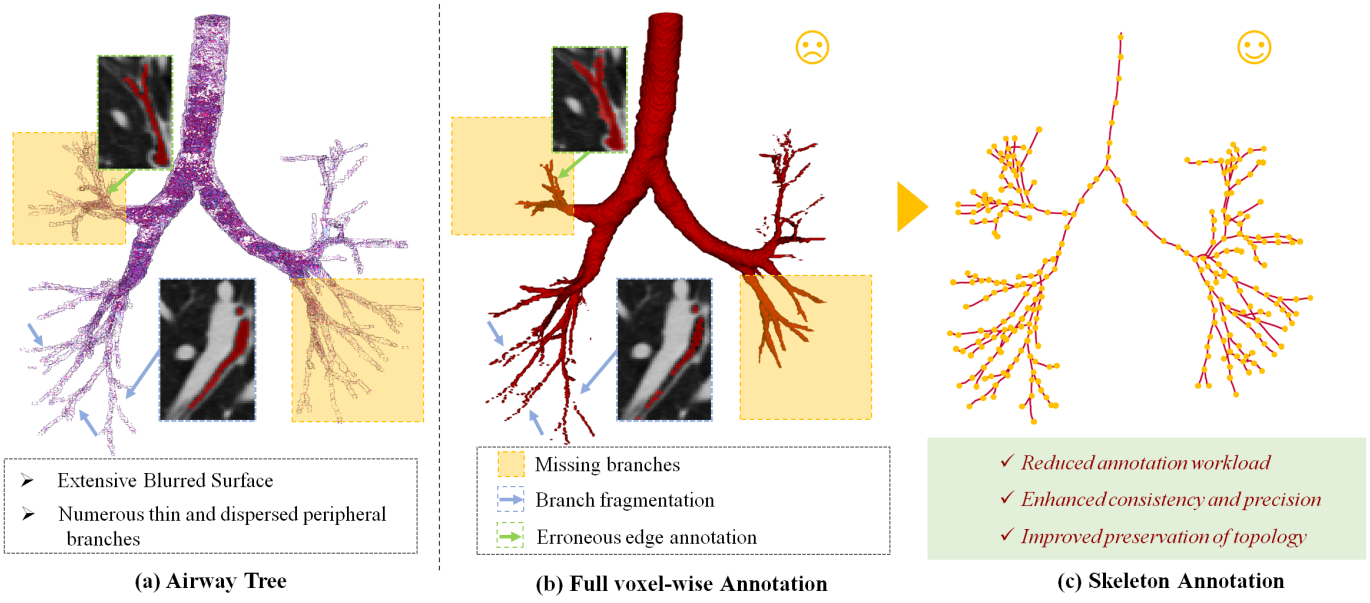

Fig. 1: **Motivation** for introducing the skeleton annotation for airway segmentation.

such as point-level (Kim et al., 2023; Li et al., 2023b; Wang and Bai, 2024), scribble-level (Zhou et al., 2023a; Han et al., 2023; Zhang and Zhuang, 2022b; Liang et al., 2022; Zhang and Zhuang, 2022a; Chen and Hong, 2022; Lee and Jeong, 2020; Luo et al., 2022; Wu et al., 2023), box-level (Lu et al., 2024; Wei et al., 2022), text-level (Xie et al., 2024) and image-level (Yin et al., 2024; Zhang et al., 2024). Despite that, most methods are tailored for convex or near-convex structures, differing fundamentally from continuous branching structures.

Label-efficient techniques for elongated branching structures, such as road, airway, or vessel, are relatively underexplored due to their inherent structural properties including topological connectivity, dispersed branching patterns, and both intra-class and inter-class imbalance. Koziński *et al.* (Koziński et al., 2020) proposed a projection-supervision strategy that enables training 3D segmentation networks using only 2D Maximum Intensity Projections (MIP) annotations. By introducing a projection-based loss inspired by space carving, their method significantly reduces annotation time by approximately 50% without compromising segmentation performance. Similarly, Xu *et al.* (Xu et al., 2021) achieved efficient vessel segmentation via local 2D patch annotations, and Dang *et al.* (Dang et al., 2022) reduced annotation burden by leveraging weak patch-level labels instead of pixel-wise supervision. In addition, Zhang *et al.* (Zhang et al., 2023b) utilized partial branch annotations combined with local-to-global label propagation for weakly supervised vessel tree delineation.

Although these approaches have alleviated the burdens of full annotation to some extent, they still face two major challenges: 1) unavoidable annotation noise at edges and 2) the inability to ensure the topological integrity of annotations. The topological integrity of elongated structures is crucial for accurate structural quantitative evaluation and bronchoscopy navigation. The delicate distal branches often indicate early pathological changes, making them crucial for the early detection and prevention of diseases. To tackle this, we present a topological-level annotation strategy, focusing on a top-down, control-point-based annotation of elongated branching structures. It substantially alleviates the annotation burden of elongated structures compared to traditional pixel-level comprehensive annotation, thereby simplifying the annotation workflow.

2.2. *Sparsely-supervised Segmentation*

Existing sparse supervision learning methods (Kim et al., 2023; Li et al., 2023b; Wang and Bai, 2024; Wang et al., 2023b; Wei and Ji, 2021; Chen and Hong, 2022; Lee et al., 2021; Zhang and Zhuang, 2022a,b; Zhou et al., 2023b; Luo et al., 2022; Lee and Jeong, 2020; Liang et al., 2022; Wu et al., 2023; Zhou et al., 2023a; Han et al., 2023) focus on establishing a reliable label propagation pathway from sparse annotations to achieve dense prediction tasks, which are primarily based on *regularization loss, auxiliary task, consistency learning, and dynamic pseudo-label learning*. The regularization methods (Grandvalet and Bengio, 2004; Tang et al., 2018) encourage the model to produce more robust segmentation by imposing regularization constraints on the prediction results. Auxiliary task methods (Wei and Ji, 2021; Chen and Hong, 2022) incorporate auxiliary task branches, such as edge detection branches, into the original segmentation framework to improve the model's edge awareness in unannotated regions. Consistency learning methods (Zhang and Zhuang, 2022a,b; Zhou et al., 2023b) typically apply different image augmentations to the original input image and construct consistency losses as constraints for the model to maintain segmentation consistency before and after augmentations. In addition, dynamic pseudo-label learning, as the most popular approach, leverages techniques such as model ensembling (Luo et al., 2022; Lee and Jeong, 2020; Zhou et al., 2023b), tree filter (Liang et al., 2022), Gaussian mixture model (Wu et al., 2023), and Bayesian inference (Zhou et al., 2023a) to achieve

reliable dynamic pseudo-label updates, thereby continuously refining the algorithm's segmentation performance.

Although existing sparse supervision methods have achieved performance comparable to fully supervised approaches in both natural and medical imaging domains, they are predominantly tailored to convex structures. When applied to airway segmentation, these methods are susceptible to the absolute sparsity of annotations or the dendritic structure of airway, potentially leading to a significant detection loss of distal branches or even complete training collapse. Thus, our Skeleton2Mask strategy is designed according to the properties of annotations and targets to propagate knowledge from sparsely annotated voxels to unannotated voxels effectively.

2.3. Hierarchical Geometry Learning

Hierarchical geometry learning is a structure-aware representation learning paradigm that aims to extract and organize geometric information from visual data across multiple levels of abstraction. It encompasses a diverse range of representations—such as contours, surfaces, centroids, skeletons, and connectivity graphs. It broadly supports structure-aware tasks by hierarchically encoding spatial semantics, geometric consistency, and anatomical plausibility. It has been widely applied in medical image segmentation (Zhang et al., 2022; Wang and Bai, 2024; Wang et al., 2022b; Cui et al., 2021) and 3D scene understanding (Liu et al., 2022; Wong and Vong, 2021; Yao et al., 2023), particularly for objects with complex and highly-structured objects.

For example, Zhang *et al.* (Zhang et al., 2022) introduced a hierarchical topology learning framework to jointly predict keypoints, centerlines, and voxel connectivity for vascular segmentation. Cui *et al.* (Cui et al., 2021) proposed a morphology-guided hierarchical model for dental instance segmentation, capturing tooth-level geometry from centroids to root landmarks. In natural scene understanding, Liu *et al.* (Liu et al., 2022) embedded persistent homology into point cloud segmentation networks to enforce topological correctness via differentiable topology-aware losses.

While prior explorations have provided valuable insights, most existing approaches still rely on geometric elements explicitly extracted from ground truth annotations. However, such annotations are typically unavailable in sparsely supervised learning paradigms. Moreover, these geometric elements are often sparse in nature, which limits their effectiveness in enhancing dense prediction tasks like semantic segmentation. In this work, we propose a hierarchical geometry-aware learning framework tailored for sparse supervision. On one hand, we exploit the topological characteristics inherently encoded in skeleton annotations to guide the model toward topologically consistent predictions. On the other hand, we leverage the semantic and positional priors of skeleton points by introducing a Gaussian geodesic distance transformation, which transforms discrete skeletons into continuous spatial cues. This dense structural encoding provides fine-grained guidance and effectively compensates for the lack of scale information inherent in sparse skeleton supervision.

3. Skeleton Annotation Strategy

Distal bronchi often indicate early structural changes associated with lung diseases. Nevertheless, full Voxel-wise Annotation (VA) is prone to overlooking distal branches or producing noisy edge annotations, especially when the available annotation time is limited. To address this, we propose the skeleton annotation (SKA) strategy, which prioritizes topological integrity and minimizes annotation bias across multi-scale branches, thereby maximizing the coverage of distal branches, as visually demonstrated in Fig. 2. Instead of voxel-wise annotation, SKA implements **control-point-wise** labeling in a **branch-by-branch** manner. Specifically, given a 3D CT image and its corresponding views in three different orientations, the annotator can flexibly perform top-down, branch-by-branch skeleton annotation on any appropriate view, until all branches are comprehensively covered. It does not involve any background marking or boundary prompts. The final result of the skeleton annotation can be exported in a structured text format. Fig. 3 illustrates the data structure of the annotation, where each curve (*i.e.,* anatomical branch) is represented by a unique identifier along with an ordered list of physical coordinates corresponding to its control points. With this structured skeleton annotation, we can: 1) infer the topological structure of the airway tree, including the orientation of each branch, the inter-branch relationship and the topological role of each control point (starting point, endpoint, or bifurcation point); 2) when combined with the airway segmentation results, enables downstream tasks such as airway quantification and hierarchical classification, which are essential for the diagnosis and prognosis of respiratory diseases.

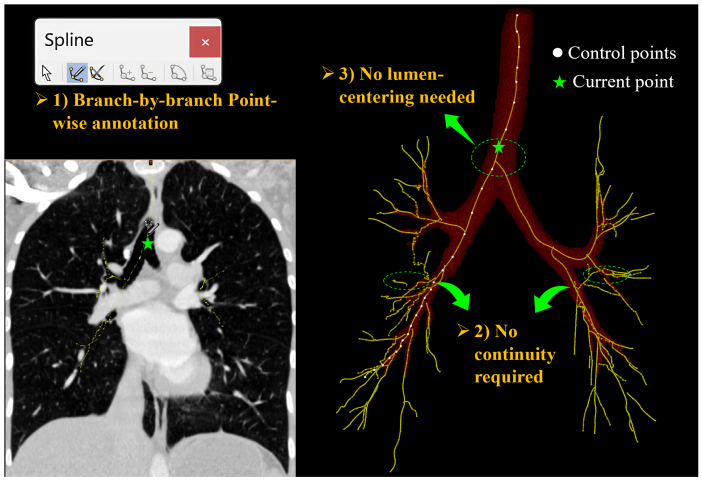

Fig. 2: **Skeleton annotation conducted with MIMICs software.**

Moreover, it is noteworthy that:

1) SKA is implemented on a per-point basis. Each branch is annotated as curves composed of control points rather than on a per-pixel basis. This branch-wise annotation can be efficiently achieved using curve annotation tools available in med-

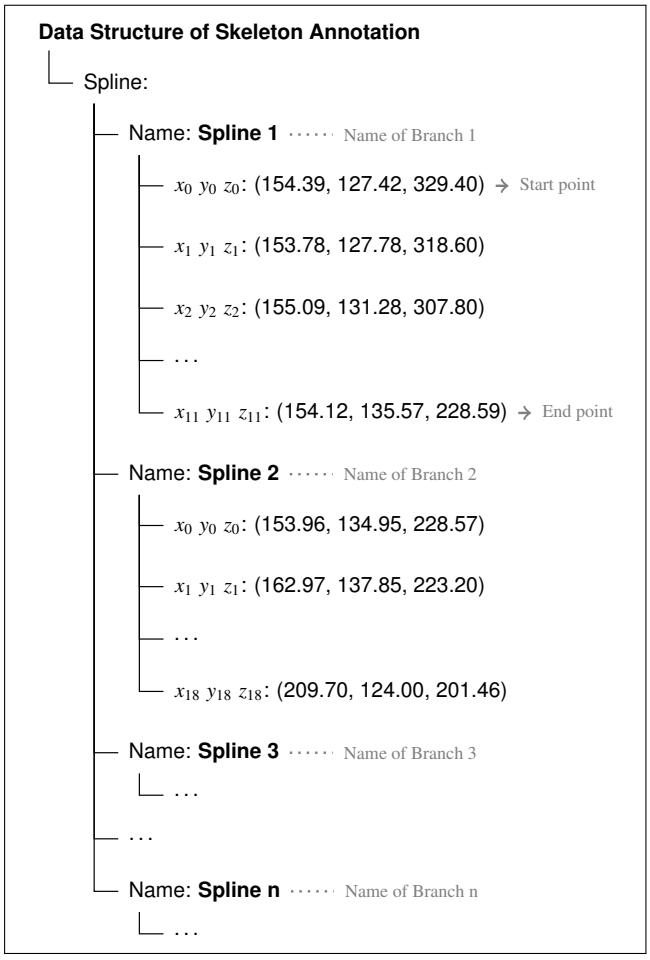

Fig. 3: **Schematic diagram of the data structure for skeleton annotation results.**

ical imaging software, such as MIMICs [1], Simvascular [2], and 3D Slicer [3]. Given the variability in airway branching patterns across individuals, the number of control points is adaptive, with annotation density increasing in proportion to the local branch curvature.

2) The continuity of branch annotations is not required. The software previously mentioned can automatically fit curves between control points using spline interpolation algorithms, thereby forming continuous branch annotations, which are displayed in real time within the software interface, as illustrated in Fig. 2. This significantly reduces the cost of 3D annotations.

3) Strictly positioned at the lumen center is not required. In practice, it is sufficient for the annotator to ensure that the added control points and the corresponding fitted curve lie within the lumen, which avoid the noisy annotations often associated with edge-based labeling.

[1]https://www.materialise.com/en/healthcare/mimics-innovation-suite/mimics

[2]https://simvascular.github.io/

[3]https://www.slicer.org/

4. Skeleton2Mask

As shown in Fig. 4, the Skeleton2Mask algorithm, supervised by SKA, employs a two-stage progressive propagation learning strategy to achieve airway mask prediction, including: 1) Dual-stream buffer propagation for initial label propagation; and 2) Hierarchical geometry-aware propagation learning for further dense prediction. Suppose a CT volume $\boldsymbol{x}$ with the corresponding SKA $\boldsymbol{y_s}$, Skeleton2Mask aims to learn a mapping $\mathcal{F} : \boldsymbol{x} \mapsto \hat{\boldsymbol{y}}$ based on SKA, where $\hat{\boldsymbol{y}}$ is the predicted probability map. Given any voxel $i \in \Omega$, where Ω represents the set of all voxels in the image domain, we define the annotated skeleton voxel set as $\Omega_s = \{i \mid \boldsymbol{y_s}(i) = 1\}$, containing only the sparsely annotated voxels from the SKA.

4.1. Dual-stream Buffer Propagation

To address the absolute sparsity of the SKA, we propose Dual-stream Buffer Propagation (DBP) to propagate the knowledge from SKA to more unlabeled voxels, by fully exploiting the spatial and gradient cues. As shown in Fig. 4, it consists of Gaussian Geodesic Distance Buffer Propagation (G^2BP) and Euclidean Distance Buffer Propagation (EBP).

4.1.1. Gaussian Geodesic Distance Buffer Propagation

In practice, we observe that 1) SKA tends to be located closer to the center of the lumen rather than the periphery. We define the center offset rate as the ratio between the distance from each control point to the airway centerline and the local lumen radius. Our clinical annotation analysis reveals a mean center offset rate of 24.8%, suggesting a clear spatial bias of the SKA. Points close to the SKA likely belong to the airway foreground, while points far from the SKA likely lie in the background. Moreover, 2) SKA is generally located in regions where foreground features are prominent, further reinforcing the feature cues of the SKA.

Inspired by the above two advantageous characteristics, we propose a Dual-stream Buffer Propagation strategy to generate the initial mask proposals. Geodesic distance, as a metric reference, considers the pixels' spatial distance relationship and their appearance similarity, and is therefore commonly used as a low-level feature to assist segmentation (Wang et al., 2018; Yu et al., 2022). Naturally, leveraging image information combined with the aforementioned SKA characteristics, geodesic distance representation effectively facilitates initial label propagation, even in the presence of severe intra-scale imbalance in the trachea. However, the geodesic distance-based label propagation relies on a strong prerequisite: intra-class pixels have similar grayscale values and show flat gradients. To achieve this, Gaussian geodesic distance computation, integrating Gaussian smoothing and geodesic distance calculation, is applied to enhance label consistency across heterogeneous regions, thus establishing a solid prerequisite for reliable label

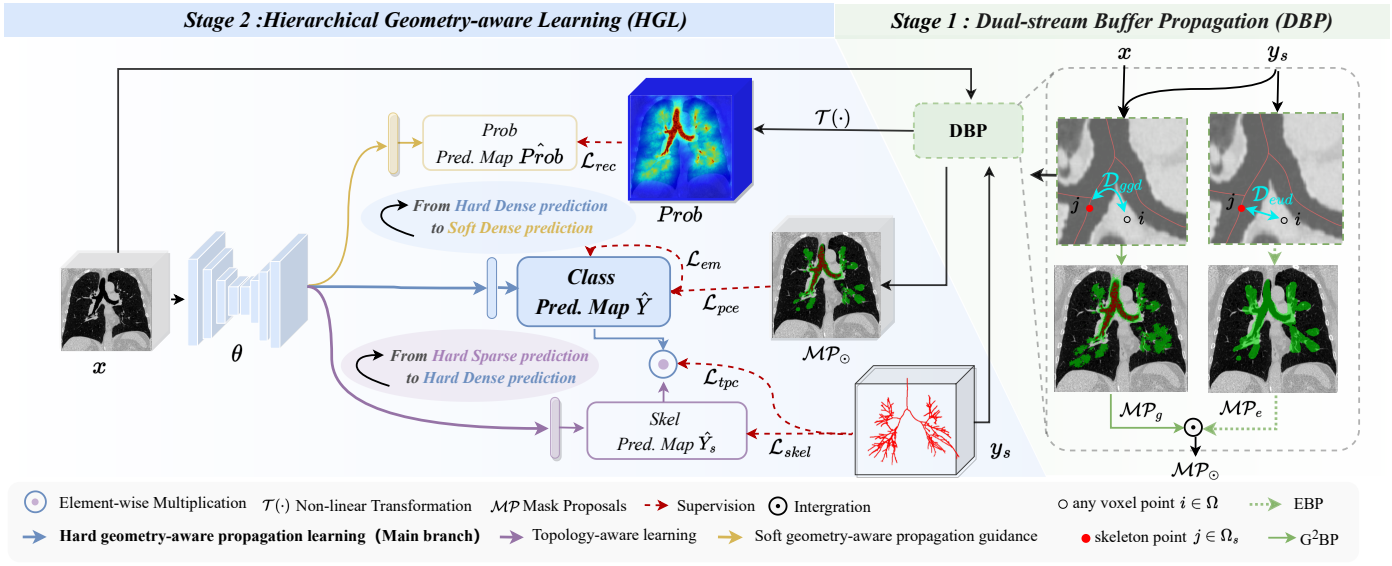

Fig. 4: **Overview of skeleton-supervised airway segmentation method - Skeleton2Mask. Training:** In stage 1, starting from CT images $\boldsymbol{x}$ and skeleton annotations $\boldsymbol{y}_s$, we employ a dual-buffer inference strategy to obtain initial label propagation, *i.e.,* the mask proposal $\mathcal{MP}_{\odot}$, where red highlights the foreground, green marks uncertain regions, while the rest is treated as background region. In stage 2, the mask proposal serves as the supervision for the main segmentation branch. By introducing topology-aware learning and soft geometry-aware propagation guidance as auxiliary tasks, a hierarchical geometric-aware learning paradigm is formed, facilitating accurate and dense airway predictions. **Inference:** Only the main segmentation branch within the HGL framework (*i.e.,*$\boldsymbol{x} \rightarrow \hat{\boldsymbol{y}}$) is utilized to generate airway predictions.

propagation. Specifically, the 3D image is represented as a graph $G = (V, E, w)$, where V represents the voxels, E is the set of edges connecting each voxel to its 26 neighbors, w is the edge weight defined as the mean intensity of the two connected voxels. The Gaussian geodesic distance $\mathcal{D}_{ggd}$ from each voxel $i \in \Omega$ in the image domain to each annotated skeleton voxel $j \in \Omega_s$ is formulated as:

$$\mathcal{D}_{ggd}[i, j, \boldsymbol{x}] := \mathcal{D}_{ged}[i, j, g_\sigma(\boldsymbol{x})], \tag{1}$$

where the Gaussian filtering $g_\sigma(\cdot)$ with a standard deviation of σ smooths the intensity variations to reduce noise interference. The geodesic distance $\mathcal{D}_{ged}$ between voxel i and voxel j in image m is given by:

$$\mathcal{D}_{ged}[i, j, m] := \min_{p \in \mathcal{P}_{i \to j}} \int_0^1 \|\nabla m(p(\xi)) \cdot \mathbf{u}(\xi)\| \, d\xi, \tag{2}$$

where $\mathcal{P}_{i \to j}$ denotes the set of all possible paths from voxel i to voxel j, and $p \in \mathcal{P}_{i \to j}$ is a feasible path parameterized by $\xi \in [0, 1]$. The vector $\mathbf{u}(\xi)$ is a unit tangent vector along the direction of the path p. Thus, $\mathcal{D}_{ged}$, obtained by Fast Marching algorithm (Sethian, 1999), as a path integral ensures that the computed geodesic distance captures both local intensity gradients and global topological structures. By incorporating the Gaussian filtering, $\mathcal{D}_{ggd}$achieves better robustness against noise and ensures a smooth yet topology-preserving geodesic metric.

To facilitate the label expansion of the foreground and background, we exploit bilateral G^{2}BI. Let Ω_{f_1} and Ω_{b_1} denote the foreground and background, respectively, then:

$$\Omega_{f_1} := \left\{ i | i \in \Omega \ \& \ \mathcal{D}_{ggd}(i) < \delta_1 max(\mathcal{D}_{ggd}) \right\}, \tag{3}$$

$$\Omega_{b_1} := \left\{ i | i \in \Omega \ \& \ \mathcal{D}_{ggd}(i) > \delta_2 max(\mathcal{D}_{ggd}) \right\}, \tag{4}$$

where $\delta_1, \delta_2 \in [0, 1]$ are hyper-parameters indicating the degree of expansion and $\delta_1 < \delta_2$. As shown in Fig. 4, the generated mask proposal $\mathcal{MP}_g(i)$ takes the value of 1 if $i \in \Omega_{f_1}$, 0 if $i \in \Omega_{b_1}$, and remains unknown otherwise.

4.1.2. Euclidean Distance Buffer Propagation

The tubular structure of the airways and the positional characteristics of the annotations naturally lead us to use a predefined diameter to compensate for the potentially insufficient label diffusion in G^{2}BP. Therefore, we further introduce Euclidean distance, an absolute spatial metric, based on the scale characteristics of the airways, thereby establishing an additional constraint for label propagation. Yet, with regard to the foreground, the narrowest airway branches might be merely a single pixel wide, leading to a high risk of label leakage for further diffusion using Euclidean distance. Consequently, we implement unilateral EBP to achieve background expansion:

$$\Omega_{b_2} := \{ i | i \in \Omega - \Omega_s \ \& \ \mathcal{D}_{eud}(i) > \gamma max(\mathcal{D}_{eud}) \}, \tag{5}$$

where $\mathcal{D}_{eud}(i) = \min_{j \in \Omega_s} \|i - j\|_2$ and $\gamma \in [0, 1]$ is a hyper-parameter to control the degree of background expansion. Similarly, as depicted in Fig. 4, mask proposal obtained by EBP $\mathcal{MP}_e$ is set to 0 within the region Ω_{b_2} and remains unknown in the rest of the area.

Thus far, the foreground and background after propagation can be represented as $\Omega_B = \Omega_{b_1} \cup \Omega_{b_2}$ and $\Omega_F = \Omega_{f_1}$, respectively. Fig. 4 gives an example of the final mask proposal $\mathcal{MP}_{\odot}$, which takes the value of 1 if $i \in \Omega_F$, 0 if $i \in \Omega_B$, and remains unknown otherwise. For simplicity, we denote the set of labeled voxels as $\Omega_L = \Omega_B \cup \Omega_F$ and the voxel set with unknown labels as $\Omega_U = \Omega - \Omega_L$.

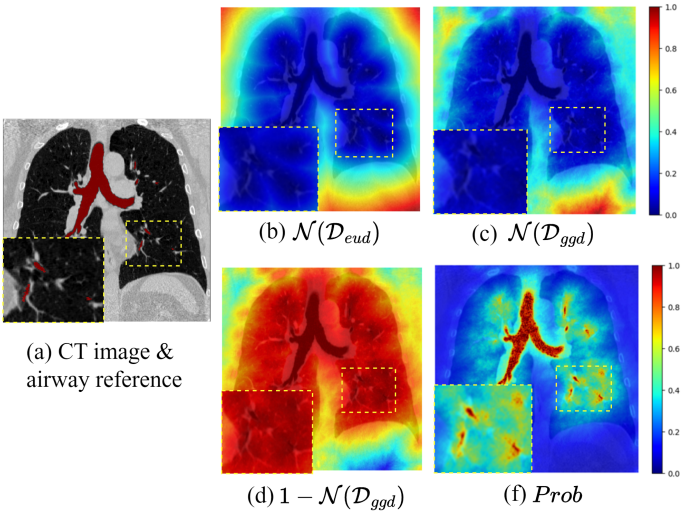

Fig. 5: **Visualization of intermediate results in label propagation.** (a) is the original CT image overlaid with airway reference. (b)-(f) are intermediate maps derived from distance-based computations. To enhance contrast and comparability across intermediate results, we apply min-max normalization, denoted as $\mathcal{N}(\cdot)$. Insets highlight regions with thin or peripheral branches to illustrate the effectiveness of different intermediate steps. Warmer colors (e.g., red) indicate higher values, while cooler colors (e.g., blue) represent lower values.

4.2. Hierarchical Geometry-aware propagation Learning

DBP significantly mitigates the challenge posed by sparse annotation, and on this foundation, Hierarchical Geometry-aware Learning (HGL) framework is incorporated to facilitate more accurate mask prediction. As illustrated in Fig. 4, it is performed in a triple-head manner to improve hierarchical airway structure representations learning at three levels, *i.e.,* topology-aware learning, hard geometry-aware propagation learning, and soft geometry-aware propagation guidance. The **hard geometry-aware propagation learning** acting as the main branch colored by blue is dedicated to voxel-wise airway prediction, supervised by the proposal masks $\mathcal{MP}_{\odot}$. The **topology-aware learning** (colored by purple) leverages the topological integrity of skeleton annotations to enhance the model's topology-aware learning while simultaneously improving the topological consistency learning of the primary segmentation branch. By utilizing fine-grained geometric perception for pixel-level probabilistic modeling, **soft geometry-aware propagation guidance** (colored by orange) enhances the structural granularity awareness of the model, thus resulting in superior segmentation performance.

4.2.1. Hard Geometry-aware Propagation Learning

The dual-stream buffer propagation strategy in stage one achieves initial annotation expansion, effectively avoiding the model learning collapse caused by the absolute sparsity of skeleton annotations. Similarly to (Chen and Hong, 2022; Zhou et al., 2023b; Wei and Ji, 2021; Zhai et al., 2023), we impose the partial cross-entropy loss solely on the labeled pixels (*i.e.*, $i \in \Omega_L$):

$$\mathcal{L}_{pce} = -\frac{1}{|\Omega_L|}\sum_{i\in\Omega_L}(\hat{\boldsymbol{y}}(i)log(\mathcal{MP}_{\odot}(i)) + (1-\hat{\boldsymbol{y}}(i))log(1-\mathcal{MP}_{\odot}(i))). \tag{6}$$

Besides, to encourage the model to produce high-confidence predictions for unannotated voxels, entropy minimization (Grandvalet and Bengio, 2004) is applied to the class probability prediction maps, with the corresponding supervised loss formulated as:

$$\mathcal{L}_{em} = -\frac{1}{|\Omega_U|}\sum_{i\in\Omega_U}\hat{\boldsymbol{y}}(i)log(\hat{\boldsymbol{y}}(i)). \tag{7}$$

4.2.2. Topology-aware Learning

Skeleton-level annotation debiases the annotation scale while preserving excellent topological integrity. Inspired by this, we introduce topology-aware learning. On one hand, it enhances the model's structural perception along the branching directions by incorporating an auxiliary skeleton segmentation task supervised by $\mathcal{L}_{skel}$. On the other hand, it aims to enhance the topological completeness of airway segmentation by introducing topological consistency learning, *i.e.*, $\mathcal{L}_{tpc}$. These components collectively define the overall topology-aware learning loss as follows:

$$\mathcal{L}_{topo} = \underbrace{Focal(\hat{\boldsymbol{y}}_s(i), \boldsymbol{y}_s(i))}_{\mathcal{L}_{skel}} + \underbrace{Focal(\hat{\boldsymbol{y}}(i) * \hat{\boldsymbol{y}}_s(i), \boldsymbol{y}_s(i))}_{\mathcal{L}_{tpc}}, \tag{8}$$

where $Focal$ represents the Focal loss (Ross and Dollár, 2017), applied to mitigate the severe inter-class imbalance caused by skeleton sparsity. Here, $i \in \Omega$ and $\hat{\boldsymbol{y}}_s$ denotes the skeleton prediction.

4.2.3. Soft Geometry-aware Propagation Guidance

Despite that initial label propagation from SKAs partially alleviates the problem of sparse supervision, the model still lacks the fine-grained structural perception of the airway. In light of this, we are considering *whether a probabilistic modeling approach can be introduced to provide fine-grained learning guidance for the model's predictions.* To this end, we first visualize $\mathcal{D}_{eud}$ and $\mathcal{D}_{ggd}$ in Fig. 5, as introduced in Sec. 4.1, in an effort to uncover potential insights. Encouragingly, the Gaussian geodesic distance map $\mathcal{D}_{ggd}$ introduced aligns with the multi-scale characteristics of the airway tree and the structural properties of SKAs, comprehensively considering the spatial and geodesic distances of each unannotated pixel $i \in \Omega \setminus \Omega_s$ to the nearest skeleton annotated pixel $j \in \Omega_s$, thereby providing fine-grained semantic guidance. The naive approach to probabilistic modeling is $Prob(i) = \mathbf{1} - \mathcal{N}(\mathcal{D}_{ggd}(i))$, in which $\mathcal{N}$ denotes min-max normalization. Nevertheless, as illustrated in Fig.5(d), the weakening of edge gradients in distal branches results in significant similarity activation between the lung parenchyma and airway lumen, which is detrimental to learning distinctive foreground features. Here, we ingeniously utilize nonlinear transformations $\mathcal{T}(\cdot)$ to construct effective probabilistic modeling, accentuating inter-class differences while avoiding over-segmentation. It can be formulated as:

$$Prob(i) = \mathcal{T}(\mathcal{D}_{ggd}(i)) = \frac{\mathbf{1}}{\mathcal{N}(\mathcal{D}_{ggd}(i)) + c}, \tag{9}$$

where $c = 1$ ensures $Prob \in [0, 1]$ while avoiding division by zero. Compared with Fig. 5(d) and Fig.5(e), the inverse transformation significantly enhances the airway branches within the

lung, thereby providing fine-grained and robust auxiliary information for branch segmentation. Such probabilistic distribution learning can be viewed as an augmented reconstruction task, consequently, Mean Squared Error (MSE) loss is utilized as the supervision loss function.

$$\mathcal{L}_{rec} = \frac{1}{n} \left\| Prob(i) - \hat{Prob}(i) \right\|_2^2, \tag{10}$$

where $i \in \Omega$, $\hat{Prob}$ represents the probability prediction map.

4.2.4. Overall Framework

Taking into account the commonalities and specificities among the three tasks in the HGL architecture, we designed a customized triple-head network to achieve hierarchical Geometry-aware Propagation Learning. Specifically, as depicted in Fig. 4, a modified 3D U-Net was employed as the backbone to extract task-shared latent representations $f_\theta = f(X;\theta)$. Each ConvBlock in the encoder and decoder stage is composed of two convolutional layers followed by Instance Normalization (Ulyanov, 2016) and ReLU activation function. Upon this, additional ConvBlocks are introduced as task-dependent layers, followed by a linear projector as the supervision head. Concretely, the prediction outputs of three branches can be formulated as:

$$\hat{y}_s = Sigmoid(Conv1(f_\theta)), \tag{11}$$

$$\hat{Prob} = Conv1(ConvBlock1(f_\theta)), \tag{12}$$

$$\hat{y} = Sigmoid(Conv1(ConvBlock2(ConvBlock1(f_\theta)))), \tag{13}$$

where Conv1 denotes $1 \times 1 \times 1$ convolution. ConvBlock1 and ConvBlock2 share the same structure, comprising $3 \times 3 \times 3$ ConvLayers with Instance Normalization and activation layers. $Sigmoid$ means the Sigmoid function. The skeleton-level prediction $\hat{y}_s$ provides topological cues for the subsequent two tasks, while the probability modeling output $\hat{Prob}$ offers fine-grained, multi-scale information guidance to the primary segmentation branch. Together, these components complement the segmentation branch to establish a progressive label propagation framework, which can be trained by minimizing:

$$\mathcal{L}_{total} = \mathcal{L}_{pce} + \lambda_1 \mathcal{L}_{em} + \lambda_2 \mathcal{L}_{rec} + \lambda_3 \mathcal{L}_{topo}, \tag{14}$$

where $\lambda_k, k = 1, 2, 3$ are the trade-off weights. Thus far, this hierarchical multi-granularity structure-aware learning fully accommodates the characteristics of the target and annotations, thereby achieving precise skeleton-to-mask propagation.

5. Experiments

5.1. Datasets

Experiments are implemented on the public Binary Airway Segmentation (BAS) Dataset (Qin et al., 2020), consisting of 90 CT scans and ATM22 Dataset (Zhang et al., 2023a) consisting of 299 CT scans. The BAS dataset consists of 70 CT scans from the LIDC dataset (Armato III et al., 2011) and 20 from the EXACT09 dataset (Lo et al., 2012). The in-plane spacing of all images ranges from 0.50 to 0.82 mm, with slice thicknesses varying between 0.45 and 1.80 mm. The ATM22 dataset includes 300 training cases released by the ATM22 Challenge [4], of which one case was excluded following the organizer's notification due to a mismatch between the image and its corresponding label. The in-plane spacing of all the CT volumes ranges from 0.51 to 0.92 mm, and slice thicknesses span 0.50 to 5.00 mm. To ensure the reliability of experimental results, both datasets were randomly shuffled and split into training and testing sets at a ratio of 4:1. Specifically, the BAS dataset was divided into 72 training cases and 18 testing cases, while the ATM22 dataset was split into 239 training cases and 60 testing cases. Since all datasets included fully annotated masks, we opted to use simulated skeleton annotations for ease of implementation in our experimental setup. The simulated SKA annotations were generated by extracting centerlines from 3D meshes of actual airway trees with MIMICs software and were subsequently refined by clinical experts. *The source code and simulated SKAs will be available at https://github.com/MorineZ/Skeleton2Mask for further research.*

5.2. Implementation Details

Data Preprocessing During the preprocessing stage, the pixel values of the CT scans were clamped to $[-1000, 400]$ Hounsfiled Unit and then rescaled to $[0, 255]$. All images were cropped to the lung region using the Lungmask algorithm (Hofmanninger et al., 2020) to reduce irrelevant background. Sliding windows with a size of $[96, 96, 96]$ and a stride of 64 were implemented to generate the input patches. Skeleton2Mask was implemented using the PyTorch framework and trained in a distributed manner on 2 NVIDIA GeForce RTX 3090 GPUs.

Optimization Procedure We employed an on-the-fly data augmentation, including random contrast adjustment, Gaussian smoothing, axis flipping, and affine transformations to enhance the robustness and generalization of the model during training. Adam optimizer was employed with an initial learning rate of 3e-3. For the BAS dataset, the learning rate was reduced by a factor of 10 at the 20th and 60th epochs using a multi-step scheduler, with training up to 100 epochs. For the ATM22 dataset, the learning rate decay occurred at the 20th and 40th epochs, with training up to 80 epochs. Additionally, for both datasets, training was terminated early if the average change in training loss over 10 consecutive epochs was less than 1e-3 after epoch 60 (BAS) or epoch 40 (ATM22).

Evaluation Metrics Following (Qin et al., 2021), we adopt both volumetric-level and topology-level metrics to comprehensively evaluate the segmentation performance. The volumetric-level metrics include Dice Similarity Coefficient (DSC), True Positive Rate (TPR), and False Positive Rate (FPR), which assess voxel-wise overlap between the predicted segmentation and the reference standard. To evaluate topological preservation, we use three widely-adopted centerline-based metrics: Tree-length Detected (TD) (Lo et al., 2012), Branches Detected Rate(BD) (Lo et al., 2012), and its stricter version BD*. As centerline-based metrics, the topological metrics use the centerline derived from SKAs during computation. TD measures

[4] https://atm22.grand-challenge.org/

the fraction of total tree length in the reference that is correctly detected. BD quantifies the fraction of reference standard branches that are detected, where a branch is considered detected if at least one voxel of it is correctly segmented (Qin et al., 2021). BD* is additionally introduced as a stricter version of BD, where a branch is considered detected only if at least 80% of voxels is correctly segmented. Notably, all evaluations are performed after thresholding the predicted probabilities at 0.5 and extracting the largest connected component. Each experiment is repeated three times to report average performance.

5.3. Clinical Annotation Practice

To systematically assess the clinical feasibility of skeleton annotation (SKA) and its potential benefits compared to full voxel-wise annotation (VA), we established three annotation protocols in our clinical workflow: SKA, fully manual voxel-wise annotation (MVA), and Semi-automatic Voxel-wise Annotation (SVA, involving the adaptive brush, local threshold and point seed methods). Specifically, we invited four radiologists from XXX Hospital in China (a tertiary-level hospital), each with over five years of professional experience, to perform both full annotations and skeleton annotations on 12 CT scans randomly selected from the BAS dataset. The detailed annotation criteria are presented in Table 1. To ensure a fair comparison, both annotation methods were implemented from 7 PM to 9 PM every evening on the same workstation using mouse clicks. Additionally, aside from CT images, no pre-segmented results were available as references throughout the annotation process.

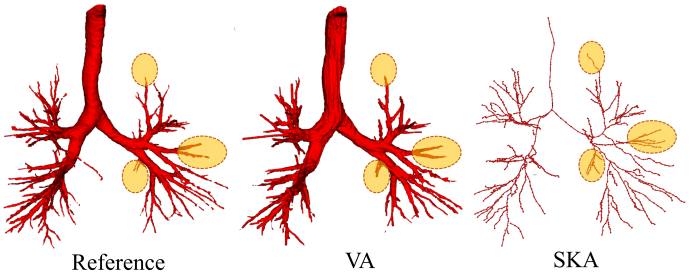

Fig. 6: **Visual comparison between VA and SKA.** The yellow regions highlight the additional branches annotated by SKA compared to voxel-wise and reference annotations.

5.3.1. Visual comparison: SKA vs. VA

Fig. 6 presents the visual comparative results of VA and SKA. We observed that: **1)** VA is prone to introducing edge noise and is more susceptible to branch annotation loss or fragmentation within constrained time limits. Note that the gold standard of airway segmentation is defined by the largest connected component of the full annotation (Zhang et al., 2023a). This implies a higher risk of branch missing annotation, which significantly hampers the evaluation of segmentation models and the quantitative analysis of airway parameters; **2)** As shown in the yellow region, skeleton annotations often include more true branches than fully voxel-wise or even reference annotations. This is because annotators can focus more on topological structures during the skeleton annotation process, which facilitates more comprehensive delineation of branching patterns. This also implicitly suggests that the reference labels may miss some true branches.

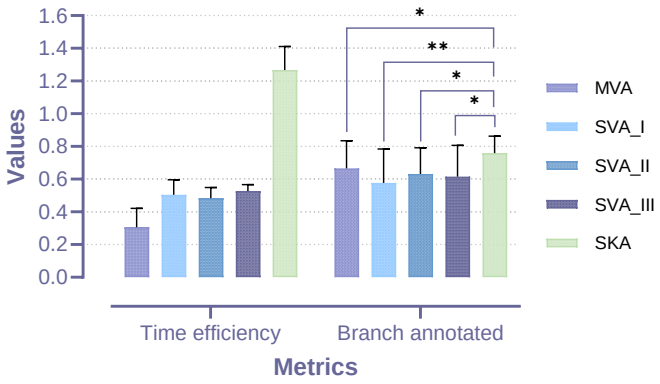

Fig. 7: **A statistical quantitative comparison between VA (including MVA& SVA) and SKA.** SVA_I/II/III represent the three semi-automatic voxel-level annotation strategies: adaptive brush, local threshold, and point seed, respectively."Time efficiency" means the number of subjects annotated per hour. "Branch annotated", similar to the BD metric, refers to the proportion of annotated branches relative to the total number of branches. Statistical significance: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

5.3.2. Quantitative comparison: SKA vs. VA

As a quantitative analysis, Fig. 7 compares the time efficiency and the proportion of annotated branches to actual branches under different annotation strategies from a statistical perspective. Overall, SKA significantly improves annotation efficiency while marking substantially more airway branches. **1)** Compared to MVA, SKA improves annotation efficiency by **4.18×**(0.305→ 1.267) while labeling more tracheal branches, saving approximately 80% of time costs, demonstrating its superiority. In fact, during the annotation process, we observed that replacing mouse clicks with a touchscreen pen increased the annotation efficiency of SKA by 1.5 times. **2)** Even compared to SVA, the proposed SKA achieves **2×** improvement (0.504/0.483/0.527→ 1.267) in efficiency, accompanied by a significantly higher branch labeling rate. This is primarily because, although various semi-automatic methods attempt to enhance annotation efficiency by leveraging local intensity distributions, contrast information, or prompt-based guidance, their effectiveness remains limited when applied to airway trees characterized by diffuse boundaries and abundant fine branches. Moreover, these approaches are particularly susceptible to image noise, further constraining their applicability in complex anatomical structures. **3)** Given the numerous airway branches, it is challenging to annotate all branches in a single pass. After an interruption in the annotation workflow, all annotators found that SKA was more conducive than VA for adding previously missed branches; **4)** As SKA constructs the airway structure on a branch-by-branch basis, it enables precise branch-level analysis and parameterized analysis, such as evaluating the number of bronchial branches beyond the sixth generation and their average diameters, offering valuable insights for the early detec-

Table 1: **Implementation details of annotations in the clinical practice experiment.**

Anno. (Annotation)	MVA	SVA			SKA
		Adaptive brush	Local threshold	Point seed	
Anno. dimension	3D	3D			3D
Anno. software	ITK-SNAP	3D slicer			MIMICs
Min anno. unit/Avg interval	voxel/single voxel	voxel/single voxel			control point/3.89mm
Secondary anno. unit	slice-by-slice	slice-by-slice			branch-by-branch
Anno. time	From 7–9 p.m. daily (after work)				
Anno. device	Lenovo Thinksation P360 (with mouse clicks)				

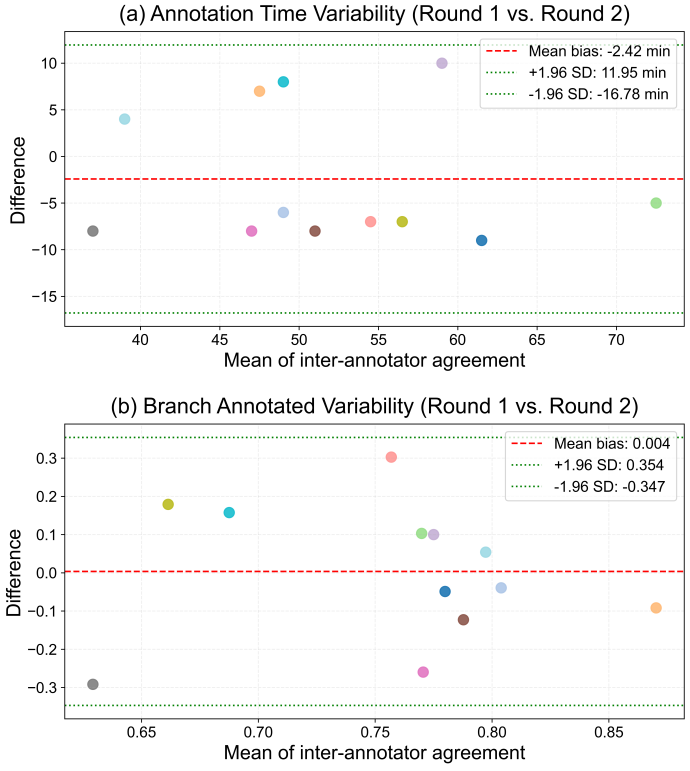

Fig. 8: **Inter-annotator consistency analysis (Bland-Altman plot) of skeleton annotations.** The x-axis shows the average annotation time (min) or rate of branches annotated per sample across two rounds, while the y-axis displays the corresponding differences (Round 2 – Round 1) in time or branch count.

tion of respiratory disorders.

5.3.3. *Inter-annotator variability in SKA*

To further evaluate the robustness of SKA, we analyzed inter-annotator variability using Bland-Altman plots (Fig. 8) and Intraclass Correlation Coefficient (ICC) analysis. **1)** In terms of annotation time, both Bland-Altman (Fig. 8(a) mean bias: -2.42 min; all points within the 95% limits of agreement (LoA): -16.78 to 11.95 min) and ICC analyses (0.859, 95% CI: 0.51-0.96) confirmed good inter-annotator consistency. The observed variation range may reflect reasonable differences in annotation rhythm, potentially influenced by factors such as software familiarity or control-point density in SKA. **2)** Interestingly, despite the variation in annotation time, Fig. 8(b) demonstrated remarkable consistency in the proportion of annotated branches with a negligible mean bias of 0.004 and tight limits of agreement (LoA: ± 0.35). **3)** These results highlight the robustness of SKA: despite observable differences in annotation time, its topology-focused nature ensures high inter-annotator consistency in the structural preservation of annotations. This robustness is essential for downstream quantitative analysis and clinical utility, reinforcing SKA's suitability for tubular object annotations.

5.4. *Comparative Test*

5.4.1. *Quantitative results*

Table 2 presents quantitative comparison results on the test set of the BAS dataset and ATM22 dataset, showcasing the segmentation performance of our algorithm under SKA supervision alongside other sparsely supervised algorithms, as well as the segmentation results of existing popular fully supervised algorithms. For fair comparison, all competing algorithms were reproduced based on publicly available code and were trained and tested using the same data splits. Compared to voxel-level metrics, the differences observed in topological-level metrics are more pronounced. This can be attributed to the intrinsic tree-like structure of the airways and the substantial intra-class and inter-class imbalance characteristics.

Firstly, we compare our approach with other sparse-supervised learning methods, including popular regularization learning (Grandvalet and Bengio, 2004; Tang et al., 2018; Zhai et al., 2023) and dynamic pseudo-label learning (Luo et al., 2022; Lee and Jeong, 2020). Notably, apart from PA-Seg (Zhai et al., 2023), all other comparison methods adopt an end-to-end sparse supervision strategy based on scribble annotations without involving pseudo-labels. As mentioned in Sec. 1, scribble annotations exhibit greater randomness in labeling locations, which, as more samples are annotated, can provide the model with diverse semantic information from varying positions. Consequently, the density of supervisory information is higher. In contrast, SKA exhibits stronger positional preference, and directly applying SKA for end-to-end propagation learning in such algorithms often leads to training failures or suboptimal results due to excessively sparse supervision. Therefore, when comparing with these algorithms, we use the mask proposal $\mathcal{MP}_{\odot}$ as the supervision signal to replace SKA (denoted by †). Results on the BAS dataset demonstrate that our method achieved superior performance, outperforming the 2nd approach (Grandvalet and Bengio, 2004) with considerable margins (+11.26% BD*, +8.27% BD, +9.96% TD, +7.30% TPR, +1.88% DSC, $p < 0.01$). This is more pronounced on the ATM22 dataset. We believe these methods are

Table 2: **Quantitative results of different methods under skeleton annotation (SKA, <3% airway labeled) and full voxel-wise annotation (VA, 100% airway labeled) based on BAS90 Dataset (Qin et al., 2020) and ATM22 Dataset (Zhang et al., 2023a).** The **bold** and underline highlight the best and second-best results among sparse supervision algorithms for each dataset. Anno. Time is estimated based on 12 manually labeled CT scans by experienced clinicians.

Dataset	Anno. Num.	Anno. Time ↓	Anno. voxels ↓	Methods	BD *(%) ↑	BD(%) ↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑	FPR(%) ↓
BAS	72 VAs	286.6h	100.0%	U-Net (Çiçek et al., 2016)(Upper bound)	$62.75_{\pm 2.4}$	$79.52_{\pm 1.1}$	$75.10_{\pm 1.6}$	$90.53_{\pm 0.4}$	$91.13_{\pm 0.7}^{***}$	$.032_{\pm .007}^{***}$
				SwinUnetr (Hatamizadeh et al., 2021)	$62.42_{\pm 2.4}$	$79.04_{\pm 1.2}$	$74.76_{\pm 1.8}$	$91.59_{\pm 0.5}$	$89.71_{\pm 0.6}$	$.019_{\pm .006}$
				NaviAirway (Wang et al., 2022a)	$60.34_{\pm 2.4}$	$78.85_{\pm 1.1}$	$72.09_{\pm 2.1}$	$86.28_{\pm 0.4}$	$90.77_{\pm 0.9}$	$.060_{\pm .001}$
				WingNet (Zheng et al., 2021b)	$62.84_{\pm 2.5}$	$79.88_{\pm 1.4}$	$75.73_{\pm 1.8}$	$91.68_{\pm 0.5}$	$89.69_{\pm 0.7}$	$.018_{\pm .007}$
	72 SKAs	**56.8h**	**1.96%**	pCE_{only} (Lin et al., 2016)†	$40.70_{\pm 2.9}^{***}$	$60.21_{\pm 3.3}^{***}$	$53.06_{\pm 3.4}^{***}$	$86.21_{\pm 0.3}^{***}$	$78.99_{\pm 0.7}^{***}$	$.010_{\pm .004}^{***}$
				EWMA (Lee and Jeong, 2020)†	$41.25_{\pm 0.1}^{***}$	$61.26_{\pm 0.1}^{***}$	$54.48_{\pm 0.2}^{***}$	$84.77_{\pm 0.3}^{***}$	$77.21_{\pm 0.1}^{***}$	$.012_{\pm .001}^{***}$
				Luo *et al.* (Luo et al., 2022) †	$41.18_{\pm 0.1}^{***}$	$61.33_{\pm 0.1}^{***}$	$54.60_{\pm 0.3}^{***}$	$84.74_{\pm 0.3}^{***}$	$77.20_{\pm 0.1}^{***}$	$\mathbf{.010_{\pm .001}}^{***}$
				Gatedcrf (Tang et al., 2018)†	$45.93_{\pm 2.5}^{***}$	$65.35_{\pm 2.3}^{***}$	$58.55_{\pm 2.5}^{***}$	$84.92_{\pm 0.7}^{***}$	$79.84_{\pm 2.6}^{***}$	$.025_{\pm .010}^{***}$
				PA-Seg (Zhai et al., 2023)	$48.55_{\pm 0.6}^{***}$	$67.43_{\pm 0.4}^{***}$	$61.13_{\pm 0.8}^{***}$	$85.45_{\pm 0.1}^{***}$	$81.05_{\pm 0.7}^{***}$	$.030_{\pm .005}$
				EM (Grandvalet and Bengio, 2004)†	$51.64_{\pm 1.3}^{***}$	$69.69_{\pm 1.2}^{***}$	$63.93_{\pm 1.4}^{***}$	$87.62_{\pm 0.6}^{**}$	$81.20_{\pm 0.7}^{***}$	$\mathbf{.010_{\pm .001}}^{***}$
				Skeleton2Mask (Ours)	$\mathbf{62.90_{\pm 1.4}}$	$\mathbf{77.96_{\pm 0.5}}$	$\mathbf{73.89_{\pm 0.8}}$	$\mathbf{89.50_{\pm 0.3}}$	$\mathbf{88.50_{\pm 0.6}}$	$.028_{\pm .001}$
ATM22	239 VAs	951.4h	100.0%	U-Net (Çiçek et al., 2016)(Upper bound)	$78.44_{\pm 2.7}^{***}$	$90.22_{\pm 1.5}^{**}$	$88.49_{\pm 1.5}^{***}$	$91.61_{\pm 0.6}^{***}$	$94.38_{\pm 0.7}^{***}$	$.037_{\pm .007}^{***}$
				SwinUnetr (Hatamizadeh et al., 2021)	$79.09_{\pm 3.0}$	$90.70_{\pm 1.6}$	$88.89_{\pm 2.0}$	$88.94_{\pm 0.7}$	$94.63_{\pm 1.0}$	$.060_{\pm .010}$
				WingNet (Zheng et al., 2021b)	$79.97_{\pm 2.7}$	$90.40_{\pm 1.6}$	$89.26_{\pm 2.0}$	$92.94_{\pm 0.7}$	$93.32_{\pm 0.8}$	$.025_{\pm .010}$
	239 SKAs	**188.6h**	**2.67%**	pCE_{only} (Lin et al., 2016)†	$54.67_{\pm 1.1}^{***}$	$82.60_{\pm 0.4}^{***}$	$72.82_{\pm 0.7}^{***}$	$82.88_{\pm 0.7}^{***}$	$76.27_{\pm 0.8}^{***}$	$.028_{\pm .002}$
				Gatedcrf (Tang et al., 2018)†	$58.34_{\pm 1.5}^{***}$	$84.10_{\pm 0.7}^{***}$	$75.48_{\pm 1.1}^{***}$	$82.54_{\pm 0.7}^{***}$	$76.54_{\pm 0.4}^{***}$	$.032_{\pm .010}$
				PA-Seg (Zhai et al., 2023)	$57.18_{\pm 0.5}^{***}$	$84.07_{\pm 0.3}^{***}$	$74.68_{\pm 0.4}^{***}$	$82.91_{\pm 0.2}^{***}$	$76.22_{\pm 0.3}^{***}$	$.027_{\pm .002}$
				EM (Grandvalet and Bengio, 2004)†	$59.64_{\pm 0.5}^{***}$	$84.41_{\pm 0.4}^{***}$	$76.36_{\pm 0.2}^{***}$	$86.31_{\pm 0.2}^{***}$	$77.68_{\pm 0.3}^{***}$	$\mathbf{.007_{\pm .001}}^{***}$
				Skeleton2Mask (Ours)	$\mathbf{75.00_{\pm 1.8}}$	$\mathbf{89.41_{\pm 0.6}}$	$\mathbf{86.32_{\pm 1.1}}$	$\mathbf{89.07_{\pm 0.7}}$	$\mathbf{87.97_{\pm 0.7}}$	$.023_{\pm .005}$

† denotes methods which utilize our mask proposal $\mathcal{MP}_{\odot}$ during training instead of the SKA, thus avoiding the crash of direct learning.
*indicate the significance level between our and other sparse supervision algorithms.

Table 3: **Time consumption and performance evaluation under VA and SKA modes on the BAS dataset (Qin et al., 2020).**

Method	Anno.	Labeled subjects	Time Cost(h)	BD *(%) ↑	BD(%) ↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑	FPR(%) ↓
U-Net (Çiçek et al., 2016)	**VA**	14 (20%)	55.7	$51.19_{\pm 1.5}$	$69.15_{\pm 1.3}$	$62.88_{\pm 1.4}$	$87.17_{\pm 0.2}$	$86.31_{\pm 0.8}$	$.031_{\pm .003}$
		29 (40%)	115.4	$55.26_{\pm 2.0}$	$73.14_{\pm 1.2}$	$67.70_{\pm 1.5}$	$88.70_{\pm 0.2}$	$88.51_{\pm 0.4}$	$.030_{\pm .003}$
		72 (All)	286.6	$62.75_{\pm 2.4}$	$79.52_{\pm 1.1}$	$75.10_{\pm 1.6}$	$90.53_{\pm 0.4}$	$91.13_{\pm 0.7}$	$.032_{\pm .007}$
Ours	**SKA**	72 (All)	56.8	$62.90_{\pm 1.4}$	$77.96_{\pm 0.5}$	$73.89_{\pm 0.8}$	$89.50_{\pm 0.3}$	$88.50_{\pm 0.6}$	$.028_{\pm .001}$

primarily applied in the segmentation of convex structures, emphasizing the reliability and robustness of the segmentation results. This may hinder the network from achieving higher recall on distal branches, as these fine branches are often challenging to capture or exhibit lower confidence. Contrarily, our approach fully considers the target's scale characteristics and the annotations' positional characteristics. By leveraging hierarchical geometry-aware learning, we achieved a more accurate and structure-friendly airway segmentation. Additionally, we compared with fully supervised algorithms, including the classical segmentation architecture (U-Net (Çiçek et al., 2016) and SwinUnetr (Hatamizadeh et al., 2021)) and the popular airway segmentation methods (NaviAirway (Wang et al., 2022a) and WingNet (Zheng et al., 2021b)). The algorithm attained segmentation performance comparable to full supervision in both topological and voxel-level metrics while saving about 80% of the annotation cost.

To further evaluate the trade-off between annotation efficiency and segmentation performance under two annotation modes (VA *vs.* SKA), we compare our method with the fully supervised performance under different training sample proportions of VA on the BAS dataset in Table 3. With comparable annotation costs (100% SKAs: 55.7h *vs.* 20% VAs: 56.8h), our algorithm demonstrates a significant performance advantage, showing an 11.7% improvement in BD*, an 11% increase in TD, and a 2.33% enhancement in DSC. Additionally, we consider a hypothetical scenario where the interval between control points in skeleton annotations is halved, which would approximately double the annotation cost (*i.e.,* 56.8 → 113.6h). Even under this assumption, our method with 100% SKA annotations still demonstrates a clear segmentation advantage compared to using 40% VA annotations with a similar time budget. This indicates a better balance between annotation efficiency and segmentation performance.

5.4.2. *Qualitative results*

Fig. 9 gives an intuitive performance comparison for the above methods across multiple CT scans (I → III indicate easy → hard). It can be observed that 1) segmentation discrepancies between different algorithms are mainly evident in the fine distal branches. Our method, while maintaining limited false-positive detections (even in the noisy image III), accurately identifies more distal branches; 2) additionally, we achieve comparable performance with the fully supervised method and even superior topological metrics in some easy cases. In Case II, the BD metric demonstrates a comparison of 0.74 vs 0.80. The above findings can be attributed to: 1) the SKA strategy simplifies the annotation process while better preserving topological completeness, thereby providing the model with more distal branch references; 2) the incorporation of two auxiliary branches in the HGL framework introduces fine-grained structural information and enhances the topological consistency of

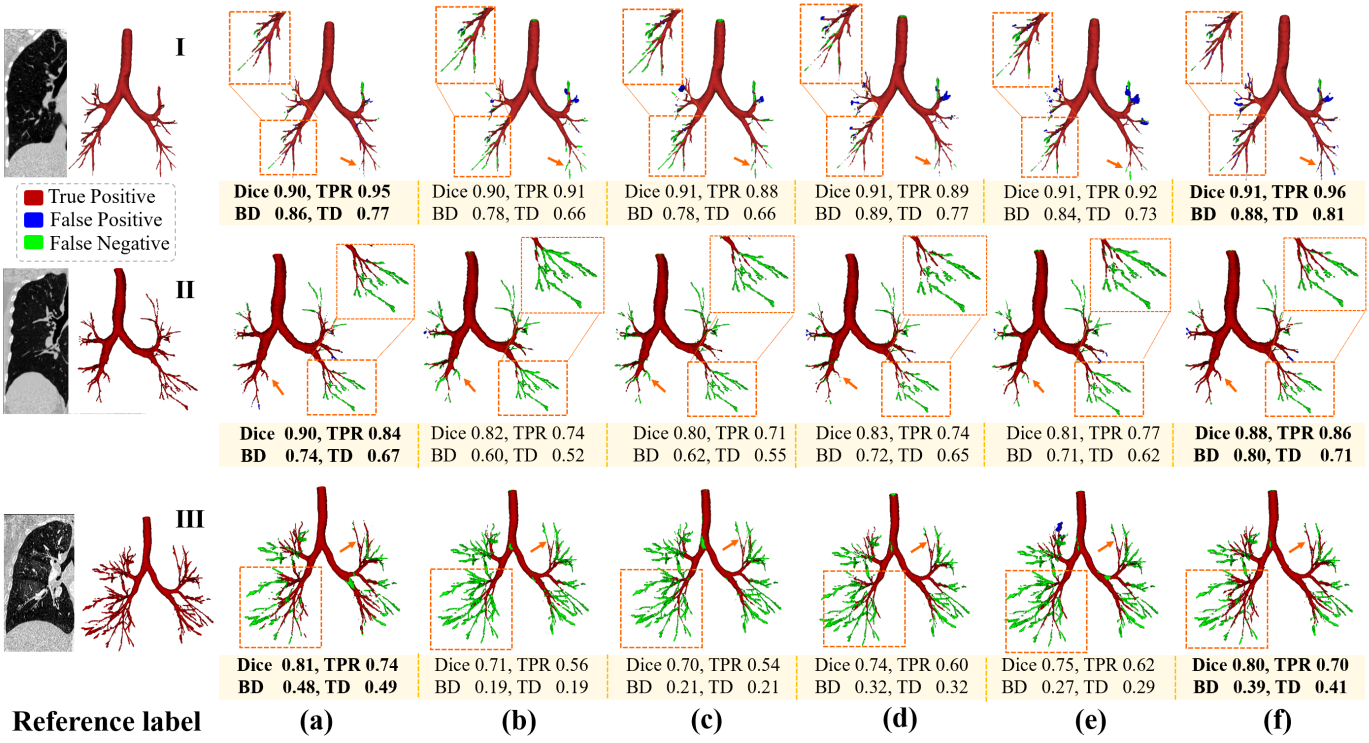

Fig. 9: **Segmentation results of different methods across three groups of CT scans (I-III).** (a-f) indicate the Segmentation results of 3D UNet (upper bound)(a) (Çiçek et al., 2016), pCE$_{only}$(b) (Lin et al., 2016), EWMA(c) (Lee and Jeong, 2020), EM(d) (Grandvalet and Bengio, 2004), PA-Seg(e) (Zhai et al., 2023) and Our Skeleton2Mask(f). Orange boxes and arrows highlight the branches with significant segmentation discrepancies.

the segmentation results.

5.5. Ablation Study

Ablation studies are provided to evaluate the importance of the two key designs of the Skeleton2Mask method: 1) Dual-stream Buffer Propagation and 2) Hierarchical Geometry-aware propagation Learning.

5.5.1. DBP strategy

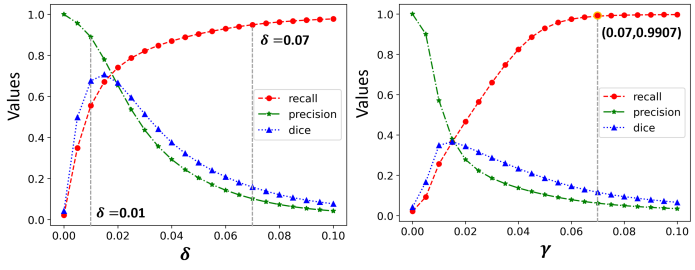

Fig. 10: **Effect of hyperparameter δ and γ on label expansion on the BAS90 dataset.** Left: Metric measures between the foreground region and the gold standard under different δ in the G^2BI strategy. Right: Metric measures between the foreground region and the gold standard under different γ in the *EBI* strategy.

Table 4 presents the component composition of the DBP strategy and the impact of hyperparameter adjustments on segmentation performance. It can be observed that direct geodesic

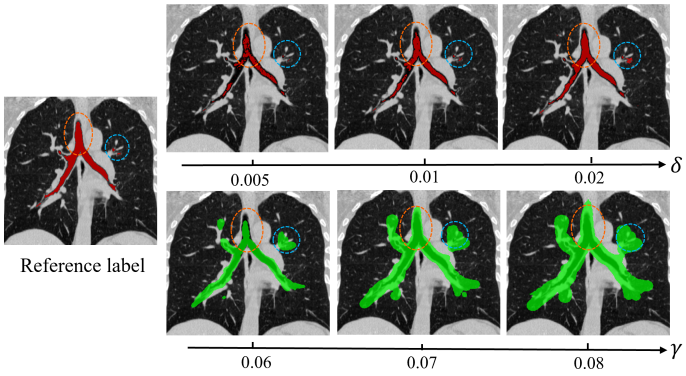

Fig. 11: **Visualization of foreground expansion under different δ settings (UPPER)and background expansion under different γ settings (LOWER).** The red region represents the foreground, green indicates uncertainty, and all other areas correspond to the background. The orange dashed circles highlight the large-scale primary bronchi, while the cyan dashed circles emphasize the distal bronchioles.

distance buffering inference (*i.e.,* GBP) results in poor segmentation performance, largely due to the intrinsic noise properties of the images, which hinder the generation of reliable mask proposals. The introduction of Gaussian smoothing ($GBP \rightarrow G^2BP$), however, greatly improves label consistency within classes, enabling geodesic distance to serve as a more dependable metric for label propagation, thus leading to a notable improvement of 9.77% in DSC and 16.95% in TD. The EBP strategy, which only considers spatial distance, struggles to handle targets with significant intra-class scale variations. However,

Table 4: **Ablation study for DBP strategy with loss function $\mathcal{L} = \mathcal{L}_{pce}$ for default.** *GBP* denotes Geodesic distance-based buffer propagation without Gaussian smoothing. The gray shading indicates our settings.

Setting			Params.		Metrics					
GBP	G^2BP	*EBP*	δ_1	δ_2	BD *(%) ↑	BD(%) ↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑	FPR(%) ↓
✓			0.01	0.07	$24.34_{\pm 2.2}$	$37.39_{\pm 3.7}$	$33.59_{\pm 2.9}$	$75.91_{\pm 0.7}$	$77.66_{\pm 1.4}$	$.280_{\pm .059}$
	✓		0.01	0.07	$38.69_{\pm 1.3}$	$57.63_{\pm 1.6}$	$50.54_{\pm 1.5}$	$85.68_{\pm 0.9}$	$78.70_{\pm 1.3}$	$.009_{\pm .002}$
		✓	-	-	$34.27_{\pm 1.7}$	$52.55_{\pm 2.0}$	$44.83_{\pm 2.7}$	$56.03_{\pm 2.2}$	$47.51_{\pm 2.0}$	$.051_{\pm .010}$
	✓	✓	0.005	0.07	$36.89_{\pm 5.3}$	$55.75_{\pm 6.9}$	$49.08_{\pm 6.7}$	$83.30_{\pm 0.4}$	$73.24_{\pm 0.9}$	$\mathbf{.004_{\pm .002}}$
	✓	✓	0.02	0.07	$\mathbf{53.74_{\pm 4.3}}$	$\mathbf{71.42_{\pm 3.7}}$	$\mathbf{65.00_{\pm 4.1}}$	$72.25_{\pm 2.3}$	$\mathbf{85.40_{\pm 1.6}}$	$.250_{\pm .047}$
	✓	✓	0.01	0.06	$39.90_{\pm 0.8}$	$59.28_{\pm 1.0}$	$52.48_{\pm 1.0}$	$82.94_{\pm 0.5}$	$74.97_{\pm 0.3}$	$.020_{\pm .001}$
	✓	✓	0.01	0.08	$41.62_{\pm 1.6}$	$60.58_{\pm 2.1}$	$53.58_{\pm 1.8}$	$81.20_{\pm 3.1}$	$78.85_{\pm 2.0}$	$.060_{\pm .018}$
	✓	✓	0.01	0.07	$40.70_{\pm 2.9}$	$60.21_{\pm 3.3}$	$53.06_{\pm 3.4}$	$\mathbf{86.21_{\pm 0.3}}$	$78.99_{\pm 0.7}$	$.010_{\pm .004}$

Table 5: **Ablation study for HGL framework with DBP strategy for default.**The gray shading indicates our settings.

$\mathcal{L}_{pce}$	$\mathcal{L}_{em}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{topo}$	λ_1	λ_2	λ_3	BD *(%) ↑	BD(%) ↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑	FPR(%) ↓
✓				-	-	-	$40.70_{\pm 2.9}$	$60.21_{\pm 3.3}$	$53.06_{\pm 3.4}$	$86.21_{\pm 0.3}$	$78.99_{\pm 0.7}$	$.010_{\pm .004}$
✓	✓			1.5	-	-	$51.64_{\pm 1.3}$	$69.69_{\pm 1.2}$	$63.93_{\pm 1.4}$	$87.62_{\pm 0.6}$	$81.20_{\pm 0.7}$	$\mathbf{.010_{\pm .001}}$
✓	✓	✓		1.5	20	-	$58.65_{\pm 1.1}$	$75.82_{\pm 1.1}$	$70.55_{\pm 1.0}$	$88.79_{\pm 0.3}$	$87.60_{\pm 0.8}$	$.030_{\pm .001}$
✓	✓		✓	1.5	-	0.01	$63.46_{\pm 1.0}$	$77.45_{\pm 1.2}$	$73.61_{\pm 0.9}$	$87.22_{\pm 0.4}$	$84.10_{\pm 0.7}$	$.026_{\pm .001}$
✓	✓	✓	✓	1.5	20	0.01	$\mathbf{62.90_{\pm 1.4}}$	$\mathbf{77.96_{\pm 0.5}}$	$\mathbf{73.89_{\pm 0.8}}$	$\mathbf{89.50_{\pm 0.3}}$	$\mathbf{88.50_{\pm 0.6}}$	$.028_{\pm .001}$
✓	✓	✓	✓	1	20	0.01	$59.83_{\pm 1.4}$	$75.49_{\pm 1.0}$	$71.00_{\pm 1.2}$	$89.44_{\pm 1.1}$	$86.88_{\pm 1.1}$	$.020_{\pm .001}$
✓	✓	✓	✓	2	20	0.01	$58.64_{\pm 3.5}$	$74.32_{\pm 3.7}$	$70.17_{\pm 3.5}$	$89.11_{\pm 0.9}$	$86.38_{\pm 2.7}$	$.020_{\pm .001}$
✓	✓	✓	✓	1.5	10	0.01	$65.67_{\pm 1.8}$	$78.74_{\pm 1.5}$	$75.33_{\pm 1.5}$	$87.96_{\pm 0.3}$	$86.11_{\pm 0.8}$	$.028_{\pm .001}$
✓	✓	✓	✓	1.5	40	0.01	$59.02_{\pm 0.8}$	$75.05_{\pm 0.6}$	$70.30_{\pm 0.7}$	$89.42_{\pm 0.4}$	$85.11_{\pm 1.0}$	$.022_{\pm .001}$
✓	✓	✓	✓	1.5	20	0.005	$62.94_{\pm 1.6}$	$77.40_{\pm 1.3}$	$73.44_{\pm 0.8}$	$88.98_{\pm 0.5}$	$87.98_{\pm 0.7}$	$.030_{\pm .002}$
✓	✓	✓	✓	1.5	20	0.02	$78.88_{\pm 1.3}$	$85.96_{\pm 1.7}$	$86.03_{\pm 1.4}$	$83.41_{\pm 0.9}$	$91.28_{\pm 1.1}$	$.088_{\pm .002}$

as an additional constraint, it aids in identifying more background regions that could easily be confused with the foreground, thereby enhancing the overall performance of the algorithm (e.g., DSC from 85.68% to 86.21%, $p < 0.01$). In this context, to identify the optimal values for parameters δ and γ in DBP for both G^2BP and EBP, we use SKA as the initial seed point. By gradually increasing δ and γ , we observe the evolution of the Dice, precision, and recall metrics of the seed expansion region in comparison to the ground truth. As depicted in Fig. 10, as they increase, both the left and right subplots display similar patterns. Specifically, when δ and γ are small, the expansion region is small, with a low recall and high precision. As they increase, the expansion region grows, leading to an increase in recall for the target region, while precision gradually decreases. This leads to an overall trend in Dice, which first increases and then decreases. To provide further insights, the seed expansion results under various δ and γ settings are visualized in Fig. 11. When δ serves as the foreground expansion factor, $\delta = 0.005$ can obtain a foreground region with high precision. Nonetheless, the limited label expansion provides only weak supervision. Conversely, when $\delta = 0.02$, although the supervisory signals become more strong, excessive label expansion leads to the accumulation of incorrect labels. Notably, this results in significant label leakage in distal bronchi. Correspondingly, as observed from Table 4, excessive foreground expansion tends to induce over-segmentation of the airways, resulting in significant false positive predictions. Conversely, insufficient expansion leads to under-segmentation, characterized by a reduced TPR. To address this, $\delta = \delta_1 = 0.01$ is a more suitable choice, yielding the optimal DSC performance. Similarly, when δ acts as the background expansion factor, smaller values may risk incorrectly classifying true airway regions as background, whereas larger values lead to insufficient expansion of the background label. Therefore, we set $\delta = \delta_2 = 0.07$ as the background expansion factor. Likewise, γ functions as a background expansion parameter, requiring a trade-off between sufficient expansion of the background region with the potential risk of foreground intrusion (reflected by precision and recall in Fig. 10). Consequently, γ is set to 0.07.

5.5.2. *HGL framework*

Leveraging reliable pseudo-labels obtained through the DBP strategy, Table 5 presents the impact of different loss functions and their weight configurations on segmentation performance within the HGL architecture. By promoting higher-confidence predictions, entropy regularization markedly enhances the accuracy of mask proposals in certain regions. However, its effect on the uncertain branch edges is relatively constrained, leading to substantial improvements in topology-level metrics but limited gains in voxel-level metrics, like TPR. As a fine-grained reconstruction auxiliary task, the integration of $\mathcal{L}_{rec}$ brings an improvement of both topology-level metrics (+7.01% BD*,

+6.62% TD) and voxel-level metrics (+6.40% TPR, +1.17% DSC, $p < 0.01$) by supplying detailed structural information from low-level features, thus mitigating the loss of distal fine branch detection caused by regularization. Topology-aware learning effectively leverages the topological integrity of skeleton-level annotations, significantly enhancing the network's performance in topological predictions. Comparing the two auxiliary tasks, it is evident that the reconstruction task, as a global fine-grained perception, shows various improvements in voxel-level metrics (*e.g.*, TPR: 81.20% → 87.60% and DSC: 87.62% → 88.79%, $p < 0.01$). In contrast, the $\mathcal{L}_{topo}$, which serves as a supervision of topological level perception, produces more pronounced enhancements in topological metrics (*e.g.*, BD*: 51.64% → 63.46% and TD: 63.93% → 73.61%). These two tasks complement each other and, along with the segmentation branch, form a hierarchical sparse supervision paradigm that transitions from skeleton to mask, and from soft to hard, significantly boosting dense prediction under sparse supervision.

To determine the loss weight configurations, we conducted a series of hyperparameter experiments in Table 5. Given the intrinsic scale and gradient magnitudes of these loss functions, we assigned relatively larger weights to $\mathcal{L}_{rec}$ (*e.g.*, $\lambda_2 = 20$) to account for its small gradient magnitude, while using smaller weights for $\mathcal{L}_{topo}$ (*e.g.*, $\lambda_3 = 0.01$) to avoid the model's overemphasis on sparse topological structures at the expense of peripheral voxel predictions. As the weight of the entropy minimization loss, λ_1 was set to a moderate value 1.5, as overly large or small values tended to suppress valid uncertain regions or cause fragmented predictions, both leading to reduced TPR and BD. Furthermore, we analyzed the influence of the reconstruction loss weight λ_2 and topology-aware learning loss weight λ_3 on model behavior. λ_2 and λ_3 play complementary roles in guiding the segmentation process. While λ_2 provides multi-scale and boundary cues for voxel-level dense prediction, λ_3 contributes topological-level structural cues. When a smaller value of λ_2 or a larger value of λ_3 is used, the topological loss $\mathcal{L}_{topo}$ tends to dominate the optimization, leading to an increased number of predicted branches (*i.e.*, BD: 62.90% → 65.67%/78.88%) and frequent false positives. In contrast, a larger λ_2 shifts the model's focus toward voxel-wise reconstruction, resulting in nearly unchanged DSC but a reduction in predicted branches. Interestingly, a smaller λ_3 (*e.g.*, 0.005) shows minimal impact on segmentation performance, suggesting that the topological loss can be assigned a relatively low weight to avoid the significant degradation in voxel-level accuracy caused by larger values. Based on this, we set $\lambda_2 = 20$ and $\lambda_3 = 0.01$ to maintain a balanced optimization between topology-level and voxel-level accuracy.

5.6. More Analysis

5.6.1. Quality analysis of mask proposals

To better showcase the DBP strategy's effectiveness in generating mask proposals and its adaptability to airway structures, we conduct visual ablation and comparative analyses of pseudo-label quality.

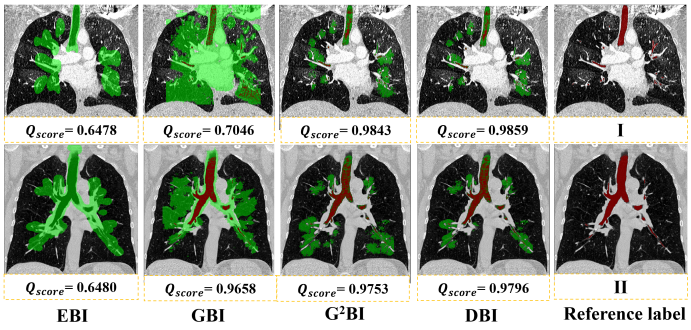

Fig. 12: **Mask proposal Visualization and Quality Evaluation under Different Buffer Inference Strategies.** Red: foreground area, green: uncertain area, otherwise: background area.

Fig. 12 visualizes the mask proposals obtained by different buffer propagation strategies. Furthermore, we assess the quality of these pseudo-labels by measuring two pivotal factors in segmentation: the precision of the certain region and its proportion within the 3D volume. It can be defined as:

$$Q_{score} := Avg(\underbrace{Pre(\mathcal{MP}_{\odot}(i), y(i))}_{precision} + \underbrace{\frac{|\Omega_L|}{|\Omega|}}_{proportion}) \quad (15)$$

Compared to EBP and GBP, G^{2}BI effectively serves as a geometry-aware metric learning that enhances label consistency while accommodating the multi-scale characteristics of branches. Meanwhile, DBP achieves further background expansion after the introduction of the EBP strategy, improving the quality of pseudo-labels while effectively suppressing more false positive predictions.

In Fig.13, we further compare the DBP strategy with existing pseudo-labeling methods, including classical graph-based segmentation algorithms (Superpixel (Chen and Hong, 2022; Zhou et al., 2023b) and ScRoadExtractor (Wei and Ji, 2021))and popular foundation model segmentation techniques (SAM (Kirillov et al., 2023) and SAM-Med2D (Cheng et al., 2023)). ScRoadExtractor (Wei and Ji, 2021) integrates graph-based segmentation with EBP to obtain pseudo-labels for road extraction. The foundation model segmentations are performed using point prompts on SKAs. Obviously, segmentation based on superpixels and foundation models tends to suffer from severe over-expansion in distant branches, leading to segmentation leakage or even complete failure. ScRoadExtractor (Wei and Ji, 2021) adopts a more conservative approach, but its EBP strategy only considers spatial distance, resulting in very limited foreground expansion. In contrast, our method provides a more reasonable label propagation scheme for such tree-like structures, effectively mitigating the issue of absolute label sparsity and enhancing the model's feature representation learning.

5.6.2. In-depth analysis of the auxiliary branches

In this section, we provide a further exploration and analysis of the two auxiliary branches in Skeleton2Mask, *i.e.,* $\mathcal{L}_{rec}$ and $\mathcal{L}_{topo}$.

As mentioned in Sec.2.2, existing sparsely-supervised methods based on auxiliary tasks primarily incorporate edge detec-

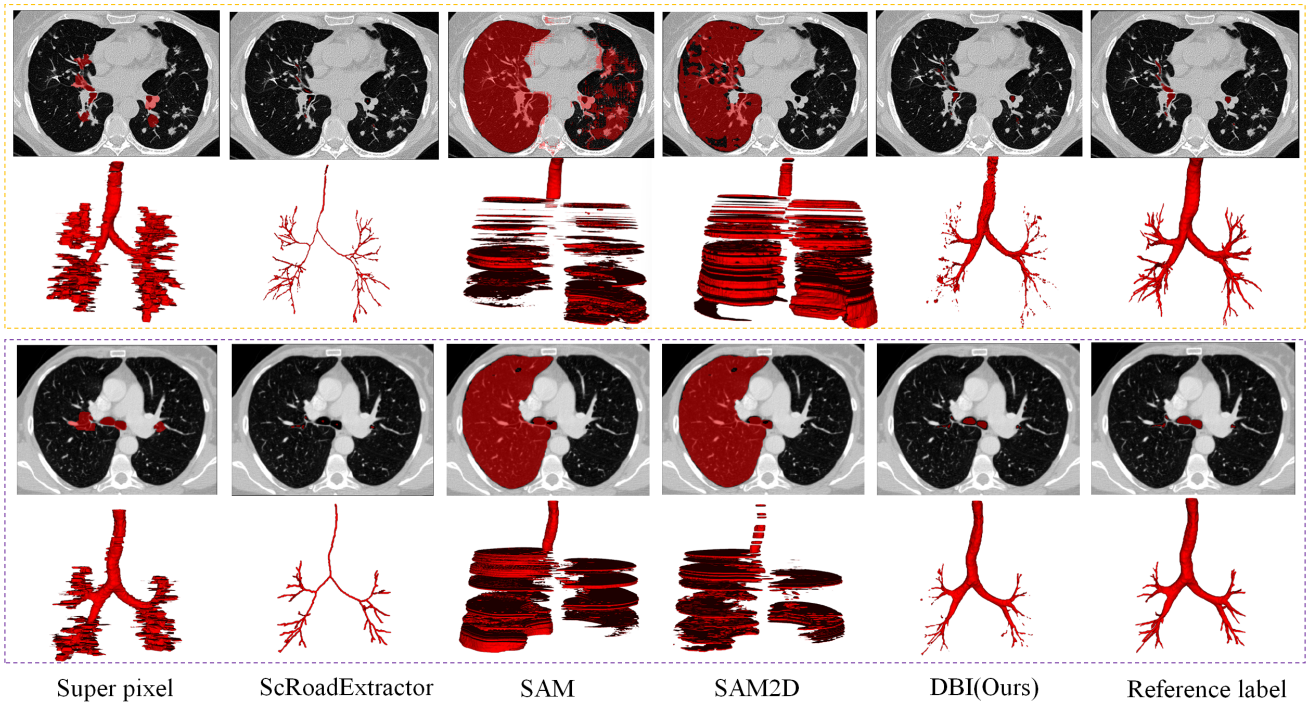

Fig. 13: **Expanding visualization comparison with the same foreground seed (*i.e.*, SKA).**

tion to furnish boundary cues for label propagation. This motivates a deeper delve into the auxiliary reconstruction pathway: *1) Is this fine-grained reconstruction task superior to the edge detection auxiliary task? 2) Can this global probabilistic modeling be equivalent to an edge-aware mechanism, merely providing hints for the uncertain regions in the pseudo-labels?* To this end, in addition to the segmentation branch (*i.e.*, *Baseline* : $\mathcal{L}_{seg} = \mathcal{L}_{pce} + \lambda_1\mathcal{L}_{em}$), we successively introduce two auxiliary tasks for comparison with the global fine-grained reconstruction task. For 1), following (Chen and Hong, 2022; Wei and Ji, 2021), we employ the holistically nested edge detector (HED) (Xie and Tu, 2015) pre-trained on the BSDS500 dataset (Arbelaez et al., 2010) to obtain coarse boundary masks (denoted as *HED*), which is utilized as an auxiliary edge detection target. For 2), we compute the reconstruction loss only on the uncertainty areas, *i.e.*,$i \in \Omega_U$. Comparing the first two rows in Table 6, the inclusion of an edge detection auxiliary task improves the DSC metric but clearly reduces the topological metrics. This can be attributed to the inherent scale imbalance of the airway. Edge detection aids in identifying low-level branches with a larger voxel ratio, but it is less effective for distal branches that dominate the topological structure. A comparison of rows 3 and 4 in the table reveals that when the reconstruction task is confined solely to the uncertain region, the model lacks probabilistic reference information from the certain region. This makes it challenging for the model to effectively collaborate with the segmentation branch in enhancing inter-class separability within the uncertain region. Adding $\mathcal{L}_{topo}$ significantly narrows the gap between local and global reconstructions in topological metrics, though a slight discrepancy remains.

Furthermore, we conducted a deeper investigation into the topological consistency loss $\mathcal{L}_{topo}$. Under the sparse supervision regime based on skeleton annotations, both components of $\mathcal{L}_{topo}$(*i.e.*, $\mathcal{L}_{skel}$ and $\mathcal{L}_{tpc}$) contributed to significant improvements in topological performance. Notably, $\mathcal{L}_{tpc}$ establishes a strong link between the main segmentation task and the skeleton prediction task. Its inclusion markedly enhanced the prediction of branch completeness (i.e., BD* improved from 55.36% to 62.90%). We also examined the effect of introducing $\mathcal{L}_{topo}$ in the context of fully supervised learning, by disabling the reconstruction branch (as listed in the last four rows in Table 6). The results show that the integration of the topology-aware learning branch consistently enhanced the topological metrics across both datasets. Besides, the slight decline in DSC might be due to an overemphasis on topological preservation, which could potentially compromise voxel-wise boundary accuracy. This suggests that the weight of $\mathcal{L}_{\text{topo}}$ should be carefully tuned to balance topological consistency and boundary precision.

5.6.3. *Robustness analysis of annotation shifts*

To investigate the impact of potential positional shifts in real-world annotations on experimental outcomes, we conducted a simulated robustness assessment on the BAS dataset. Specifically, we applied random elastic displacements to the simulated SKAs used in Sec.5.1 along the radial direction away from the lumen, with an average offset of $\tau * r$, where r represents the lumen radius and $\tau \in [0, 1]$ denotes the degree of offset. As shown in Table 7, the topological-level metrics and the composite voxel-level metric (DSC) of the algorithm exhibit stability even as the degree of annotation offset increases. Interestingly, within a certain range of annotation offsets (*e.g.* $\tau \leq 0.5$), there is a slight enhancement in topological metrics compared to $\tau = 0$. We attribute this to the model receiving additional

Table 6: **In-depth analysis of the auxiliary branches: Soft Geometry-aware Propagation Guidance $\mathcal{L}_{rec}$ and Topology-aware Learning $\mathcal{L}_{\text{topo}}$.** *HED* denotes boundary masks obtained from Holistically-Nested Edge Detector (Xie and Tu, 2015). Ω and Ω_U refer to regions where the auxiliary supervision is applied to the entire volume and only to the uncertain regions, respectively. The gray shading indicates the baseline results of fully supervised learning (with U-Net) on the two datasets, as reported in Table 2.

Anno.	Dataset	$\mathcal{L}_{topo}$		$\mathcal{L}_{rec}$	BD *(%) ↑	BD(%)↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑
		$\mathcal{L}_{skel}$	$\mathcal{L}_{tpc}$	*Prob*					
SKA	BAS				51.64	69.69	63.93	87.62	81.20
				$HED_i, i \in \Omega$	49.07	68.43	61.04	88.30	83.91
				$Prob_i, i \in \Omega_U$	48.43	68.09	61.85	88.65	83.90
				$Prob_i, i \in \Omega$	58.65	75.82	70.55	88.79	87.60
		✓	✓	$Prob_i, i \in \Omega_U$	61.03	75.78	72.10	88.85	86.48
		✓		$Prob_i, i \in \Omega$	55.36	78.93	71.07	89.13	85.85
		✓	✓	$Prob_i, i \in \Omega$	62.90	77.96	73.89	89.50	88.50
VA	BAS				62.75	79.52	75.10	90.53	91.13
		✓	✓		64.48	80.57	76.33	90.29	91.62
	ATM22				78.44	90.22	88.49	91.61	94.38
		✓	✓		80.01	89.82	89.62	91.49	94.22

Table 7: **Robustness analysis of annotation shifts.** $\tau \in [0, 1]$ denotes the degree of offset, with larger τ values corresponding to greater annotation shifts away from the airway centerline.

τ	BD *(%) ↑	BD(%) ↑	TD(%)↑	DSC(%) ↑	TPR(%) ↑	FPR(%) ↓
@0.3	$62.93_{\pm 2.0}$	$77.64_{\pm 1.2}$	$73.63_{\pm 1.3}$	$89.13_{\pm 0.1}$	$88.62_{\pm 0.8}$	$.029_{\pm .004}$
@0.5	$62.88_{\pm 1.7}$	$78.15_{\pm 0.5}$	$74.09_{\pm 1.3}$	$89.13_{\pm 0.1}$	$88.46_{\pm 0.8}$	$.030_{\pm .003}$
@0.7	$62.22_{\pm 0.9}$	$77.25_{\pm 0.8}$	$73.10_{\pm 1.2}$	$89.12_{\pm 0.1}$	$86.58_{\pm 0.9}$	$.024_{\pm .002}$

topological guidance as a result of the accumulation of training samples with varied annotation preferences, including potential distal branch skeletons that were initially unannotated.

5.6.4. Analysis of computational efficiency.

The proposed model contains 4.13 million parameters. On 2 NVIDIA GeForce RTX 3090 GPUs, the model achieves an average inference time of 4 seconds per CT case. When performing sliding window inference with $96\times96\times96$ patches, the computational cost per patch is approximately 53.68 GFLOPs. These results indicate that the proposed method maintains a reasonable computational cost while achieving high segmentation accuracy, making it suitable for practical deployment in clinical settings.

6. Conclusion

Tailored to label-effective and topology-preserving airway segmentation, this paper proposes a novel skeleton annotation strategy (SKA) that not only enhances the preservation of topological integrity but also reduces annotation time. Its effectiveness and reliability have been validated through **clinical experiments**, highlighting its potential to streamline airway segmentation tasks. While SKA proves effective, current airway segmentation methods based on such sparse supervision (*i.e.,* SKA) face limitations. To address these, we further propose a reliable sparse supervision learning method, Skeleton2Mask, built upon SKA. This approach introduces a novel dual-stream buffer propagation strategy to facilitate reliable initial label diffusion, mitigating the issue of extremely sparse supervision and preventing the collapse of direct training. Aligning with the tree-like structure of the airway, its hierarchical geometry-aware learning explicitly integrates structural information at different levels through a tri-head supervision manner.

Extensive experiments demonstrate that our algorithm significantly outperforms other popular algorithms, particularly in topological metrics, achieving performance comparable to fully supervised methods while reducing standard costs by about 80%.

Nevertheless, the proposed approach has certain limitations that merit further investigation. The choice of hyperparameters in the DBP strategy (*i.e.,* δ and γ) is susceptible to the noise characteristics of the dataset, which directly determines the quality of the initial proposals. Since these proposals are fixed throughout training and serve as the main supervisory signal, suboptimal propagation may bias model optimization and hinder performance. This issue becomes especially important when adapting the method to other imaging contexts or elongated branching structures, where variations in image contrast, noise levels, and structural scale are more pronounced. To this end, in future work, we will focus on 1) integrating adaptive Gaussian geodesic distance calculation tailored to varying noise levels across images, aiming for more reliable foreground diffusion and global reconstruction; and 2) dynamically updating pseudo-labels by incorporating additional semantic features into the existing low-level features, with particular emphasis on maintaining the accuracy of distal branches.

Acknowledgments

This work is supported by Natural Science Foundation of China under Grant 62271465, Suzhou Basic Research Program under Grant SYG202338, and the National Natural Science Foundation of China (No. 82430065).

References

Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., 2010. Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence 33, 898–916.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics 38, 915–931.

Blum, H., 1967. A Transformation for Extracting New Descriptors of Shape, in: Models for the Perception of Speech and Visual Form, pp. 362–380.

Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y., 2023. 3D medical image segmentation with sparse annotation via cross-teaching between 3D and 2D networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 614–624.

Chen, Q., Hong, Y., 2022. Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–243.

Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al., 2023. SAM-Med2D. arXiv preprint arXiv:2308.16184 .

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 424–432.

Cui, Z., Zhang, B., Lian, C., Li, C., Yang, L., Wang, W., Zhu, M., Shen, D., 2021. Hierarchical morphology-guided tooth instance segmentation from CBCT images, in: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer. pp. 150–162.

Dang, V.N., Galati, F., Cortese, R., Di Giacomo, G., Marconetto, V., Mathur, P., Lekadir, K., Lorenzi, M., Prados, F., Zuluaga, M.A., 2022. Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation. Medical Image Analysis 75, 102263.

De Jong, A., Pardo, E., Rolle, A., Bodin-Lario, S., Pouzeratte, Y., Jaber, S., 2020. Airway management for COVID-19: a move towards universal videolaryngoscope? The Lancet Respiratory Medicine 8, 555.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. Advances in neural information processing systems 17.

Guo, J., Fu, R., Pan, L., Zheng, S., Huang, L., Zheng, B., He, B., 2022. Coarse-to-fine airway segmentation using multi information fusion network and CNN-based region growing. Computer Methods and Programs in Biomedicine 215, 106610.

Han, M., Luo, X., Liao, W., Zhang, S., Zhang, S., Wang, G., 2023. Scribble-based 3D multiple abdominal organ segmentation via triple-branch multi-dilated network with pixel-and class-wise consistency, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 33–42.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, in: International MICCAI brainlesion workshop, Springer. pp. 272–284.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G., 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental 4, 1–13.

Kim, B., Jeong, J., Han, D., Hwang, S.J., 2023. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11360–11370.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

Koziński, M., Mosinska, A., Salzmann, M., Fua, P., 2020. Tracing in 2D to reduce the annotation effort for 3D deep delineation of linear structures. Medical image analysis 60, 101590.

Lee, H., Jeong, W.K., 2020. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 14–23.

Lee, J.H., Kim, C., Sull, S., 2021. Weakly supervised segmentation of small buildings with point labels, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7406–7415.

Li, H., Xu, Z., Zhou, M., Shi, X., Kang, Y., Bu, Q., Lv, H., Li, M., Lin, M., Cui, L., et al., 2023a. Segment membranes and nuclei from histopathological images via nuclei point-level supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 539–548.

Li, S., Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y., 2022. PLN: Parasitic-like network for barely supervised medical image segmentation. IEEE Transactions on Medical Imaging 42, 582–593.

Li, W., Yuan, Y., Wang, S., Zhu, J., Li, J., Liu, J., Zhang, L., 2023b. Point2mask: Point-supervised panoptic segmentation via optimal transport, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 572–581.

Liang, Z., Wang, T., Zhang, X., Sun, J., Shen, J., 2022. Tree energy loss: Towards sparsely annotated semantic segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16907–16916.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3159–3167.

Liu, W., Guo, H., Zhang, W., Zang, Y., Wang, C., Li, J., 2022. TopoSeg: Topology-aware segmentation for point clouds., in: IJCAI, pp. 1201–1208.

Lo, P., Van Ginneken, B., Reinhardt, J.M., Yavarna, T., De Jong, P.A., Irving, B., Fetita, C., Ortner, M., Pinho, R., Sijbers, J., et al., 2012. Extraction of airways from CT (EXACT'09). IEEE Transactions on Medical Imaging 31, 2093–2107.

Lu, J., Deng, J., Zhang, T., 2024. BSNet: Box-supervised simulation-assisted mean teacher for 3d instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20374–20384.

Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S., 2022. Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 528–538.

Nadeem, S.A., Hoffman, E.A., Sieren, J.C., Comellas, A.P., Bhatt, S.P., Barjaktarevic, I.Z., Abtin, F., Saha, P.K., 2020. A CT-based automated algorithm for airway segmentation using freeze-and-grow propagation and deep learning. IEEE transactions on medical imaging 40, 405–418.

Qin, Y., Gu, Y., Zheng, H., Chen, M., Yang, J., Zhu, Y.M., 2020. 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. pp. 809–813.

Qin, Y., Zheng, H., Gu, Y., Huang, X., Yang, J., Wang, L., Yao, F., Zhu, Y.M., Yang, G.Z., 2021. Learning tubule-sensitive CNNs for pulmonary airway and artery-vein segmentation in CT. IEEE transactions on medical imaging 40, 1603–1617.

Ross, T.Y., Dollár, G., 2017. Focal loss for dense object detection, in: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2980–2988.

Selvan, R., Kipf, T., Welling, M., Juarez, A.G.U., Pedersen, J.H., Petersen, J., de Bruijne, M., 2020. Graph refinement based airway extraction using mean-field networks and graph neural networks. Medical image analysis 64, 101751.

Sethian, J.A., 1999. Fast marching methods. SIAM review 41, 199–235.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y., 2018. On regularized losses for weakly-supervised cnn segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 507–522.

Ulyanov, D., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 .

Wang, A., Tam, T.C.C., Poon, H.M., Yu, K.C., Lee, W.N., 2022a. Naviairway: a bronchiole-sensitive deep learning-based airway segmentation pipeline. arXiv preprint arXiv:2203.04294 .

Wang, A., Xu, M., Zhang, Y., Islam, M., Ren, H., 2023a. S^2ME: Spatial-spectral mutual teaching and ensemble learning for scribble-supervised polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 35–45.

Wang, D., Zhang, Z., Zhao, Z., Liu, Y., Chen, Y., Wang, L., 2022b. Pointscatter: Point set representation for tubular structure extraction, in: European conference on computer vision, Springer. pp. 366–383.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T.,

David, A.L., Deprest, J., Ourselin, S., et al., 2018. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence 41, 1559–1572.

Wang, T., Bai, Y., 2024. BAISeg: Boundary assisted weakly supervised instance segmentation. arXiv preprint arXiv:2406.18558 .

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2023b. BLPSeg: Balance the label preference in scribble-supervised semantic segmentation. IEEE Transactions on Image Processing .

Wei, J., Hu, Y., Li, G., Cui, S., Kevin Zhou, S., Li, Z., 2022. BoxPolyp: boost generalized polyp segmentation using extra coarse bounding box annotations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 67–77.

Wei, Y., Ji, S., 2021. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–12.

Wong, C.C., Vong, C.M., 2021. Persistent homology based graph convolution network for fine-grained 3D shape segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7098–7107.

Wu, L., Zhong, Z., Fang, L., He, X., Liu, Q., Ma, J., Chen, H., 2023. Sparsely annotated semantic segmentation with adaptive gaussian mixtures, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15454–15464.

Xiao, D., Chen, Z., Wu, S., Huang, K., Xu, J., Yang, L., Xu, Y., Zhang, X., Bai, C., Kang, J., et al., 2020. Prevalence and risk factors of small airway dysfunction, and association with smoking, in china: findings from a national cross-sectional study. The Lancet Respiratory Medicine 8, 1081–1093.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: ICCV, pp. 1395–1403.

Xie, Y., Zhou, T., Zhou, Y., Chen, G., 2024. SimTxtSeg: Weakly-supervised medical image segmentation with simple text cues, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 634–644.

Xu, Y., Xu, X., Jin, L., Gao, S., Goh, R.S.M., Ting, D.S., Liu, Y., 2021. Partially-supervised learning for vessel segmentation in ocular images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer. pp. 271–281.

Yao, T., Li, Y., Pan, Y., Mei, T., 2023. HGNet: Learning hierarchical geometry from points, edges, and surfaces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21846–21855.

Yin, X., Im, W., Min, D., Huo, Y., Pan, F., Yoon, S.E., 2024. Fine-grained background representation for weakly supervised semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology .

Yu, W., Zheng, H., Zhang, M., Zhang, H., Sun, J., Yang, J., 2022. Break: Bronchi reconstruction by geodesic transformation and skeleton embedding, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

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

Zhai, S., Wang, G., Luo, X., Yue, Q., Li, K., Zhang, S., 2023. PA-Seg: Learning from point annotations for 3D medical image segmentation using contextual regularization and cross knowledge distillation. IEEE transactions on medical imaging 42, 2235–2246.

Zhang, B., Yu, S., Wei, Y., Zhao, Y., Xiao, J., 2024. Frozen CLIP: A strong backbone for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3796–3806.

Zhang, K., Zhuang, X., 2022a. CycleMix: A holistic strategy for medical image segmentation from scribble supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11656–11665.

Zhang, K., Zhuang, X., 2022b. ShapePU: A new pu learning framework regularized by global consistency for scribble supervised cardiac segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 162–172.

Zhang, M., Gu, Y., 2023. Towards connectivity-aware pulmonary airway segmentation. IEEE Journal of Biomedical and Health Informatics .

Zhang, M., Wu, Y., Zhang, H., Qin, Y., Zheng, H., Tang, W., Arnold, C., Pei, C., Yu, P., Nan, Y., et al., 2023a. Multi-site, multi-domain airway tree modeling. Medical image analysis 90, 102957.

Zhang, M., Yu, X., Zhang, H., Zheng, H., Yu, W., Pan, H., Cai, X., Gu, Y., 2021. FDA: Feature decomposition and aggregation for robust airway segmentation, in: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3, Springer. pp. 25–34.

Zhang, X., Zhang, J., Ma, L., Xue, P., Hu, Y., Wu, D., Zhan, Y., Feng, J., Shen, D., 2022. Progressive deep segmentation of coronary artery via hierarchical topology learning, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 391–400.

Zhang, Z., Zhang, X., Qi, Y., Yang, G., 2023b. Partial vessels annotation-based coronary artery segmentation with self-training and prototype learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 297–306.

Zhao, T., Yin, Z., Wang, J., Gao, D., Chen, Y., Mao, Y., 2019. Bronchus segmentation and classification by neural networks and linear programming, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer. pp. 230–239.

Zheng, H., Qin, Y., Gu, Y., Xie, F., Sun, J., Yang, J., Yang, G.Z., 2021a. Refined local-imbalance-based weight for airway segmentation in CT, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 410–419.

Zheng, H., Qin, Y., Gu, Y., Xie, F., Yang, J., Sun, J., Yang, G.Z., 2021b. Alleviating class-wise gradient imbalance for pulmonary airway segmentation. IEEE transactions on medical imaging 40, 2452–2462.

Zhou, C., Xu, C., Cui, Z., 2023a. Progressive bayesian inference for scribble-supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3751–3759.

Zhou, M., Xu, Z., Zhou, K., Tong, R.K.y., 2023b. Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 137–147.