Leveraging Foundation Models, Knowledge Distillation, and Pseudo-Labeling for Robust Lung Segmentation in Computed Tomography Scans

Written Component for the UCLA Physics & Biology in Medicine Qualifying Examination

Jin Kim
Center for Computer Vision & Imaging Biomarkers

Committee:
Dr. Matthew S. Brown
Dr. Dan Ruan
Dr. John M. Hoffman

Abstract

Accurate segmentation of anatomical structures in Computed Tomography (CT) scans, particularly the lungs, is crucial for detecting and monitoring various pulmonary diseases. However, manual segmentation has several drawbacks, including being labor-intensive, subject to variability between different observers, and requiring a significant amount of time to complete. We propose a novel approach that combines the strengths of foundation models, knowledge distillation, and pseudo-labeling techniques for lung CT segmentation, while also considering the context of surrounding anatomical structures to enhance segmentation performance. Our approach leverages publicly available datasets to develop a robust foundation model using a dual-loss training approach, combining a cosine similarity loss between image and text embeddings and a distillation loss using MedSAM as a teacher model. Building upon this foundation model, we will generate high-resolution, multi-channel probability maps for various anatomical regions in lung CT scans using the model's zero-shot classification capabilities. These probability maps will be used to produce pseudo-segmentation labels for our in-house lung CT dataset, which will then be employed to fine-tune the MedSAM model using the H-SAM architecture. The fine-tuned model will be evaluated against state-of-the-art methods using in-house data, assessing its performance on both lung segmentation and the segmentation of other relevant anatomical structures. Our hypothesis is that the proposed approach, combining foundation models, knowledge distillation, and pseudo-labeling techniques, will result in a robust and efficient CT segmentation model that outperforms current state-of-the-art methods, demonstrating its potential for clinical translation and advancing the field of pulmonary medicine by providing a comprehensive understanding of the anatomical structures within lung CT scans.

Specific Aims

Accurate segmentation of anatomical structures in Computed Tomography (CT) scans is crucial for detecting and monitoring various pulmonary diseases. However, manual segmentation is not only tedious and exhausting but also susceptible to variations in results due to differences in perception among those performing the task. Deep learning-based segmentation models, while powerful, often face limitations in medical applications due to the scarcity of sizable, annotated datasets required for effective training. Foundation models like MedSAM have shown promise in adapting to new tasks with limited data, but their performance on lung CT segmentation and the segmentation of other relevant anatomical structures remains underexplored. We hypothesize that by leveraging public datasets and pseudo-labeling techniques, we can develop a robust and efficient CT segmentation model that outperforms current state-of-the-art methods, providing a comprehensive understanding of the anatomical structures within lung CT scans.

SA-1: Develop a foundation model for lung CT segmentation by leveraging public datasets and knowledge distillation from MedSAM.

We will curate a comprehensive public dataset of 60,000 lung CT images and associated medical reports from various sources, including the Zenodo and NIH. The dataset will be pre-processed, ensuring uniformity and compatibility across all images. Using this data, we will build a CLIP-like model with a dual-loss training approach. The first loss will be a cosine similarity loss between image and text embeddings, encouraging the model to learn meaningful representations. The second loss will be a distillation loss utilizing MedSAM as the teacher model to guide the learning of the image encoder. The hypothesis for this aim is that the combination of CLIP-style training and MedSAM-assisted distillation, along with a diverse and well-curated dataset, will result in a foundation model with strong generalization capabilities for lung CT segmentation.

SA-2: Generate high-resolution probability maps for lung regions using the foundation model's zero-shot classification capabilities.

Building upon the foundation model developed in SA-1, we will create an inference pipeline to generate high-resolution, multi-channel probability maps for various anatomical regions in lung CT scans. This pipeline will leverage the zero-shot classification capabilities of the foundation model and a patch-based approach, where input CT images are divided into small, overlapping patches. Each patch will be processed by the image encoder to generate patch-level embeddings, which will be compared with text embeddings of anatomical regions using cosine similarity to assign probability scores. The probability maps will be refined through post-processing techniques, including normalization, Gaussian smoothing, and heuristic thresholding. In parallel, we will collect an in-house dataset of 250 lung CT scans, which will undergo quality control and pre-processing steps to ensure data consistency and reliability. The hypothesis for this aim is that the zero-shot classification capabilities of the foundation model, combined with the patch-based processing approach, can effectively generate high-quality probability maps for lung regions and other anatomical structures, even on unseen in-house data.

SA-3: Produce pseudo-segmentations using the probability maps, fine-tune MedSAM, and evaluate on expert-annotated in-house data.

We will generate pseudo-segmentation labels for our in-house lung CT dataset by combining MedSAM predictions guided by the probability maps from SA-2. The pseudo-labeled dataset will be augmented using various techniques to introduce more variability. The MedSAM model will be fine-tuned using the H-SAM architecture, which includes a LoRA-adapted image encoder and a hierarchical pixel decoder. To evaluate the fine-tuned model, we will use a held-out portion of the in-house dataset with expert-annotated segmentation labels. The model's performance will be assessed using various metrics and benchmarked against state-of-the-art methods, including the original MedSAM, nnU-Net, and SAM. The hypothesis is that the pseudo-labeling approach, combined with MedSAM fine-tuning on diverse in-house data, will result in a model that outperforms current methods, demonstrating its robustness and generalizability.

Background & Significance

Respiratory diseases, such as chronic obstructive pulmonary disease (COPD), lung cancer, and interstitial lung diseases, persist as significant contributors to global health burdens and deaths[7]. The World Health Organization (WHO) revealed that COPD was the third most common cause of mortality worldwide, responsible for around

3.2 million deaths[8]. Lung cancer continues to be the most fatal cancer, with projections of 2.5 million new diagnoses and 1.8 million deaths in 2022[9]. The current COVID-19 pandemic has further emphasized the crucial need for timely identification and treatment of lung diseases, as pre-existing respiratory conditions are linked to a higher likelihood of severe disease and death[10]. Computed tomography (CT) is the primary imaging modality for diagnosing and monitoring lung diseases, providing high-resolution visualization of the lung parenchyma, airways, and vasculature[11]. However, the interpretation of lung CT scans is a complex and time-consuming task that requires expertise in radiology. The increasing volume of CT scans performed in clinical practice, coupled with the shortage of trained radiologists in many parts of the world, has led to a substantial burden on healthcare systems[12].

Accurate segmentation of lung regions in CT scans is a critical step in the quantitative analysis of lung diseases. It enables the extraction of clinically relevant features, such as lung volumes, densities, and textures, which can aid in the diagnosis, staging, and monitoring of various lung conditions[13]. However, manual segmentation is a tedious and labor-intensive process that is prone to inter-observer variability[14]. Recent advances in deep learning have shown promising results in automating lung CT segmentation, with convolutional neural networks (CNNs) achieving state-of-the-art performance[15]. Despite the promising results of these models, their effectiveness is largely dependent on access to extensive, labeled datasets for training purposes, which can be difficult to acquire in the healthcare field[16]. The acquisition of annotated medical imaging data is a significant bottleneck in the development of deep learning models for lung CT segmentation. Manual annotation requires extensive time and expertise from trained radiologists, making it difficult to scale up the dataset size[17]. Moreover, patient privacy concerns and institutional data sharing policies often limit the accessibility of medical imaging data[18]. As a result, many existing deep learning models for lung CT segmentation are trained on relatively small, single-institution datasets, which may limit their generalizability to diverse patient populations and scanning protocols[19].

To address the limitations of current deep learning approaches, foundation models, such as Segment Anything Model (SAM)[20] or MedSAM[21], have emerged as a promising solution. Foundation models are large, pre-trained models that capture rich representations of medical images and can be adapted to various downstream tasks with limited fine-tuning data[22]. MedSAM, in particular, has demonstrated impressive performance in several medical image segmentation tasks, including brain tumor and liver segmentation. However, its potential for lung CT segmentation remains largely unexplored. A recent study[23] evaluated the performance of SAM on various medical image segmentation tasks, finding that its accuracy was significantly lower than state-of-the-art algorithms specifically designed for medical images. The study also identified several factors that may affect SAM's accuracy in medical images, such as segmentation difficulty, image dimension, target region size, and contrast. The adaptation of foundation models to lung CT segmentation faces several challenges. First, the domain shift between the pre-training and target datasets may limit the transferability of learned features[24]. Second, the fine-tuning of foundation models on limited annotated data may lead to overfitting and poor generalization[25]. Third, the high computational requirements of foundation models may hinder their deployment in clinical settings with limited resources[26].

To address these challenges, we propose a novel approach that combines the strengths of foundation models, knowledge distillation[27], and pseudo-labeling techniques[28] for lung CT segmentation, while also exploring the potential benefits of incorporating surrounding anatomical structures to enhance lung segmentation accuracy. Our approach leverages publicly available datasets, such as the NIH DeepLesion dataset[29], and RAD-ChestCT[30], to mitigate the need for extensive in-house data collection and annotation. The NIH DeepLesion dataset consists of over 32,000 annotated lesions identified on CT images from 4,400 unique patients. The RAD-ChestCT dataset, provided by Zenodo, includes 35,747 chest CT scans from 19,661 adult patients, with each CT volume annotated with a matrix of 84 abnormality labels and 52 location labels. These datasets contain a diverse range of lung CT scans from multiple institutions and patient populations, enabling the development of robust and generalizable models.

We will employ knowledge distillation techniques to transfer the learned representations from MedSAM to a task-specific lung CT segmentation model. Knowledge distillation is a process in which a smaller student model is trained to mimic the behavior of a larger teacher model. By distilling the knowledge from MedSAM, we can leverage its pre-trained features while reducing the computational requirements and improving the efficiency of the segmentation model[31]. This approach has the potential to overcome the domain shift and limited data challenges associated with foundation model adaptation. Furthermore, we will explore the use of pseudo-labeling techniques to generate high-quality probability maps for lung regions using the foundation model's zero-shot classification capabilities. Pseudo-labeling is a semi-supervised learning approach in which a model's predictions on unlabeled data are used as training labels for subsequent iterations. By generating pseudo-labels for our in-house dataset, we can significantly expand the training data and improve the model's performance on unseen data[32]. This approach can also help in reducing the annotation burden and the reliance on expert-annotated

data.

The proposed lung CT segmentation model will be rigorously evaluated against state-of-the-art methods, such as nnU-Net[33] and UNETR[34], using expert-annotated in-house data. nnU-Net is a self-configuring deep learning framework that has shown excellent performance in various medical image segmentation tasks. UNETR is a transformer-based architecture that has achieved state-of-the-art performance in several medical image segmentation benchmarks. By comparing our model's performance against these established methods, we can demonstrate its effectiveness and potential for clinical translation. The successful development of a robust and efficient lung CT segmentation model using foundation models, knowledge distillation, and pseudo-labeling techniques can have significant implications for the diagnosis and management of lung diseases. The model can be integrated into clinical workflows to assist radiologists and pulmonologists in the accurate and timely assessment of lung conditions. It can also facilitate the development of quantitative imaging biomarkers for disease progression and treatment response monitoring. Moreover, the proposed approach can serve as a paradigm for adapting foundation models to other medical image segmentation tasks, such as cardiac and abdominal segmentation. The combination of publicly available datasets, knowledge distillation, and pseudo-labeling techniques can enable the rapid development of high-performance segmentation models in various clinical domains, ultimately improving patient care and advancing the field of medical image analysis.

Research Design and Methods

SA-1: Develop a foundation model for lung CT segmentation using public datasets and knowledge distillation from MedSAM

Dataset Curation and Preprocessing

To develop a robust foundation model for lung CT segmentation, we will curate a comprehensive public dataset from two primary sources: the NIH DeepLesion dataset[29] and the RAD-ChestCT dataset[30]. The NIH DeepLesion dataset consists of over 32,120 slices with annotated lesions from 4,427 unique patients, while the RAD-ChestCT dataset, provided by Zenodo, includes 35,747 chest CT scans from 19,661 adult patients, with each CT volume annotated with a matrix of 84 abnormality labels and 52 location labels. By combining these datasets, we can leverage a diverse range of approximately 60,000 lung CT scans and their corresponding medical reports to train our foundation model, ensuring its generalizability across multiple institutions and patient populations. The choice of 60,000 CT scans is based on a finding[1] demonstrated that using a subset as small as 5% of the ImageNet dataset (approximately 700,000 images) was sufficient for pre-training without compromising transfer performance. Given that the number of slices per CT scan varies from 3 to 600, with the NIH DeepLesion dataset having around 3 slices on average and the RAD-ChestCT dataset having 500 slices on average[35], we estimate that our curated dataset of 60,000 CT scans will yield approximately 700,000 slices after trimming, which aligns with the recommended subset size for effective pre-training.

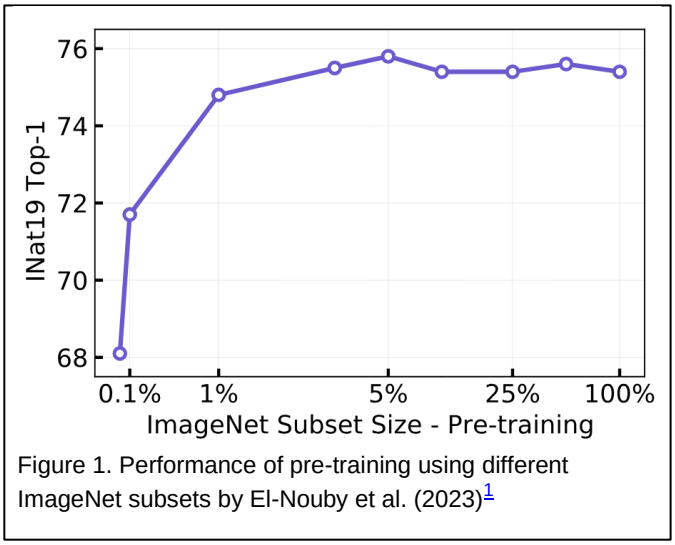

Figure 1. Performance of pre-training using different ImageNet subsets by El-Nouby et al. (2023)[1]

To ensure the quality and relevance of the data used for training, we will perform a trimming process on the CT scans. As the middle slices of a CT scan typically contain the most important information about the patient's lungs, we will focus on these slices and discard those that do not contain significant information. Specifically, we will trim the RAD-ChestCT dataset CT scans down to 20 slices per scan, effectively reducing the total number of slices by 96%. This trimming process will not only help us focus on the most informative slices but also reduce computational requirements during the pre-training phase.

We will perform several preprocessing steps to ensure uniformity and compatibility across the collected data. First, all CT scans will be converted to a standardized format, such as NIfTI, to maintain consistency. Next, we will resize the images to a fixed resolution of 512x512 pixels to facilitate efficient processing by our model. The CT scans will then undergo intensity normalization to ensure that the pixel values fall within a consistent range,

typically [-1000, 400] Hounsfield Units (HU), which covers the range of lung tissue densities. Furthermore, we will preprocess the associated medical reports and annotations to extract relevant information by tokenizing the text, removing stop words, and applying lemmatization to normalize the data. The processed reports and annotations will be used as input to the text encoder in our CLIP-like model. By curating a diverse and well-preprocessed dataset from the NIH DeepLesion and RAD-ChestCT sources, and by applying a trimming process to focus on the most informative slices, we aim to provide our foundation model with a comprehensive and relevant set of lung CT images and their corresponding medical reports and annotations, enabling the model to learn robust and generalizable features for lung CT segmentation, laying a strong foundation for the subsequent stages of our research.

Model Architecture and Training

Our proposed foundation model for lung CT segmentation will adopt a CLIP[36]-like structure, consisting of an image encoder and a text encoder. The image encoder will be based on either CNN architectures, such as ResNet[37], or Vision Transformer(ViT)[38], pretrained on a large-scale dataset like ImageNet[39]. For the text encoder, we will employ a state-of-the-art language model, specifically LLaMA[40], which has shown impressive performance on various natural language processing tasks. The two encoders will be jointly trained using a dual-loss approach, combining a cosine similarity loss between image and text embeddings and a distillation loss using MedSAM as a teacher model.

The cosine similarity loss will be calculated based on the difference between the image embedding generated by the image encoder and the text embedding generated by the text encoder. This loss will encourage the model to learn meaningful representations that align the visual and textual information. The distillation loss, on the other hand, will be computed as the difference between the embedding from the image encoder and the embedding from the MedSAM encoder. By minimizing this loss, we will force our image encoder to output embeddings similar to those produced by MedSAM, effectively transferring the knowledge from the pre-trained MedSAM model to our foundation model. This teacher-student-like approach will enable our model to benefit from the rich representations learned by MedSAM while being specifically tailored for lung CT segmentation.

The training process will involve fine-tuning both the image encoder and the text encoder on our curated dataset of lung CT images and their associated medical reports. We will employ techniques such as data augmentation, including random cropping, flipping, and rotation, to enhance the model's robustness and generalization ability. The model will be trained using an optimization algorithm like Adam[41] with a learning rate scheduler to ensure stable convergence. We will monitor the model's performance on a validation set during training and employ early stopping to prevent overfitting. By leveraging the dual-loss approach and the knowledge distillation from MedSAM, our foundation model will learn to effectively capture the relationship between lung CT images and their corresponding medical reports, providing a strong basis for the subsequent segmentation tasks.

Dual-Loss Training Approach: Combining CLIP-Style and MedSAM-Assisted Distillation

To effectively train our foundation model for lung CT segmentation, we propose a dual-loss training approach that combines a cosine similarity loss between image and text embeddings, inspired by the CLIP method, and a distillation loss using MedSAM as the teacher model to guide the learning of the image encoder. This unique combination of losses enables our model to learn meaningful representations that align visual and textual information while benefiting from the rich representations learned by the pre-trained MedSAM model.

The cosine similarity loss, as employed in the CLIP approach, encourages the model to learn a multi-modal embedding space where related image and text pairs are mapped closely together. Given an image embedding I_e generated by the image encoder and the corresponding text embedding T_e produced by the text encoder, the cosine similarity loss is calculated as follows:

$$L_{cos} = 1 - \frac{I_e \cdot T_e}{||I_e|| \cdot ||T_e||} \qquad \textbf{Eq. 1}$$

where $I_e \cdot T_e$ represents the dot product between the image and text embeddings, and I_e and T_e denote the Euclidean norms of the respective embeddings. By minimizing this loss, our model learns to associate visual patterns in lung CT images with relevant textual descriptions from the associated medical reports. This alignment of visual and textual representations enhances the model's ability to capture semantically meaningful features that are relevant to lung CT segmentation.

Concurrently, we employ a distillation loss that leverages MedSAM as a teacher model to guide the learning of our image encoder. Let I_s denote the embedding generated by our image encoder (student) and I_t represent the embedding produced by the MedSAM encoder (teacher) for the same input image. The distillation loss is computed as the mean squared error between these embeddings:

$$L_{distill} = \frac{1}{n}\sum_{i=1}^{n}\left(I_s^{(i)} - I_t^{(i)}\right)^2 \qquad \textbf{Eq. 2}$$

where n is the number of samples in the batch, and $I_s^{(i)}$ and $I_t^{(i)}$ are the student and teacher embeddings for the i-th sample, respectively. By minimizing this loss, we encourage our image encoder to output embeddings that are similar to those learned by MedSAM, effectively transferring the knowledge from the pre-trained MedSAM model to our foundation model. This teacher-student-like approach allows our model to benefit from the rich representations learned by MedSAM while being specifically tailored for the task of lung CT segmentation.

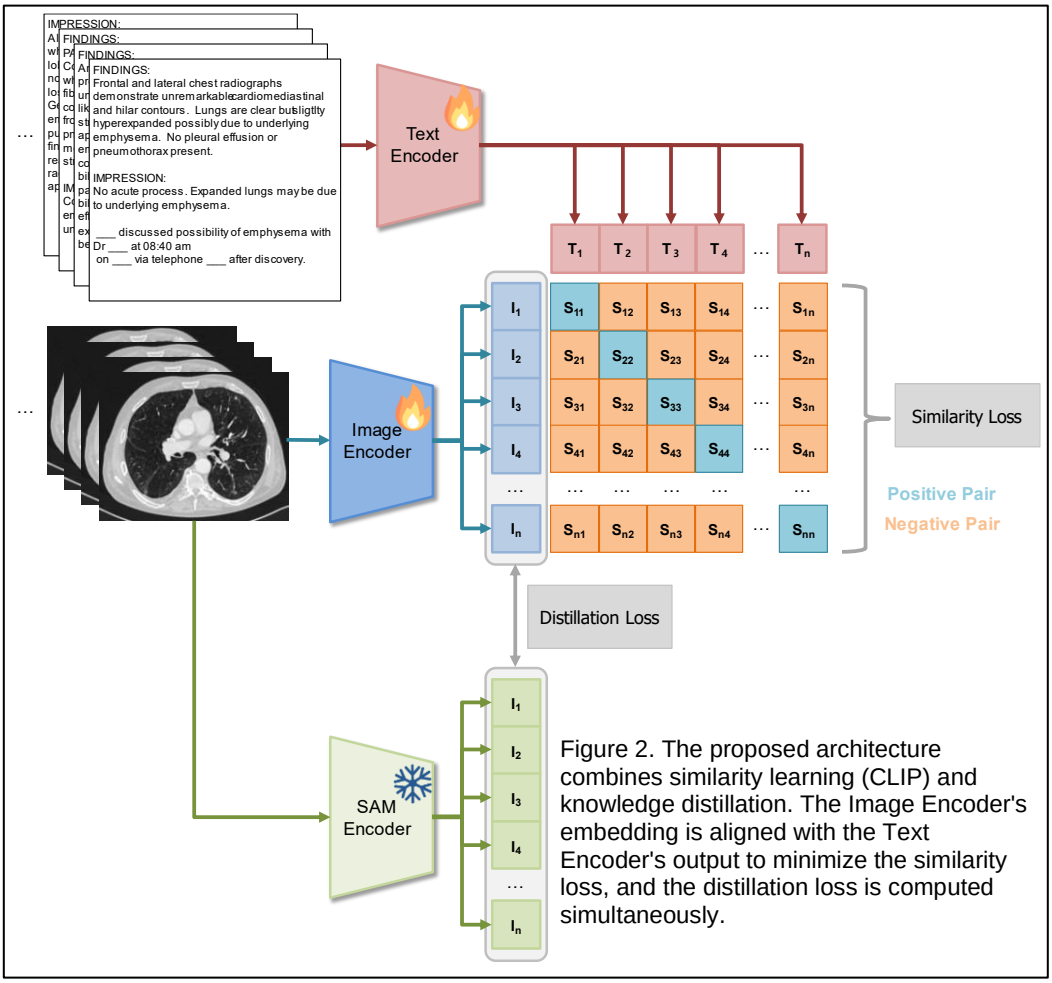

Figure 2. The proposed architecture combines similarity learning (CLIP) and knowledge distillation. The Image Encoder's embedding is aligned with the Text Encoder's output to minimize the similarity loss, and the distillation loss is computed simultaneously.

The distillation process is inspired by the DeiT-LT approach[27], which demonstrates the effectiveness of distilling knowledge from a CNN teacher to a ViT student in the context of long-tailed recognition. In our case, we adapt this technique to distill knowledge from the MedSAM model to our foundation model's image encoder. To enhance the student model's ability to learn generalizable features that are robust to variations in lung CT scans, we pass out-of-distribution (OOD) images through the teacher model during distillation. These OOD images are generated by applying strong augmentations, such as RandAugment[42], to the original lung CT images. The augmented images are then fed to the MedSAM teacher model to obtain the corresponding embeddings, which serve as the targets for the distillation loss. By learning to mimic the teacher's behavior on these OOD samples, our foundation model becomes more resilient to potential variations and artifacts in real-world lung CT scans.

The overall training objective of our foundation model is a weighted combination of the cosine similarity loss and the distillation loss, defined in **Eq.1** and **Eq.2**:

$$L_{total} = \alpha L_{cos} + \beta L_{distill} \qquad \textbf{Eq. 3}$$

where α and β are hyperparameters that control the relative importance of each loss term. By optimizing this dual-loss objective, our foundation model learns a rich and informative representation of lung CT images that captures both local and global features crucial for accurate segmentation. The cosine similarity loss ensures that the learned features are aligned with relevant textual descriptions, facilitating the association between visual patterns and their semantic meaning. Meanwhile, the distillation loss leverages the knowledge encoded in the

MedSAM model, allowing our foundation model to benefit from its pre-trained representations and adapt them specifically for lung CT segmentation. This synergistic combination of losses empowers our model to achieve robust and reliable performance on the challenging task of lung region segmentation in CT scans.

SA-2: Generate high-resolution lung probability maps using the foundation model's zero-shot classification capabilities

Patch-based Probability Map Generation with Multi-Class Prompts

Building upon the foundation model developed in SA-1, we will create an inference pipeline to generate high-resolution, multi-channel probability maps for various anatomical regions in lung CT scans. This pipeline will leverage the zero-shot classification capabilities of the foundation model, enabling it to effectively identify and localize different anatomical structures without the need for additional training or fine-tuning. Our approach draws inspiration from patch-based object localization techniques, where an image is divided into small patches, and each patch is assigned a probability score based on its similarity to a given set of text prompts. To generate multi-channel probability maps, we will preprocess the input lung CT images by dividing each CT slice into a grid of small, overlapping patches. The size of these patches will be carefully chosen to balance the trade-off between spatial resolution and computational efficiency. Smaller patch sizes will capture more fine-grained details but may increase processing time, while larger patch sizes will be more computationally efficient but may sacrifice some localization accuracy. We will systematically evaluate different patch sizes to determine the optimal configuration for our specific application.

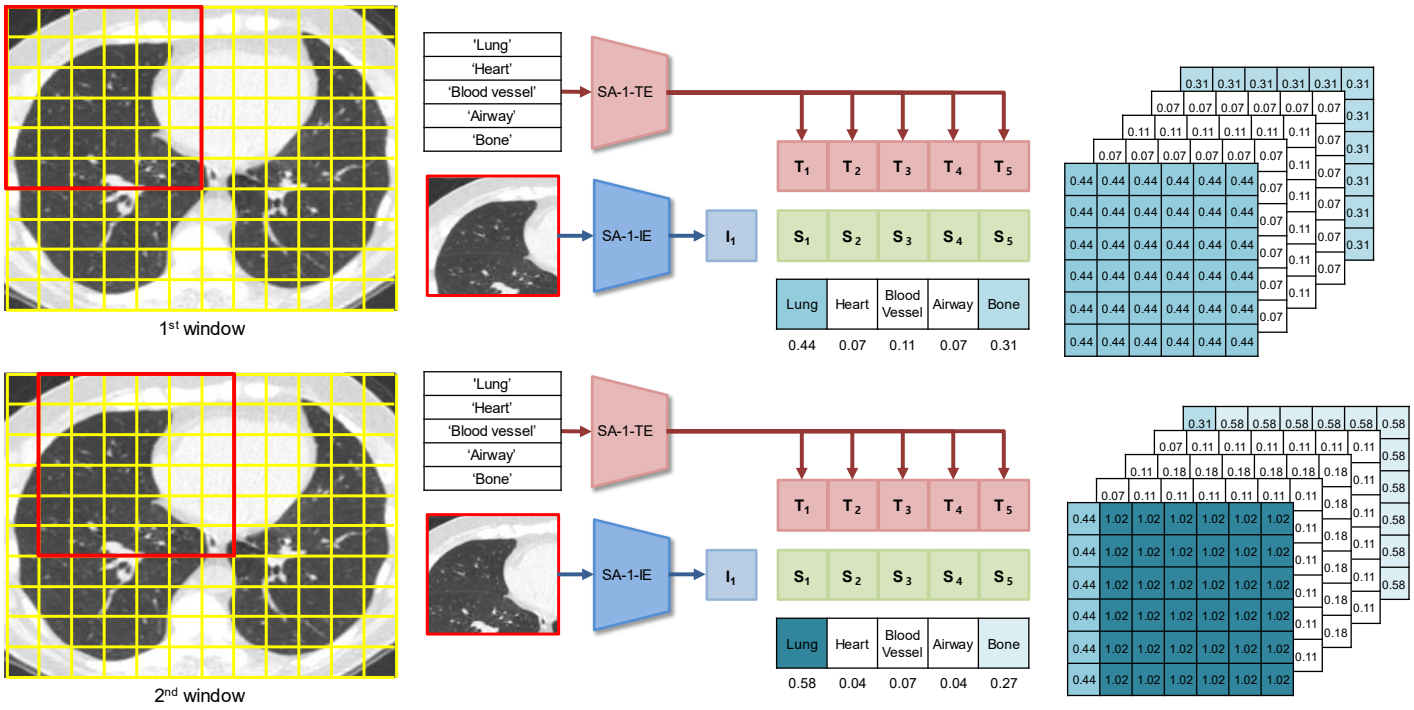

Figure 3. Illustration of the sliding window approach for patch-based probability map generation. The window slides over the image with a smaller stride than the patch size, generating overlapping patches. Each patch is processed by the foundation model to obtain probability scores for each anatomical class.

Each patch will be passed through the image encoder of the foundation model, which was trained using the dual-loss approach combining CLIP-style cosine similarity and MedSAM-assisted distillation (as described in SA-1). The image encoder will generate a patch-level embedding, denoted as I_e, which captures the salient features and visual patterns within the patch. Simultaneously, we will input a set of carefully crafted text prompts, such as "lung", "heart", "blood vessel", "airway", and "bone", to the text encoder of the foundation model. The text encoder will process these prompts and generate corresponding text embeddings, denoted as $T_{e_1}, T_{e_2}, \ldots, T_{e_n}$, where n is the number of anatomical classes (in this case, $n = 5$). These text embeddings encode the semantic meaning of each anatomical region. To assess the similarity between each patch and the

different anatomical regions, we will calculate the cosine similarity between the patch embedding I_e and each of the text embeddings $T_{e_1}, T_{e_2}, \ldots, T_{e_n}$. The cosine similarity is computed as follows:

$$\text{similarity}_i = \frac{I_e \cdot T_{e_i}}{\left||I_e|\right| \cdot \left|\left|T_{e_i}\right|\right|} \quad \textbf{Eq. 4}$$

where $I_e \cdot T_{e_i}$ represents the dot product between the patch embedding and the i-th text embedding, and $\left||I_e|\right|$ and $\left|\left|T_{e_i}\right|\right|$ denote the Euclidean norms of the respective embeddings. The similarity scores obtained will fall within the range of -1 to 1, with higher values indicating a stronger alignment between the visual content of the patch and the semantic meaning of the corresponding anatomical region. We will assign the computed similarity scores as the probability values for each patch, indicating the likelihood of it belonging to each of the anatomical regions. This process will result in a multi-channel probability map, where each channel corresponds to a specific anatomical class (lung, heart, blood vessel, airway, and bone). While the multi-channel probability map provides valuable contextual information about the surrounding anatomical structures, our primary focus is on the lung channel for binary segmentation. In the next subsection, we will extract the lung channel from the multi-channel probability map and apply refinement techniques to obtain a binary lung segmentation mask.

To generate high-resolution probability maps, we will process the patches using a sliding window approach, as illustrated in **Figure 3**. The window will move across the CT slice with a stride smaller than the patch size, ensuring that we capture fine-grained spatial information and generate smooth, continuous probability maps for each anatomical class. The detailed process is outlined in the provided **Algorithm 1**:

Algorithm 1 Multi-class Patch-based Probability Map Generation

Input:
I: Input lung CT slice of size (*h,w,c*)
T: Set of aligned text prompts (e.g., ”lung”, ”heart”, ”blood vessel”, ”airway”, ”bone”)
image encoder: Foundation model’s image encoder (e.g., ResNet or ViT)
text encoder: Foundation model’s text encoder (e.g., LLaMA)
W_i: Learned projection matrix for image embeddings
W_t: Learned projection matrix for text embeddings
t: Learned temperature parameter
patch size: Size of each patch
stride: Stride for sliding the window
Output:
prob map: Multi-channel probability map of size (*h,w,n*)
Initialization:
Initialize an empty multi-channel probability map (prob map) of size (*h,w,n*)

```
for i = 0 to h − patch size + 1 with step size stride do
    for j = 0 to w – patch size + 1 with step size stride do
        Extract the current patch from the input CT slice
        Generate patch embedding using the image encoder
        Project patch embedding to the joint embedding space using W_i
        Generate text embeddings for each prompt using the text encoder
        Project text embeddings to the joint embedding space using W_t
        Compute cosine similarities between the patch embedding and text embeddings
        Assign the computed similarities to the corresponding channels of prob map
    end for
end for
Normalize prob map by the number of overlapping patches at each pixel location
return prob map
```

As the window slides across the CT slice, the probability scores of overlapping patches for each anatomical class are aggregated by taking the average or maximum value at each pixel location. This aggregation step helps to reduce noise and promote spatial coherence in the resulting probability maps. The final output is a set of high-resolution, multi-channel probability maps, where each pixel value in a specific channel represents the likelihood of that pixel belonging to the corresponding anatomical region.

One of the key advantages of our multi-class patch-based approach is its ability to generate fine-grained, anatomically-aware probability maps. By incorporating multiple text prompts corresponding to different anatomical structures, we can obtain probability maps that highlight the relevant regions for each class. This multi-channel representation provides a rich and informative basis for subsequent segmentation and analysis tasks, enabling a more comprehensive understanding of the lung CT scans. Furthermore, our approach leverages the zero-shot classification capabilities of the foundation model, allowing it to generate probability maps for multiple anatomical regions without the need for additional training or fine-tuning. This is made possible by the dual-loss training approach employed in SA-1, which aligns the visual and textual embeddings in a shared multi-modal space. By exploiting this alignment, we can efficiently generate probability maps for various anatomical structures using only a set of text prompts, without requiring explicit annotations for each region in the CT scans.

To evaluate the effectiveness of our multi-class patch-based probability map generation pipeline, we will conduct extensive experiments on a diverse range of lung CT scans. We will assess the quality and accuracy of the generated probability maps using both quantitative and qualitative metrics. Quantitative evaluation will involve comparing the generated probability maps against ground truth segmentations, when available, using metrics such as Dice similarity coefficient (DSC) and Intersection over Union (IoU). Qualitative assessment will involve visual inspection of the probability maps by expert radiologists to evaluate their coherence, granularity, and alignment with anatomical structures. We will also investigate the impact of various hyperparameters, such as patch size, stride, and the choice of text prompts, on the performance of our pipeline. Through rigorous experimentation and analysis, we aim to demonstrate the robustness, versatility, and clinical applicability of our approach.

Post-processing Techniques for Probability Map Refinement

To further optimize the accuracy and quality of the generated lung probability map, we will investigate and implement various post-processing techniques. Although our patch-based approach generates multi-channel probability maps for various anatomical regions, we will focus specifically on the lung channel for binary segmentation. By extracting the lung channel from the multi-channel probability map, we can effectively refine the lung region probability map while maintaining the contextual information provided by the other anatomical structures. These post-processing techniques aim to refine the lung probability map by reducing noise, improving spatial coherence, and emphasizing the most relevant regions. By carefully designing and integrating these post-processing steps into our pipeline, we can significantly improve the localization accuracy and robustness of our lung region segmentation approach, described in **Figure 4**.

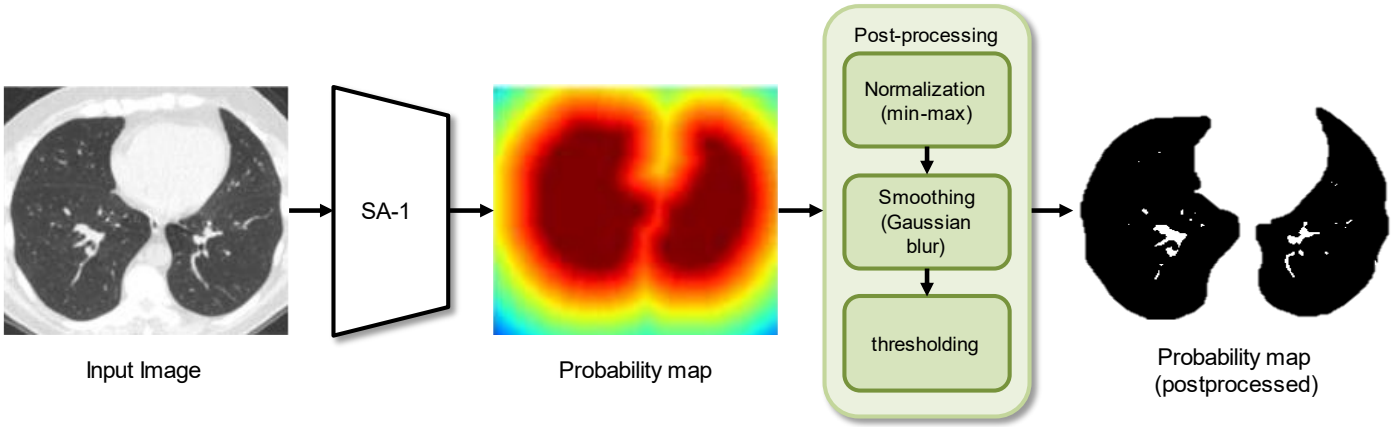

Figure 4. Overview of the probability map generation and post-processing pipeline. The input lung CT image is fed into the SA-1 model, which generates a high-resolution probability map using the foundation model's zero-shot classification capabilities. The generated probability map undergoes a series of post-processing techniques, including normalization, spatial smoothing, and heuristic thresholding, converting it from a continuous heatmap representation to a discrete binary segmentation mask.

One of the primary post-processing techniques we will explore is normalization. The probability maps generated by our patch-based approach may exhibit variations in the range and distribution of probability values across different CT scans. To ensure consistency and comparability, we will apply normalization techniques to scale the probability values to a common range, such as [0, 1]. This normalization step will involve analyzing the histogram of probability values within each map and applying suitable transformations, such as min-max scaling or sigmoid activation. Min-max scaling linearly transforms the probability values to the desired range using the following formula:

$$P_{norm} = \frac{P - min(P)}{max(P) - min(P)} \qquad \textbf{Eq. 5}$$

where P represents the original probability values, and P_{norm} denotes the normalized probability values. Alternatively, sigmoid activation can be used to squash the probability values to the range [0, 1] using the logistic function:

$$P_{norm} = \frac{1}{1 + e^{-\alpha(P-\beta)}} \qquad \textbf{Eq. 6}$$

where α and β are parameters that control the steepness and the center of the sigmoid curve, respectively. By normalizing the probability maps, we can establish a consistent interpretation of the probability values and facilitate the application of subsequent post-processing steps and thresholding operations.

The probability maps obtained from the patch-based approach may contain local noise and irregularities due to the granularity of the patches and the sliding window operation. To mitigate these issues and promote spatial coherence, we will employ Gaussian smoothing techniques. Gaussian smoothing involves convolving the probability map with a 2D Gaussian kernel, which effectively blurs the map and suppresses high-frequency noise. The 2D Gaussian kernel is defined as:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad \textbf{Eq. 7}$$

where x and y represent the spatial coordinates, and σ is the standard deviation of the Gaussian distribution. The size and standard deviation of the Gaussian kernel will be carefully tuned to strike a balance between noise reduction and preservation of important spatial details. Additionally, we will investigate the use of edge-preserving smoothing techniques, such as bilateral filtering or guided filtering, which can smooth the probability map while maintaining sharp transitions at the boundaries of the lung regions. By applying these smoothing techniques, we can obtain a more robust and spatially consistent representation of the lung regions in the probability maps.

We will develop and evaluate various heuristic thresholding strategies to binarize the refined probability maps into lung and non-lung regions. Thresholding is a crucial step in converting the continuous probability values into discrete segmentation masks. We will explore different thresholding approaches, including adaptive thresholding, multi-level thresholding, and global thresholding. Global thresholding involves selecting a single probability threshold τ and classifying all pixels above the threshold as lung regions and those below as non-lung regions:

$$Mask_{binary}(x,y) = \begin{cases} 1, & P(x,y) \geq \tau \\ 0, & otherwise \end{cases} \quad \textbf{Eq. 8}$$

where $Mask_{binary}(x,y)$ represents the binary segmentation mask, and $P(x,y)$ denotes the probability value at the pixel location (x,y). Adaptive thresholding, on the other hand, takes into account the local characteristics of the probability map and determines threshold values dynamically based on the neighborhood of each pixel. One common approach is to use the mean or median of the local neighborhood as the threshold value. Multi-level thresholding extends this concept by defining multiple probability thresholds and generating a hierarchical segmentation of the lung regions. By carefully selecting and fine-tuning these thresholding strategies, we can optimize the trade-off between sensitivity and specificity in detecting lung regions and achieve a highly accurate binary segmentation mask.

In-House Data Collection for Pre-processing for Lung CT Segmentation

To create a diverse and high-quality in-house dataset for validating and refining our lung CT segmentation framework, we will collect a retrospective cohort of 250 lung CT scans from patients who have undergone chest CT imaging for various clinical indications. This sample size was determined based on the findings of recent studies demonstrating the importance of training data diversity for robust lung segmentation performance. For example, a study[15] showed that a diverse dataset of 231 cases outperformed models trained on public datasets, achieving a mean Dice score of 0.98 compared to 0.94 for a model trained on the Lung Tissue Research Consortium dataset. Similarly, another study[43] used a multi-institutional dataset of 929 CT scans and achieved mean Dice scores of 0.985. While a larger dataset was used in the latter study, the difference in performance compared to the smaller diverse dataset was minimal, suggesting that data diversity plays a crucial role in achieving high segmentation accuracy. Given that our institution's patient population encompasses a wide range of lung appearances and pathologies, we believe that a sample size of 250 scans will capture substantial variability while remaining manageable for the scope of this study. By collecting this diverse dataset, we aim to provide a rigorous test of our segmentation framework's robustness and generalizability. The collected CT scans will be stored in the Digital Imaging and Communications in Medicine (DICOM) format, which includes complete header information. The collected CT scans will undergo a streamlined quality control and pre-processing pipeline to ensure data consistency, reliability, and optimal

performance of our segmentation algorithms, as illustrated in **Figure 5**.

The first step in our pipeline is a quality control check, where the collected CT scans will be assessed for their suitability for further analysis. This quality control step will involve a combination of automated checks and manual verification by experienced radiologists. The automated checks will screen for common issues such as incomplete or corrupted DICOM files, insufficient coverage of the lung region, and excessive noise or artifacts. Scans that pass the initial automated checks will then undergo a rapid manual verification to identify any remaining issues that may have been missed. This two-step quality control process allows us to efficiently filter out problematic scans while maintaining a high level of data quality. Scans that fail the quality control checks will be excluded from further analysis and dumped from the pipeline to maintain the integrity of the dataset.

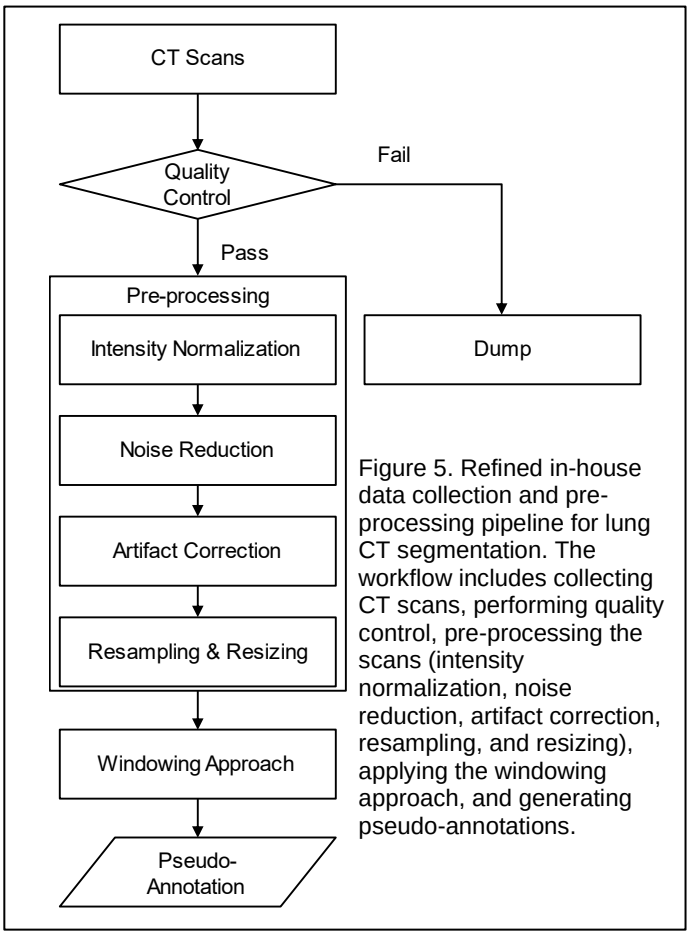

Figure 5. Refined in-house data collection and pre-processing pipeline for lung CT segmentation. The workflow includes collecting CT scans, performing quality control, pre-processing the scans (intensity normalization, noise reduction, artifact correction, resampling, and resizing), applying the windowing approach, and generating pseudo-annotations.

Scans that pass the quality control checks will then undergo a comprehensive pre-processing procedure to standardize the data and enhance the performance of our segmentation algorithms. The first step in our pre-processing pipeline is intensity normalization, where we will scale the raw CT intensity values to a standardized range. This is typically achieved by converting the raw HU to a predefined range (e.g., -1000 to 400 HU) that emphasizes the lung parenchyma and other relevant structures. Intensity normalization ensures that the intensity distributions are consistent across scans, facilitating the subsequent analysis steps. The next step is noise reduction, where we will employ advanced techniques such as non-local means filtering or block-matching and 3D filtering (BM3D) to suppress noise while preserving the fine details and edges of the lung structures. These techniques exploit the self-similarity of image patches to effectively denoise the images without over-smoothing the important features. After noise reduction, we will apply artifact correction methods to mitigate common CT artifacts such as beam hardening, partial volume effects, and motion artifacts. This may involve techniques such as projection-based metal artifact reduction, iterative reconstruction algorithms, or deep learning-based artifact suppression methods. By correcting these artifacts, we can improve the overall quality and interpretability of the CT scans. Finally, we will resample the scans to a consistent voxel spacing and resize them to a fixed image size (e.g., 512x512x512) to ensure a standardized input for our segmentation algorithms. Resampling will be performed using interpolation methods such as trilinear or higher-order spline interpolation to minimize the loss of spatial resolution. The choice of target voxel spacing and image size will be based on the trade-off between computational efficiency and segmentation accuracy, as determined by empirical evaluation.

After the completion of the pre-processing steps, the curated lung CT dataset will be ready for the application of our patch-based probability map generation pipeline, which leverages the foundation model developed in SA-1. This pipeline employs a windowing approach, where the pre-processed CT scans are divided into smaller patches, and each patch is processed by the foundation model to generate class-specific probability maps. By using multiple window sizes and combining the probability maps from different patches, we can generate high-resolution, multi-class probability maps that provide detailed localization information for various anatomical regions in the lungs. These probability maps serve as pseudo-annotations, providing a rich source of training data for our segmentation models without the need for manual annotation. The pseudo-annotations will be used in SA-3 to fine-tune our segmentation models, enabling them to adapt to the specific characteristics of our in-house dataset and improve their performance on real-world clinical data. By leveraging the power of the foundation model and the windowing approach, we can efficiently generate high-quality pseudo-annotations for a large number of scans, overcoming the limitations of manual annotation and accelerating the development of robust lung CT segmentation algorithms.

SA-3: Pseudo-Segmentation Label Generation and Model Fine-tuning

Generating Pseudo-Segmentation Labels

In SA-3, we will generate pseudo-segmentation labels for our in-house lung CT dataset by leveraging the probability maps obtained from SA-2 and the MedSAM model. The process involves overlaying a grid on the lung CT image and extracting the coordinates of each grid intersection. Using the SA-2 model, we generate and postprocess a probability map for the lung region by providing a relevant text prompt. This binary probability map focuses on distinguishing between lung and non-lung regions, while the multi-class probability maps generated in SA-2 for other anatomical structures serve as auxiliary inputs to guide the segmentation process. The grid intersections are then filtered based on their location within the probability map, creating a list of "dot prompts" that correspond to the lung region. These dot prompts, along with the lung CT image, are fed into the MedSAM model, which generates a segmentation mask for each dot prompt, as described in **Algorithm 2**.

To create the final pseudo-segmentation label, we combine the individual segmentation masks using a voting scheme. The overlap ratio between each segmentation candidate (C_k) and the probability map (P) is calculated using the formula:

$$\text{OverlapRatio} = \frac{|C_k \cap P|}{|C_k|} \qquad \boldsymbol{Eq.9}$$

where $|C_k \cap P|$ represents the intersection area between the segmentation candidate and the probability map, and $|C_k|$ represents the area of the segmentation candidate. If the overlap ratio exceeds a heuristically predefined threshold, the segmentation candidate is included in the final pseudo-segmentation label. The selected segmentation candidates are then combined using a union operation to form the final pseudo-segmentation label:

$$\text{Seg}_{\text{pseudo}} = \bigcup \left(C_k \middle| \frac{|C_k \cap P|}{|C_k|} > Threshold \right) \qquad \boldsymbol{Eq.10}$$

Algorithm 2 Pseudo-Segmentation Label Generation

Input:
I: Input lung CT image
G: Grid size for extracting dot prompts
SA-2Model: Probability map generation model from SA-2
MedSAM: MedSAM segmentation model
threshold: Overlap ratio threshold for selecting segmentation candidates
Output:
Seg_{pseudo}: Final pseudo-segmentation label
Initialization:
Initialize an empty segmentation mask (Seg_{pseudo}) of the same size as the input CT image
Generate grid intersections (*gridIntersections*) on the input CT image *I* using grid size *G*
Generate probability map (*probabilityMap*) for the lung region using *SA-2Model* with the text prompt
Filter grid intersections (*dotPrompts*) based on their location within the *probabilityMap*
Initialize an empty list (*Masks*) to store individual segmentation masks
for each *dot* in *dotPrompts* **do**
  Generate a segmentation mask (*mask_candidate*) using *MedSAM* with *I* and *dot*
  Append *mask_candidate* to *Masks*
end for
for each *candidate* in *Masks* **do**
  Calculate the *overlapRatio* between *candidate* and *probabilityMap*
  if *overlapRatio* > *threshold* **then**
    Seg_{pseudo} = Seg_{pseudo} ∪ *segmentationMask*
  end if
end for
return Seg_{pseudo}

By applying this pseudo-segmentation label generation procedure to our in-house lung CT dataset collected in SA-2, we can obtain a large number of annotated samples to fine-tune the MedSAM model. The generated pseudo-segmentation labels will serve as a valuable resource for adapting the model to the specific characteristics of our dataset, improving its ability to accurately segment lung regions in real-world clinical scenarios. The use of probability maps from SA-2 and the MedSAM model in the label generation process helps to ensure the quality and reliability of the pseudo-segmentation labels. The probability maps provide a robust estimate of the lung regions, while the MedSAM model's segmentation capabilities help to refine and localize the regions of interest. By combining these two sources of information, we can generate pseudo-segmentation labels that closely approximate the true segmentation masks, enabling effective fine-tuning of the MedSAM model.

Fine-Tuning MedSAM with Pseudo-Segmentation Labels and H-SAM

Given the limited size of our in-house dataset, data augmentation will play a vital role in expanding the training set and improving the model's ability to generalize to unseen cases. By applying various transformations to the pseudo-labeled CT scans, we can introduce more variability and simulate a wider range of lung anatomies and imaging conditions. This will help the model learn more robust features and reduce overfitting to the specific characteristics of our dataset. To augment our pseudo-labeled dataset, we will apply a combination of geometric, intensity-based, and advanced augmentation techniques. Geometric transformations, such as random rotations, translations, and elastic deformations, will be used to simulate variations in patient positioning and anatomical structures. These transformations will help the model learn to be invariant to such variations, enabling it to segment lung regions accurately across a wide range of patient scans. Intensity-based transformations, such as brightness and contrast adjustments, Gaussian noise addition, and histogram equalization, will be employed to mimic variations in image acquisition settings and enhance the model's ability to handle different imaging conditions. In addition to these standard augmentation techniques, we will incorporate advanced augmentation methods, such as Cutout[4], CutMix[5], and Mixup[6], described in **Figure 6(B)**, to further enrich our training dataset, illustrated in Figure 6. Cutout randomly masks out square regions of the input image, encouraging the model to learn more robust features and reducing overfitting. CutMix replaces the cropped regions with patches from another image, creating new combinations of visual patterns and improving the model's ability to generalize. Mixup linearly combines two input images and their corresponding labels, generating interpolated samples that help the model learn smoother decision boundaries. By applying these advanced augmentation techniques to both the input CT scans and their corresponding pseudo-segmentation labels, we can significantly expand the diversity and richness of our training dataset, leading to improved segmentation performance and robustness.

To fine-tune the MedSAM model, we will adopt the H-SAM architecture, which introduces several innovative components to improve segmentation performance, illustrated in **Figure 6(A).** H-SAM employs a LoRA-adapted image encoder, which freezes the pre-trained layers of the original SAM encoder and adds a smaller, trainable bypass composed of two low-rank matrices. This approach allows for efficient adaptation of the encoder to the specific characteristics of our lung CT dataset while preserving the valuable pre-learned knowledge. During fine-tuning, only the bypass matrices will be updated, enabling minor yet effective adjustments to the encoder's feature representations. To complement the hierarchical Transformer decoder, H-SAM also incorporates a hierarchical pixel decoder inspired by the U-Net architecture. The pixel decoder in the second stage integrates features from the first-stage pixel decoder through skip connections, allowing for the generation of high-resolution predictions. This hierarchical pixel decoder effectively handles multi-scale objects in medical images and captures intricate local details, further enhancing the segmentation quality.

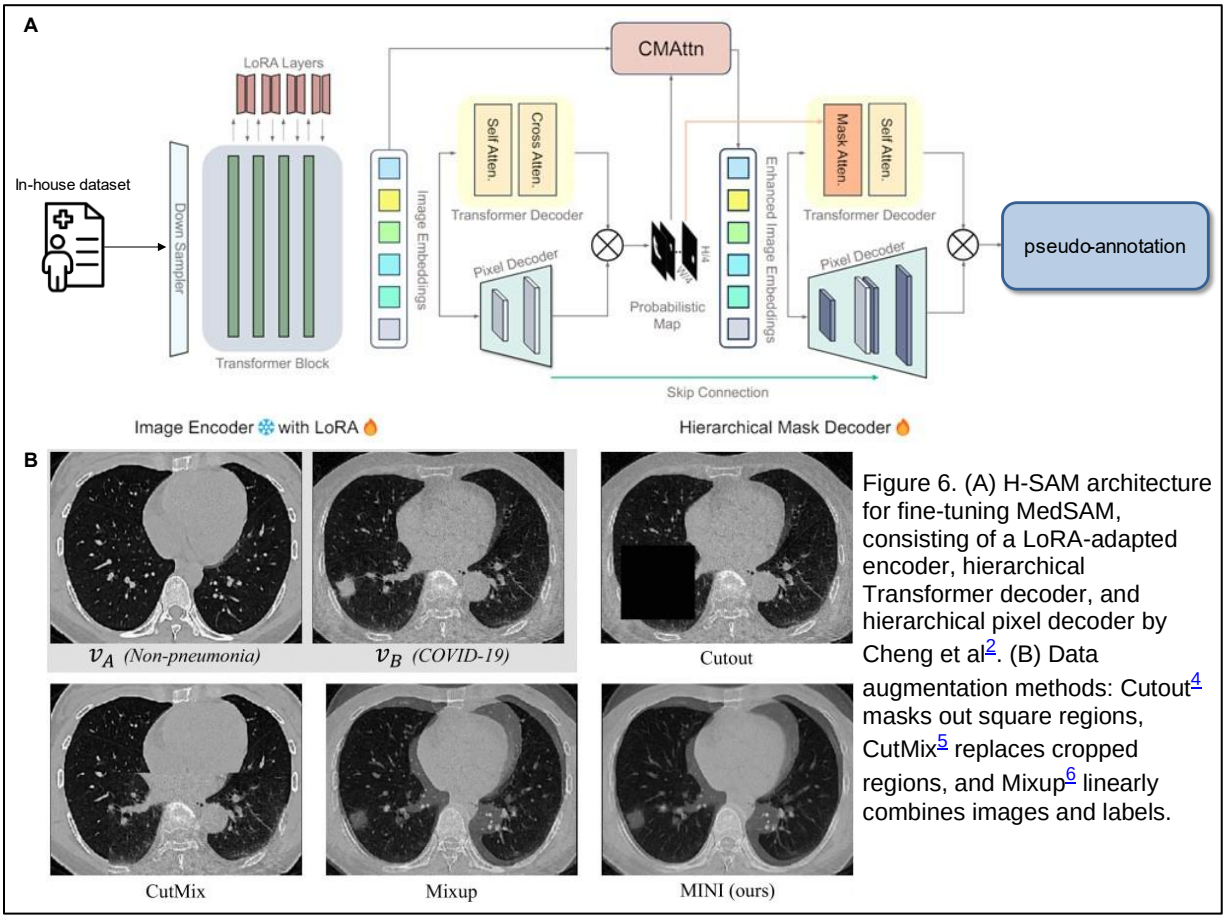

Figure 6. (A) H-SAM architecture for fine-tuning MedSAM, consisting of a LoRA-adapted encoder, hierarchical Transformer decoder, and hierarchical pixel decoder by Cheng et al[2]. (B) Data augmentation methods: Cutout[4] masks out square regions, CutMix[5] replaces cropped regions, and Mixup[6] linearly combines images and labels.

During the fine-tuning process, we will employ a combination of loss functions to optimize the MedSAM model. The training loss will consist of a pixel-wise classification loss, such as binary cross-entropy loss, and a region-based loss, such as the Dice loss. The binary cross-entropy loss is defined as:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \qquad Eq.11$$

where N is the number of pixels, y_i is the ground truth label for pixel i, and $\hat{y}_i$ is the predicted probability for pixel i.

The Dice loss, which is particularly effective for segmentation tasks with imbalanced class distributions, is defined as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \qquad Eq.12$$

where ϵ is a small constant added for numerical stability.

These losses will be applied to both stages of the hierarchical decoding procedure, with the first stage supervised using ground truth masks of reduced resolution and the second stage supervised using the original high-resolution ground truth. The total loss for each stage will be a weighted combination of the binary cross-entropy loss and the Dice loss:

$$L_{stage} = \lambda_{BCE} L_{BCE} + \lambda_{Dice} L_{Dice} \qquad Eq.13$$

where λ_{BCE} and λ_{Dice} are hyperparameters that control the relative importance of each loss term.

The final output will be an ensemble of the probabilities from both stages, leveraging the complementary information captured at different scales. The ensemble output $\hat{y}_{ensemble}$ is computed as a weighted average of the probabilities from the first stage ($\hat{y}_1$) and the second stage ($\hat{y}_2$):

$$\hat{y}_{ensemble} = \alpha \hat{y}_1 + (1 - \alpha)\hat{y}_2 \qquad Eq.14$$

where α is a hyperparameter that determines the contribution of each stage to the final output.

By incorporating deep supervision and a weighted combination of stage-specific losses, we can ensure thorough optimization of the model and achieve superior segmentation performance.

By leveraging the pseudo-segmentation labels generated from our in-house lung CT dataset, employing advanced data augmentation techniques, and adopting the H-SAM architecture for fine-tuning, we aim to develop a highly accurate and robust lung segmentation model. The combination of these strategies will enable the model to learn from a diverse range of patient scans, adapt to the specific characteristics of our dataset, and generate precise segmentations that can support clinical decision-making and research applications in the field of pulmonary medicine.

Evaluating the Fine-Tuned MedSAM Model

To assess the effectiveness of our fine-tuned MedSAM model, we will evaluate its performance on a held-out portion of the in-house lung CT dataset with expert-annotated segmentation labels. This evaluation set will consist of a diverse range of cases, including various stages of lung diseases and different patient demographics, to ensure a comprehensive assessment of the model's generalization capabilities. The ground truth segmentation labels for this evaluation set will be carefully annotated by experienced radiologists, following a standardized protocol to ensure consistency and reliability.

We will employ several widely-used evaluation metrics to quantify the model's performance and compare it against state-of-the-art methods. The primary metric will be the Dice similarity coefficient (DSC), which measures the overlap between the predicted segmentation and the ground truth. The DSC is defined as:

$$DSC = 1 - L_{\text{Dice}} \qquad Eq.15$$

where A is the predicted segmentation, and B is the ground truth segmentation. A higher DSC value indicates better segmentation accuracy, with a perfect overlap resulting in a DSC of 1. Additionally, we will calculate the Hausdorff distance (HD), which measures the maximum distance between the predicted and ground truth segmentation boundaries. The HD is defined as:

$$HD(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right\} \qquad Eq.16$$

where d(a, b) is the Euclidean distance between points a and b. A lower HD value indicates better boundary alignment and segmentation precision. We will report both the 95th percentile HD and the average symmetric surface distance (ASSD) to provide a comprehensive assessment of the model's boundary delineation performance. Other evaluation metrics, such as sensitivity, specificity, and the Jaccard index, will also be calculated to provide a more comprehensive understanding of the model's performance.

	Whole lesion	Central 50%	Off-center 50%
SAM (no fine-tuning)	0.4745±0.2138	0.7343±0.1789	0.2890±0.2076
SAM single task (fine-tuning)	0.7136±0.1277	**0.8463±0.1240**	0.5034±0.2070
SAM parallel multitask (fine-tuning)	0.6985±0.1312	0.8354±0.1178	0.4891±0.2113
SAM cascaded multitask (fine-tuning)	0.7239±0.1321	0.8363±0.1149	0.5241±0.2291
2D nnU-Net	0.6682±0.1659	0.7786±0.1740	0.4553±0.1966
3D nnU-Net	**0.7797±0.1089**	0.8341±0.1039	**0.6331±0.1973**

Table 1. Example comparison of lung CT segmentation performance between fine-tuned SAM model and state-of-the-art methods by Liu et al[3].

To benchmark our fine-tuned MedSAM model, we will compare its performance against state-of-the-art methods for lung CT segmentation. This will include the original MedSAM model, which will serve as a baseline to demonstrate the effectiveness of our fine-tuning approach. Additionally, we will compare our model against other top-performing models specifically designed for lung CT segmentation, such as nnU-Net[33] and SAM[20]. nnU-Net is a self-configuring deep learning framework that has shown excellent performance across various medical image segmentation tasks, while TotalSegmentator is a specialized model that has achieved state-of-the-art results in lung CT segmentation. By comparing our model's performance against these established methods, we can assess its relative strengths and weaknesses and demonstrate its potential for clinical application. The results of this comparison will be presented in a comprehensive table, showcasing the example evaluation metrics for each model in **Table 1**. This table will provide a clear and concise overview of our model's performance relative to the state-of-the-art, enabling readers to easily assess its effectiveness and potential impact in the field of pulmonary medicine.

Potential Pitfalls and Alternatives

While our proposed approach of leveraging foundation models, knowledge distillation, and pseudo-labeling techniques for lung CT segmentation shows promise, there are several potential pitfalls to consider. One major concern is the reliance on pre-training performance. If the pre-training phase does not yield a sufficiently robust and generalizable foundation model, the subsequent fine-tuning process may struggle to achieve optimal results. This could be due to limitations in the dataset used for pre-training, such as insufficient diversity or inadequate representation of certain anatomical variations. To mitigate this risk, it is crucial to carefully curate the pre-training dataset, ensuring that it encompasses a wide range of lung anatomies, pathologies, and imaging conditions. Additionally, exploring alternative pre-training strategies, such as contrastive learning or self-supervised learning, could potentially enhance the foundation model's ability to capture meaningful representations.

Another potential issue is the effectiveness of our approach in segmenting small or subtle structures within the lungs. While the use of a large pseudo-labeled dataset may improve the model's performance on larger, more prominent regions like the lung parenchyma, it may not necessarily translate to accurate segmentation of finer details, such as small nodules or vascular structures. This limitation could be attributed to the inherent challenges in generating high-quality pseudo-labels for these intricate regions, as well as the potential

oversimplification of their appearance in synthetic data. To address this concern, it may be necessary to incorporate additional strategies specifically targeted at refining the segmentation of small structures. This could involve the use of multi-scale architectures, attention mechanisms, or specialized loss functions that prioritize the accurate delineation of these regions. Furthermore, integrating a smaller set of expertly annotated data focusing on these challenging cases could provide valuable guidance during the fine-tuning process.

Lastly, the size of the text data in our pre-training dataset may be insufficient. While we have a substantial number of CT scans (60,000), which can be split into a large number of individual images (700,000), the corresponding text data is limited to only 60,000 instances. This disparity could hinder the foundation model's ability to learn robust associations between visual and textual features, potentially impacting its performance in downstream tasks. To overcome this limitation, we could explore techniques for augmenting the text data, such as generating synthetic reports using natural language generation models or leveraging external sources of medical text data. Additionally, investigating alternative architectures that can effectively handle the imbalance between image and text data, such as cross-modal attention mechanisms or hierarchical fusion strategies, could help mitigate the impact of limited text information.

References

1. El-Nouby, A., et al., *Are large-scale datasets necessary for self-supervised pre-training?* arXiv preprint arXiv:2112.10740, 2021.
2. Cheng, Z., et al., *Unleashing the Potential of SAM for Medical Adaptation via Hierarchical Decoding.* arXiv preprint arXiv:2403.18271, 2024.
3. Liu, Y., et al. *Universal 3D CT lesion segmentation using SAM with RECIST annotation*. in *Medical Imaging 2024: Image Processing*. 2024. SPIE.
4. DeVries, T. and G.W. Taylor, *Improved regularization of convolutional neural networks with cutout.* arXiv preprint arXiv:1708.04552, 2017.
5. Yun, S., et al. *Cutmix: Regularization strategy to train strong classifiers with localizable features*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
6. Zhang, H., et al., *mixup: Beyond empirical risk minimization.* arXiv preprint arXiv:1710.09412, 2017.
7. Soriano, J.B., et al., *Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* The Lancet Respiratory Medicine, 2020. **8**(6): p. 585-596.
8. Li, H.Y., et al., *Global, regional and national burden of chronic obstructive pulmonary disease over a 30-year period: estimates from the 1990 to 2019 global burden of disease study.* Respirology, 2023. **28**(1): p. 29-36.
9. Bray, F., et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2021.
10. Guan, W.-j., et al., *Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis.* European Respiratory Journal, 2020. **55**(5).
11. Rubin, G.D., *Lung nodule and cancer detection in computed tomography screening.* Journal of thoracic imaging, 2015. **30**(2): p. 130-138.
12. Thrall, J.H., et al., *Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success.* Journal of the American College of Radiology, 2018. **15**(3): p. 504-508.
13. Gerard, S.E., et al., *FissureNet: a deep learning approach for pulmonary fissure detection in CT images.* IEEE transactions on medical imaging, 2018. **38**(1): p. 156-166.
14. Karimi, D., et al., *Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis.* Medical image analysis, 2020. **65**: p. 101759.
15. Hofmanninger, J., et al., *Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem.* European Radiology Experimental, 2020. **4**: p. 1-13.
16. Willemink, M.J., et al., *Preparing medical imaging data for machine learning.* Radiology, 2020. **295**(1): p. 4-15.
17. Yang, J., et al., *Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017.* Medical physics, 2018. **45**(10): p. 4568-4581.
18. Kaissis, G.A., et al., *Secure, privacy-preserving and federated machine learning in medical imaging.* Nature Machine Intelligence, 2020. **2**(6): p. 305-311.
19. Mazurowski, M.A., et al., *Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI.* Journal of magnetic resonance imaging, 2019. **49**(4): p. 939-954.

20. Kirillov, A., et al. *Segment anything*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
21. Ma, J., et al., *Segment anything in medical images.* Nature Communications, 2024. **15**(1): p. 654.
22. Bommasani, R., et al., *On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258, 2021.
23. He, S., et al., *Accuracy of segment-anything model (sam) in medical image segmentation tasks.* arXiv preprint arXiv:2304.09324, 2023.
24. Raghu, M., et al., *Transfusion: Understanding transfer learning for medical imaging.* Advances in neural information processing systems, 2019. **32**.
25. Geirhos, R., et al., *Shortcut learning in deep neural networks. Nature Machine Intelligence, 2 (11), 665–673*. 2020, Number.
26. Panch, T., H. Mattie, and R. Atun, *Artificial intelligence and algorithmic bias: implications for health systems.* Journal of global health, 2019. **9**(2).
27. Rangwani, H., et al., *DeiT-LT Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets.* arXiv preprint arXiv:2404.02900, 2024.
28. Koleilat, T., et al., *MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation.* arXiv preprint arXiv:2403.20253, 2024.
29. Yan, K., et al., *DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning.* Journal of medical imaging, 2018. **5**(3): p. 036501-036501.
30. Draelos, R.L., et al., *Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes.* Medical image analysis, 2021. **67**: p. 101857.
31. Li, Y., et al. *Neural architecture search for lightweight non-local networks*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
32. Zhang, Y., et al., *Collaborative unsupervised domain adaptation for medical image diagnosis.* IEEE Transactions on Image Processing, 2020. **29**: p. 7834-7844.
33. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
34. Hatamizadeh, A., et al. *Unetr: Transformers for 3d medical image segmentation*. in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022.
35. Hamamci, I.E., et al., *A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities.* arXiv preprint arXiv:2403.17834, 2024.
36. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PMLR.
37. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
38. Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.
39. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.
40. Touvron, H., et al., *Llama: Open and efficient foundation language models.* arXiv preprint arXiv:2302.13971, 2023.
41. Loshchilov, I. and F. Hutter, *Decoupled weight decay regularization.* arXiv preprint arXiv:1711.05101, 2017.
42. Cubuk, E.D., et al. *Randaugment: Practical automated data augmentation with a reduced search space*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020.
43. Harrison, A.P., et al. *Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images*. in *Medical Image Computing and Computer Assisted Intervention−MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. 2017. Springer.