HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full length article

SwinUNeLCsT: Global–local spatial representation learning with hybrid CNN–transformer for efficient tuberculosis lung cavity weakly supervised semantic segmentation

Zhuoyi Tan [a], Hizmawati Madzin [a,*], Bahari Norafida [b], Rahmita Wirza OK Rahmat [a], Fatimah Khalid [a], Puteri Suhaiza Sulaiman [a]

[a] *Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, 43400, Malaysia*
[b] *Department of Radiology, Universit Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

ARTICLE INFO

Keywords:
Deep learning
Classification
Semantic segmentation
Weakly-supervised learning
CT tuberculosis imaging

ABSTRACT

Radiological diagnosis of lung cavities (LCs) is the key to identifying tuberculosis (TB). Conventional deep learning methods rely on a large amount of accurate pixel-level data to segment LCs. This process is time-consuming and laborious, especially for those subtle LCs. To address such challenges, firstly, we introduce a novel 3D TB LCs imaging convolutional neural network (CNN)-transformer hybrid model (SwinUNeLCsT). The core idea of SwinUNeLCsT is to combine local details and global dependencies for TB CT scan image feature representation to effectively improve the recognition ability of LCs. Secondly, to reduce the dependence on accurate pixel-level annotations, we design an end-to-end LCs weakly supervised semantic segmentation (WSSS) framework. Through this framework, radiologists need only to classify the number and the approximate location (e.g., left lung, right lung, or both) of LCs in the CT scan to achieve efficient segmentation of the LCs. This process eliminates the need for meticulously drawing boundaries, greatly reducing the cost of annotation. Extensive experimental results show that SwinUNeLCsT outperforms currently popular medical 3D segmentation methods in the supervised semantic segmentation paradigm. Meanwhile, our WSSS framework based on SwinUNeLCsT also performs best among the existing state-of-the-art medical 3D WSSS methods.

1. Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium TB that has long posed a serious threat to global public health. A key indicator in the radiological diagnosis of TB is lung cavities (LCs) (Tan et al., 2024; Dartois and Rubin, 2022; Ullah et al., 2024). Accurate identification of LCs is critical to confirm diagnosis, treatment, and monitor TB disease progression. However, due to the complexity and diversity of the LCs, it is not easy to annotate these cavities at the pixel level in CT (computerized tomography) images.

Over the past decade, deep learning models have achieved great success in the field of medical image analysis. However, due to the data-hungry nature of deep learning networks, semantic segmentation models based on the full supervision learning paradigm typically require a large amount of labor-intensive pixel-level annotated data (Lateef and Ruichek, 2019; Asgari Taghanaki et al., 2021). To address this issue, some recent methods have sought to use inexpensive labels to overcome the challenge of annotating radiological images, including self-supervised learning (Chen et al., 2021; Zhang et al., 2021; Huang et al., 2021), active learning (Kirsch et al., 2019; Yoo and Kweon, 2019), semi-supervised learning (Berthelot et al., 2019; Zheng et al., 2022; Tan et al., 2023b), and weakly-supervised learning (Ouyang et al., 2020; Chen et al., 2022b).

In the field of weakly supervised semantic segmentation (WSSS) (Zhang et al., 2020; Laradji et al., 2021; Ye et al., 2022; Sun et al., 2023; Lu et al., 2023), class activation mapping (CAM) (Selvaraju et al., 2017; Chen et al., 2022b; Wang et al., 2024) are the most common solution,

[*] Corresponding author.
E-mail address: hizmawati@upm.edu.my (H. Madzin).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2024.102012
Received 20 January 2024; Received in revised form 21 February 2024; Accepted 14 March 2024
Available online 28 March 2024

and it is also the most challenging one among all WSSS scenarios. The fundamental concept behind CAM is to create a heatmap from the network's feature maps, highlighting the regions within an image that receive the most attention during the model's classification decisions. This enables CAM-based methods to segment the deep semantic information using only image-level labels (e.g., information about what classes are present in an image) (Tan et al., 2023a), and thus solve the problem of the high cost of pixel-level annotation in radiological images to some extent. However, directly applying CAM-based methods to identify pulmonary TB cavities in CT images is not straightforward because the lack of pixel-level guidance when training classification networks often leads to an excessively broad focus on features (Tan et al., 2022), thereby reducing the segmentation accuracy of LCs using CAM. In addition, the popular WSSS methods that use image-level labels usually adopt a multi-stage framework (Sun et al., 2020; Wu et al., 2021). Specifically, the multi-stage approach starts by generating CAM as pseudo-label check-ins to train semantic segmentation tasks (Xu et al., 2023; Su et al., 2023). This process needs to involve multiple steps to perform, which increases the training complexity and reduces the efficiency.

In the early development of the WSSS field, WSSS primarily relied on a convolutional neural network (CNN) (Isensee et al., 2021; Wen et al., 2021) to generate CAM. However, with the rise of vision transformers (ViTs) (Dosovitskiy et al., 2020), ViTs have also begun to be widely applied in the generation of CAM for WSSS (Li et al., 2023b; Yu et al., 2023a; Ru et al., 2023). Unlike the local receptive field of CNN, ViTs can leverage their self-attention mechanism to capture global information more effectively in images (Ding et al., 2022). As a result, CAM generated using ViTs often display richer global features. However, the limitation lies in the fact that standard transformer architecture models are typically less sensitive to local features, which often results in their inability to effectively identify some fine LCs structures. This, in turn, impacts the precise medical diagnosis of TB.

Therefore, to address these challenges, our study focuses on the following three primary objectives:

1. Addressing the limitations of traditional transformer architectures in capturing minute structural features of LCs, we propose a novel medical 3D imaging architecture model, dubbed SwinUNeLCsT. Inspired by Dong et al. (2023), the core module of SwinUNeLCsT employs a hybrid architecture of transformer and CNN. This hybrid structure aims to integrate both global and local features in pulmonary TB CT images, enhancing the recognition of LCs. As illustrated in Fig. 1, SwinUNeLCsT demonstrates better performance in identifying LCs of different sizes compared to the standard self-attention-based architectural model.
2. To reduce the dependence on the pixel-level annotation of the LCs and improve training efficiency, we design an end-to-end (Amodei et al., 2016; Tampuu et al., 2020) WSSS framework for the LCs. In this framework, we utilize multi-class labels of CT images to generate pixel-level weakly supervised regions. These regions are then refined to produce pseudo semantic segmentation labels for LCs.
3. To optimize the gradient conflicts (Yu et al., 2020; Liu et al., 2021a) between various training tasks and improve the ability to recognize TB LCs lesion features, we construct a novel LCs WSSS training optimization strategy.

By adopting our WSSS framework, radiologists need only classify the number and approximate location (e.g., left lung, right lung, or both) of LCs in a CT scan to achieve efficient segmentation of the LCs, without the need to meticulously delineate the boundaries of the cavities in each slice, significantly reducing the cost of annotation. Our main contributions in this work are summarized as follows:

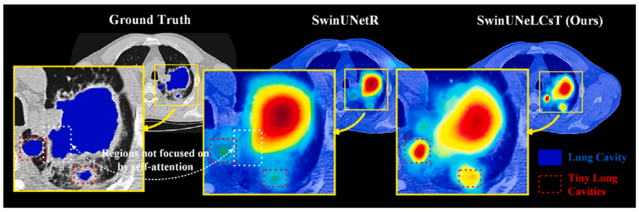


Fig. 1. Class activation mappings comparison: Standard self-attention in SwinUNetR (Tang et al., 2022) versus global–local attention in SwinUNeLCsT for recognizing lung cavities.

1. We introduce a novel 3D self-attention and convolution hybrid module and propose a novel model for TB LCs CT image analysis, dubbed SwinUNeLCsT. This model integrates local and global features to capture LCs of various sizes better. Compared with the standard self-attention module, this model not only has lower computational complexity but also improves LCs recognition efficiency.
2. We propose an end-to-end learning framework for WSSS of LCs. This framework effectively segments LCs using only classification annotations specific to these cavities. Moreover, compared to state-of-the-art WSSS methods in medical imaging, this framework has achieved the best performance in segmenting LCs. To our knowledge, this is the first framework developed specifically for the WSSS of TB-affected LCs.
3. We design a novel optimization strategy for WSSS training of the LCs. This strategy can improve the ability to extract LCs lesion features by integrating the probability maps obtained in previous iterations to refine the feature maps and alleviate the gradient conflicts generated during model training.

Overall, the contributions made in this article provide a forward-looking solution strategy for the long-standing issues of efficient LCs semantic segmentation and the high cost of pixel-level annotation in pulmonary TB CT medical imaging analysis.

2. Related work

2.1. Deep learning in pulmonary tuberculosis CT images

Deep learning (DL) has great potential in the field of medical image analysis, especially in the field of pulmonary TB from CT images. Although DL has been widely used in TB X-ray analysis and other medical imaging fields (Zhou et al., 2021b; Sainz de Cea et al., 2020; Huang et al., 2022), its applications in TB CT image processing are still relatively few. For example, for chest X-ray (CXR), Iqbal et al. (2023) proposed a TB-specific model TB-UNet, which mainly consists of two parts: the first part is a classification network consisting of five double convolutional blocks, DenseNet-169 layers, and a feature fusion module; The second part is the segmentation network, which is mainly composed of the dilated fusion block and the attention mechanism. In order to identify TB lesion tissue from CT scans, Gordaliza et al. (2019) proposed a training adaptation strategy based on the V-Net model. This strategy enables the use of complete 3D volumes to capture fine-grained features. Moreover, the model is optimized through a novel multi-task learning loss function that leverages model uncertainty to weight regression and binary classification tasks. Alebiosu et al. (2023) developed a new medical image segmentation model, DAvoU-Net, which can generate features for TB analysis by leveraging multi-scale residual blocks and receptive dense connections. Moreover, to reflect the severity of TB more accurately in the image, the model adopted a novel method, which was to input the features extracted by a three-dimensional CNN into a bidirectional Long Short-Term memory network (Bi-LSTM), which enabled the network to extract higher level

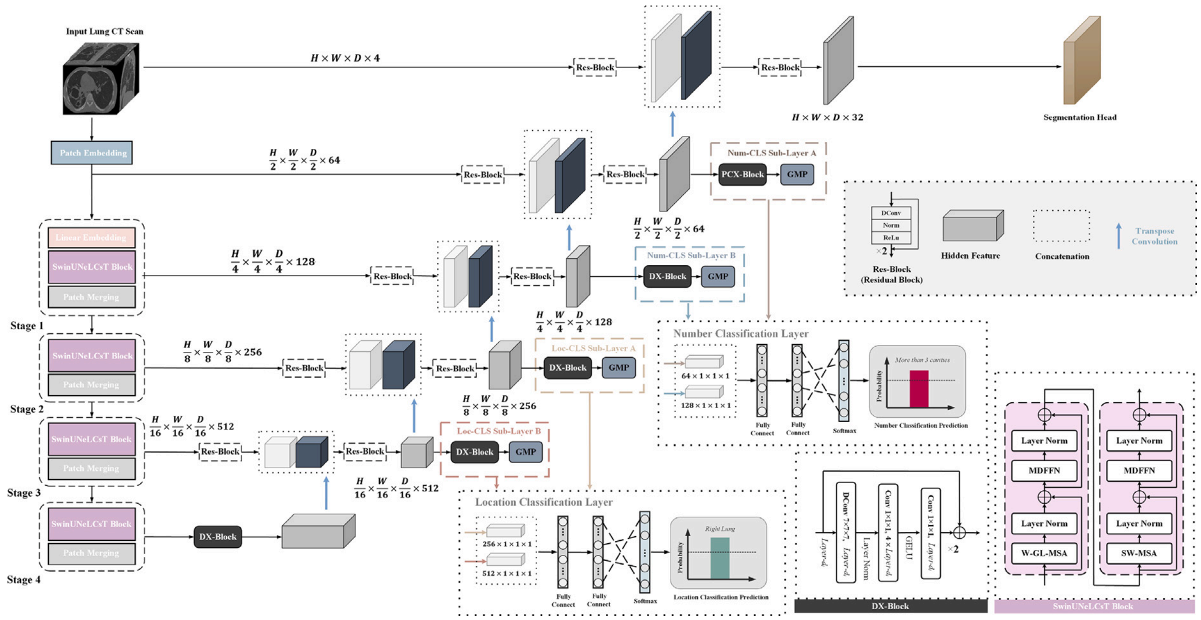


Fig. 2. Overview of the SwinUNeLCsT architecture.

discriminative features. However, the aforementioned DL approach cannot be directly applied to the identification of LCs in TB. This is due to, firstly, the small size, complex morphologies, and indistinct boundaries of LCs lesions in TB CT images, which make recognizing TB LCs more challenging compared to other application domains. Therefore, to better identify LCs in CT data, we need to develop a specialized DL strategy to overcome these challenges.

2.2. Medical image weakly supervised semantic segmentation

WSSS is a form of machine learning, characterized by its allowance for the use of incomplete, inaccurate, or inconsistent annotation data for training models. In traditional semantic segmentation, models usually require a large amount of pixel-level precise annotation data for training. However, in WSSS, models can be trained using simpler or incomplete annotations. These annotations might be bounding boxes (Khoreva et al., 2017; Song et al., 2019; Liu et al., 2020), scribbles (Lin et al., 2016; Unal et al., 2022), or just image-level labels (Wang et al., 2020; Chen et al., 2022b). The typical workflow of WSSS is first to use weak labels to train a model and then generate pseudo-segmentation masks. These masks are later refined through a series of refinement strategies and finally used to support the training of segmentation models in a standard fully supervised manner. Among the many WSSS methods, using image-level labels to train classification networks is widely adopted due to its cost-effectiveness. Chen et al. (2022a) proposed a new method suitable for medical WSSS—causal CAM. This method employs two causal chains: categorical causal chain and anatomical causal chain. Li et al. (2023a) developed a novel weakly-supervised approach, SA-MIL, for precise pixel-level segmentation of histopathological images. By treating pixels as instances in Multiple Instance Learning (MIL) and introducing a self-attention mechanism, this method enhances the global relevance among instances and segmentation performance. However, there are certain limitations in directly applying these methods to identify LCs, primarily because the activation maps obtained directly from classification networks, while highlighting the key feature regions of the LCs in images, focus on features that are often too broad. Different from previous work, we propose a weakly supervised learning method specifically for the LCs. This method not only focuses on the most distinctive features of the LCs but also more accurately identifies the contour information of the LCs.

3. SwinUNeLCsT

3.1. SwinUNeLCsT encoder

The overall SwinUNeLCsT architecture is shown in Fig. 2. Denote the input TB CT scan image as a sub-volume $\mathcal{X} \in \mathbb{R}^{H\times W\times D\times S}$ and the volumetric token with a patch resolution of (V_h, V_w, V_d) has a patch size $V_h \times V_w \times V_d \times S$. A sequence of 3D tokens is projected onto a size of $\frac{H}{V_h'} \times \frac{W}{V_w'} \times \frac{D}{V_d'} \times C$ in the patch partitioning layer, where C is the represents the dimensionality of the embedding space. Following Liu et al. (2021b), to efficiently model interactions between tokens in a 3D context, all projected sequences of embeddings are partitioned into non-overlapping windows. These windows are used to calculate local self-attention within each region. The outputs of encoder blocks in layers l and $l+1$ are computed as:

$$\begin{aligned}
\hat{z}^l &= \text{W-GL-MSA}(\text{LN}(z^{l-1})) + z^{l-1}\\
z^l &= \text{MDFFN}(\text{LN}(\hat{z}^l)) + \hat{z}^l\\
\hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^l)) + z^l\\
z^{l+1} &= \text{MDFFN}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1},
\end{aligned} \tag{1}$$

where W-GL-MSA and SW-MSA represent window-based global–local and sliding window partitioned multi-head self-attention modules respectively. $\hat{z}^l$ and $\hat{z}^l$ are the outputs of W-GL-MSA and SW-MSA; LN and MDFFN denote layer normalization and multi-DConv feed forward network, respectively.

Our encoder uses a patch size of $2 \times 2 \times 2$. Each patch is mapped to an 8-dimensional feature space (as $2\times2\times2\times1 = 8$), where 1 represents that the image has only one input channel. The initial embedding space dimension is $C = 64$ dimensions. Our encoder overall includes 4 stages, each stage containing 2 SwinUNeLCsT blocks, totaling 8 layers ($L = 8$). Between each stage, patch merging layers are used to reduce the resolution by a factor of 2. Patch merging involves grouping patches of size $2 \times 2 \times 2$ and concatenating their features, increasing the feature dimension to $4C$, but as patches are merged, the total number of tokens is reduced. After merging, a linear layer downsamples the features from $4C$ dimensions to $2C$ dimensions, thereby reducing the dimensionality of features and further lowering the resolution. In the hierarchical process, the linear embedding layer and SwinUNeLCsT block outputs from the first stage maintain a resolution of $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$, where H, W, and D represent the original height, width, and depth of the input

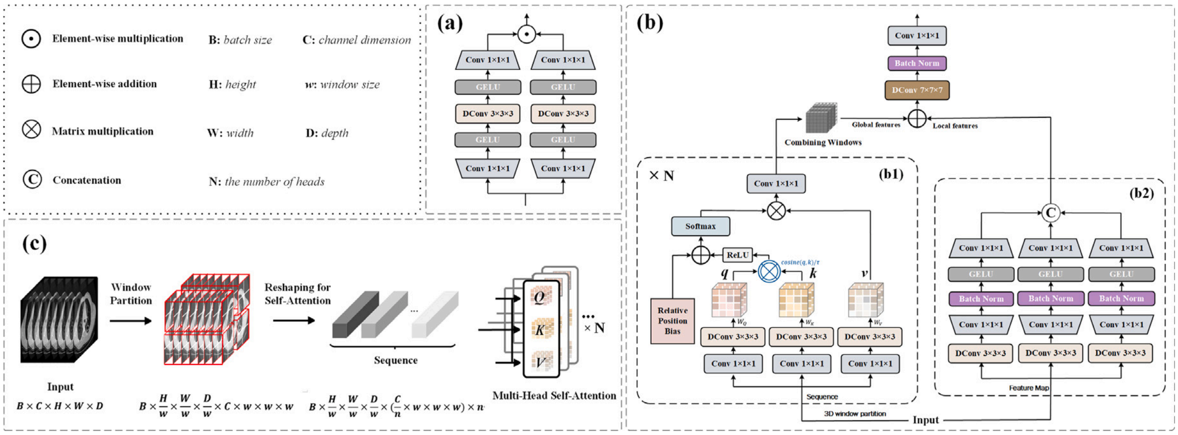


Fig. 3. Components of the SwinUNeLCsT Block: (a) multi-DConv feed forward network (MDFFN). (b) The window-based global–local multi-head self-attention (W-GL-MSA) Module, with (b1) depicting the spectral scaled cosine multi-head self-attention and (b2) illustrating the lightweight simplified inception. (c) The 3D window partition operation.

image. Subsequent stages continue this pattern, further reducing the resolution in each stage, with the second stage at $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, the third stage at $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$, and the fourth stage at $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$.

3.2. Multi-DConv feed-forward network

In traditional swin transformers (Liu et al., 2021b), the multi-layer perceptron (MLP) usually consists of a series of fully connected layers, which capture complex abstract features of the input data through stacking multiple nonlinear transformation layers. However, such a structural design can lead to unnecessary complexity and redundancy. To address these issues and enhance the network's feature extraction capabilities, we propose the multi-DConv feed-forward network (MDFFN), as shown in Fig. 3(a). MDFFN employs *depth-wise separable* convolutional layers, *DConv* 3 × 3 × 3, to optimize the efficiency of local feature processing. Furthermore, it improves the modeling of interactions between features by performing element-wise multiplication on the feature maps output by two parallel sub-networks at corresponding positions. Assume that $\mathbf{X}_l$ is the input to the lth layer of MDFFN, the operations in one branch of the network would be:

$$\mathbf{Y}_{l,DConv} = GELU(DConv_{3\times3\times3}(GELU(Conv_{1\times1\times1}(\mathbf{X}_l))))$$

$$\mathbf{Y}_{l,conv} = Conv_{1\times1\times1}(\mathbf{Y}_{l,DConv})$$

where GELU represents Gaussian Error Linear Unit (Hendrycks and Gimpel, 2016). And then the output $\mathbf{Y}_l$ of the MDFFN is obtained by an element-wise multiplication of the outputs of two such branches followed by an element-wise addition with the input:

$$\mathbf{Y}_l = \mathbf{Y}^{(1)}_{l,conv} \odot \mathbf{Y}^{(2)}_{l,conv}$$

3.3. Window-based global–local multi-head self-attention

In complex LCs semantic segmentation tasks, although the role of global context is crucial, local information is equally indispensable as a key element to maintaining rich spatial details. Traditional self-attention mechanisms excel in capturing global dependencies, as they compute relationships across all sequence elements, which is beneficial for tasks necessitating a comprehensive context awareness. On the other hand, convolutions capture local features through their local receptive fields. This characteristic often makes them more effective in processing data with strong spatial structures like TB CT scans. Therefore, to fully leverage the advantages of both architectures, we develop a window-based global–local multi-head self-attention (W-GL-MSA), as shown in Fig. 3(b). W-GL-MSA contains two parallel branches: spectral scaled cosine multi-head self-attention and lightweight simplified Inception, which are used to extract global and local context information respectively. Overall, the output of W-GL-MSA can be represented as:

$$\mathbf{O}_{\text{W-GL-MSA}} = Conv_{1\times1\times1}\left(BN(DConv_{7\times7\times7}(\mathbf{I} + \mathbf{X}))\right) \tag{2}$$

where BN stands for Batch Normalization. $\mathbf{I}$ is the output of the lightweight simplified inception module, and $\mathbf{X}$ is the output processed by the window-based spectral scaled cosine self-attention module.

3.3.1. 3D window partition

In the 3D patch partition module of the SwinUNeLCsT encoder, the input image is first divided into multiple small blocks, known as patches, which are then mapped to a high-dimensional feature space through linear embedding, as shown in Fig. 3(c). Specifically, in this process, the input tensor is segmented into distinct, non-overlapping 3D windows. Consider a 3D window of dimensions $w \times w \times w$. Consequently, the total number of such windows is calculated as $\frac{H}{w} \times \frac{W}{w} \times \frac{D}{w}$. The dimensions of each window are $B \times C \times w \times w \times w$. Subsequently, these 3D windows are transformed into 2D matrices. This transformation results in a flattened dimension of $B \times \frac{H}{w} \times \frac{W}{w} \times \frac{D}{w} \times (C \times w \times w \times w)$. Self-attention is then applied within each of these flattened windows. This involves the independent computation of *Query* (Q), *Key* (K), and *Value* (V) matrices for each window. In the multi-head attention mechanism, where each head independently processes a distinct segment of the channel dimension (C). Consequently, this approach results in a dimensionality of $B \times \frac{H}{w} \times \frac{W}{w} \times \frac{D}{w} \times (\frac{C}{h} \times w \times w \times w)$ for each head, with h representing the number of heads. Following the computation of self-attention, the output is reconstituted into the 3D space for each window. These windows are then reconstructed to match the original 3D input tensor's shape. Finally, these processed windows are reassembled into a comprehensive $B \times C \times H \times W \times D$ tensor, thereby completing the output of the 3D window partition operation with integrated self-attention.

3.3.2. Window-based spectral scaled cosine self-attention

The self-attention mechanism in transformer models is key for capturing global image features, but it is also a major source of computational overhead. In the field of medical imaging, the standard self-attention mechanism (Vaswani et al., 2017; Liu et al., 2021b) exhibits a significant increase in the interaction time and memory complexity of key-query dot product with the deepening dimensions of CT images, specifically yielding a complexity of $O(W^2H^2D^2)$ for an image with $W \times H \times D$ pixels. Although reducing the input image resolution can mitigate the computational complexity of self-attention, it results in a decline in the model's recognition performance. To solve this problem, following previous work (Wang et al., 2022; Zamir et al., 2022; Liu et al., 2022a; Dong et al., 2023), we propose a novel spectral scaled cosine multi-head self-attention (SSC-MSA) with linear complexity, as illustrated in Fig. 3 (b1). The input $X_{in} \in \mathbb{R}^{H \times W \times D \times C}$ is first embedded, yielding in the generation of the *query* vector $\mathbf{Q} = W_d^Q W_p^Q X$, the *key* vector $\mathbf{K} = W_d^K W_p^K X$, and the *value* vector $\mathbf{V} = W_d^V W_p^V X$. Here, $W_p^{(\cdot)}$ represents a $1 \times 1 \times 1$ *point-wise* convolution, and $W_d^{(\cdot)}$ signifies a *depth-wise* convolution with dimensions $3 \times 3 \times 3$. Next, the SSC-MSA method transforms the projections $\mathbf{Q}$ and $\mathbf{K}$ to produce a transposed attention map A with a size of $\mathbb{R}^{C \times C \times C}$ through their cosine similarity interaction. Overall, the SSC-MSA process is defined as:

$$\mathbf{X} = \mathbf{W}_p \cdot \mathbf{BN} \cdot \mathbf{W}_{Dc7} \cdot \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{3}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\text{ReLU}(\text{cosine}(\mathbf{Q}, \mathbf{K}))}{\tau}\right)\mathbf{V} \tag{4}$$

where $\mathbf{Q} \in \mathbb{R}^{HWD \times C}$, $\mathbf{K} \in \mathbb{R}^{C \times HWD}$, and $\mathbf{V} \in \mathbb{R}^{HWD \times C}$ are matrices obtained after reshaping tensors from their original size $\mathbb{R}^{H \times W \times D}$. $\mathbf{W}_{Dc7}$ represents a $7 \times 7 \times 7$ *point-wise* convolution. The scalar τ is a trainable parameter, which is not shared across different heads or layers. It is initialized with a value greater than 0.01.

3.3.3. Lightweight simplified inception

The lightweight simplified inception is a relatively shallow structure, primarily responsible for extracting local features in images. This module's design, which aligns with the concepts presented in Chollet (2017), comprises three parallel and identical branches. Within each branch, the feature map initially undergoes a *DConv* $3 \times 3 \times 3$. Then, it is followed by two consecutive $1 \times 1 \times 1$ convolution layers. Between these two convolution layers, a layer norm and a GELU activation function are inserted to preserve feature diversity and achieve lower latency. Finally, the feature maps from these three branches are merged. Specifically, the implementation details of the lightweight and simplified inception module are shown in Fig. 3(b2). Let $\mathbf{F}$ be the feature map input to the module. The module's output $\mathbf{I}$ can be represented as:

$$\mathbf{O}_i = Conv_{1\times1\times1}\Big(ReLU(BN(Conv_{1\times1\times1}(DConv_{3\times3\times3}(\mathbf{F}_i)))))\tag{5}$$

$$\mathbf{I} = \text{Concat}\big(\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3\big) \tag{6}$$

3.4. SwinUNeLCsT decoder

To enhance feature representation and effectively reduce the impact of processing transformer sequence lengths, this study introduces a convolution-based decoder. This decoder utilizes skip-connection technology to achieve information fusion between swin transformer and convolutional networks at various levels. In our SwinUNeLCsT model, the output sequence of each encoding stage is reorganized into feature maps of specific sizes. Specifically, the height H, width W, and depth D of the input image are reduced by a factor of 2^i, forming feature maps of size $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i} \times C$, where $i = 0, 1, 2, 3, 4$ and $C = 32, 64, 128, 256, 512$. As the network depth increases, the spatial resolution of the feature maps gradually decreases. For example, the output feature map of the first stage $i = 0$ retains the original size, while that of the fifth stage (bottleneck stage) $i = 5$ is reduced to $1/32$ of the original size. Overall, SwinUNeLCsT captures and encodes image features of different scales more effectively by progressively reducing the feature space. In lower stages (such as $i = 0$ or $i = 1$), larger feature maps help capture more details; in higher stages (such as $i = 4$ or the bottleneck stage $i = 5$), smaller feature maps tend to encode global and abstract features. The output feature map of the bottleneck stage is input into a DX block to generate the final output of the encoder. Subsequently, the decoder upsamples the bottleneck feature map through a transposed convolution layer. The output of the transposed convolution is merged with the representations of the preceding layers and input into a residual block that contains two $3 \times 3 \times 3$ *depth-wise* convolution layers, followed by instance normalization and ReLU activation functions. The final output feature map is processed through a $1 \times 1 \times 1$ convolution layer and a softmax activation function to generate probability masks for segmenting TB CT medical images.

3.4.1. DX block

To learn richer feature representations in TB CT imaging, we develop a novel large-kernel depthwise separable convolution module (as shown in Fig. 2), inspired by the ConvNeXt architecture (Liu et al., 2022b), known as the DX-block. This module aims to capture extensive spatial features and long-range dependencies in the input lung CT scan data. Specifically, the module receives an input $\mathbf{X}$ with a dimension of $layer\text{-}d_i$, where i represents a specific layer in the network. Then, the output of a basic DX-block can be expressed as:

$$\hat{\mathbf{X}} = Conv_{1\times1\times1}(GELU(Conv_{1\times1\times1}(LN(DConv_{7\times7\times7}(\mathbf{X}))))) \tag{7}$$

$$\mathbf{O}_{\text{DX-block}} = \mathbf{X} + \hat{\mathbf{X}}, \tag{8}$$

where LN stands for layer normalization.

4. Lung cavity weakly supervised semantic segmentation

In LCs WSSS, we first use the SwinUNeLCsT model to jointly train tasks for classifying the number and location of lung cavities, to obtain CAM for LCs from these two classification layers. Then, to better represent LCs features, we propose a technique called CAM feature fusion gate (CFFG), which merges the CAM of these two layers, producing CAM that simultaneously possess regions of interest for both the location and number of LCs. Given the CAM $\mathcal{M} \in \mathbb{R}^{h \times w \times d \times C}$, the overall CAM $\mathcal{M}(i, j, s)$ is defined as follows:

$$\mathcal{M}(i, j, s, :) = \text{CFFG}(\mathcal{M}^{num}(i, j, s, :), \mathcal{M}^{loc}(i, j, s, :)), \tag{9}$$

where $\mathcal{M}(i, j, s, :)$ denotes the amalgamated CAM that fuses both number classification layer's CAM ($\mathcal{M}^{num}(i, j, s, :)$) and location classification layer's CAM ($\mathcal{M}^{loc}(i, j, s, :)$).

Moreover, CAM for each classification layer is formed by merging the CAM of two sub-layers through CFFG:

$$\mathcal{M}^{num}(i, j, s, :) = \text{CFFG}(\mathcal{M}^{num}_{\text{layer-A}}(i, j, s, :), \mathcal{M}^{num}_{\text{layer-B}}(i, j, s, :)), \tag{10}$$

$$\mathcal{M}^{loc}(i, j, s, :) = \text{CFFG}(\mathcal{M}^{loc}_{\text{layer-A}}(i, j, s, :), \mathcal{M}^{loc}_{\text{layer-B}}(i, j, s, :)), \tag{11}$$

where $\mathcal{M}^{num}_{\text{layer-A}}$ and $\mathcal{M}^{num}_{\text{layer-B}}$ denote number classification sub-layers A and B. $\mathcal{M}^{loc}_{\text{layer-A}}$ and $\mathcal{M}^{loc}_{\text{layer-B}}$ denote location classification sub-layers A and B.

Subsequently, through a series of steps detailed in Section 4.2, we refine these regions of interest and generate pseudo-annotations for LCs. Finally, we use these pseudo-annotations for end-to-end WSSS of LCs, as shown in Fig. 4.

4.1. CFFG module

The CFFG module is implemented as follows: Firstly, ω_k^c represents the weights for a specific class c in the CAM A_k at the classification layer. These weights indicate the influence of the given class c on every spatial location in the A_k. Secondly, the weights ω_k^c corresponding to each layer are multiplied by the A_k of that layer. Furthermore, by summing the weighted activation values across all layers, the resulting value $\mathcal{M}^c$ represents the contribution of the given class c to the heatmap. To highlight regions of interest in the heatmap $\mathcal{M}^c$ and suppress irrelevant areas, we apply a Rectified Linear Unit (ReLU) operation to $\mathcal{M}^c$. The overall computation is expressed as follows:

$$\mathcal{M}^c = ReLU(\sum_k \omega_k^c \cdot A_k). \tag{12}$$

Finally, for CAM from two different layers $\mathcal{M}^c_{\text{layer-1}}$ and $\mathcal{M}^{\hat{c}}_{\text{layer-2}}$, the operation of CFFG can be defined as:

$$\mathcal{M}(i, j, s, :) = \max(\mathcal{M}^c_{\text{layer-1}}(i, j, s, :), \mathcal{M}^{\hat{c}}_{\text{layer-2}}(i, j, s, :)), \tag{13}$$

where the categories of classification c and $\hat{c}$ in $\mathcal{M}^c_{\text{layer-1}}$ and $\mathcal{M}^{\hat{c}}_{\text{layer-2}}$ can be *num* or *loc*, where *num* and *loc* represent LCs number and location classification, respectively.

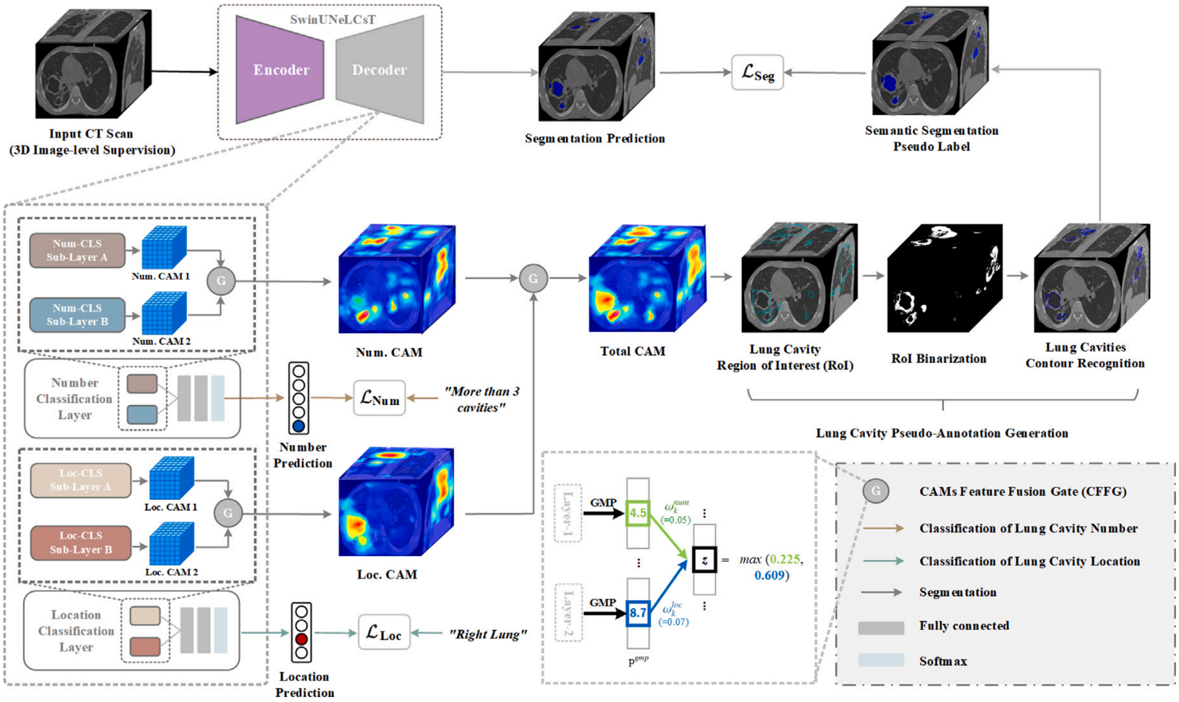


Fig. 4. Overview of the proposed lung cavities weakly supervised semantic segmentation framework. The lung cavities pseudo-annotation generation process is shown in Fig. 5(a).

4.2. Weakly supervised semantic segmentation pseudo label generation

The LCs WSSS pseudo-label generation can be divided into the following four steps:

In the first step, we extract dependable foreground and background information from the target CAM $\mathcal{M}^{i,j,s,:}$, utilizing a background score β_l (where $0 < \beta_l < 1$) (Ru et al., 2022). This process is implemented as follows:

$$\mathcal{I}(i,j,s) = \begin{cases} \mathtt{argmax}(\mathcal{M}^{i,j,s,:}), & \text{if } \mathtt{max}(\mathcal{M}^{i,j,s,:}) \geq \beta_l, \\ 0, & \text{if } \mathtt{max}(\mathcal{M}^{i,j,s,:}) < \beta_l, \end{cases} \tag{14}$$

where, $\mathcal{M}^{i,j,s,:}$ represents the channel values in the CAM at location (i,j,s), and $\mathtt{max}(\mathcal{M}^{i,j,s,:})$ represents the maximum channel value at that position. If the maximum channel value is greater than or equal to the threshold β_l, $\mathcal{I}(i,j,s)$ is set to the index of the maximum value in the corresponding channel (i.e., the predicted class label). If the maximum channel value is less than or equal to the threshold β_l, $\mathcal{I}(i,j,s)$ is set to 0, representing the background. This step aims to remove unreliable weak supervision information in the image and retain pixel information that is strongly related to the background and foreground.

In the second step, we utilize the LCs category information $\mathcal{I}(i,j,s)$ obtained from the first step to extract and generate preliminary pseudo labels for LCs from the original image $\mathcal{O}(i,j,s)$:

$$\mathcal{R}(i,j,s) = \mathbf{1}_{\{\mathcal{I}(i,j,s)=1\}} \cdot \mathcal{O}(i,j,s), \tag{15}$$

where $\mathbf{1}_{\{\mathcal{I}(i,j,s)=1\}}$ is an indicator function that is valued at 1 if and only if the pixel (i,j,s) in $\mathcal{I}(i,j,s)$ is classified as LCs; otherwise, it is valued at 0. Such processing ensures that only the pixel information belonging to the LCs category is retained, setting all other pixels to 0 to effectively remove the background.

In the third step, we perform a binary thresholding operation on $\mathcal{R}(i,j,s)$ to facilitate the subsequent LCs contour recognition. The result after binary thresholding, denoted as $\mathcal{B}(i,j,s)$, is as follows:

$$\mathcal{B}(i,j,s) = \begin{cases} 255, & \text{if } \mathcal{R}(i,j,s) > \beta_t \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

where β_t represents the threshold value, after binary thresholding, $\mathcal{B}(i,j,s)$ will have pixel values of either 0 or 255, making it suitable for further LCs contour recognition.

In the fourth step, we propose a LCs recognition algorithm to identify the LCs in the input image $\mathcal{B}(i,j,s)$ from three different dimensions, as shown in Fig. 5(b). In the implementation principle, the algorithm accepts an image $\mathcal{B}(i,j,s)$ as input, where i and j represent the pixel coordinates in the image and s represents the size or a specific slice in the image sequence. The algorithm performs contour detection in three different dimensions. For each contour c within C_d for every dimension, the algorithm evaluates whether each contour is a parent contour. If it is, that contour will be removed from C_d. This step is to remove peripheral contours or irrelevant structures to focus only on the inner LCs contours. Once all dimensions have been processed, their results are concatenated to form the final detection result.

4.3. Network training

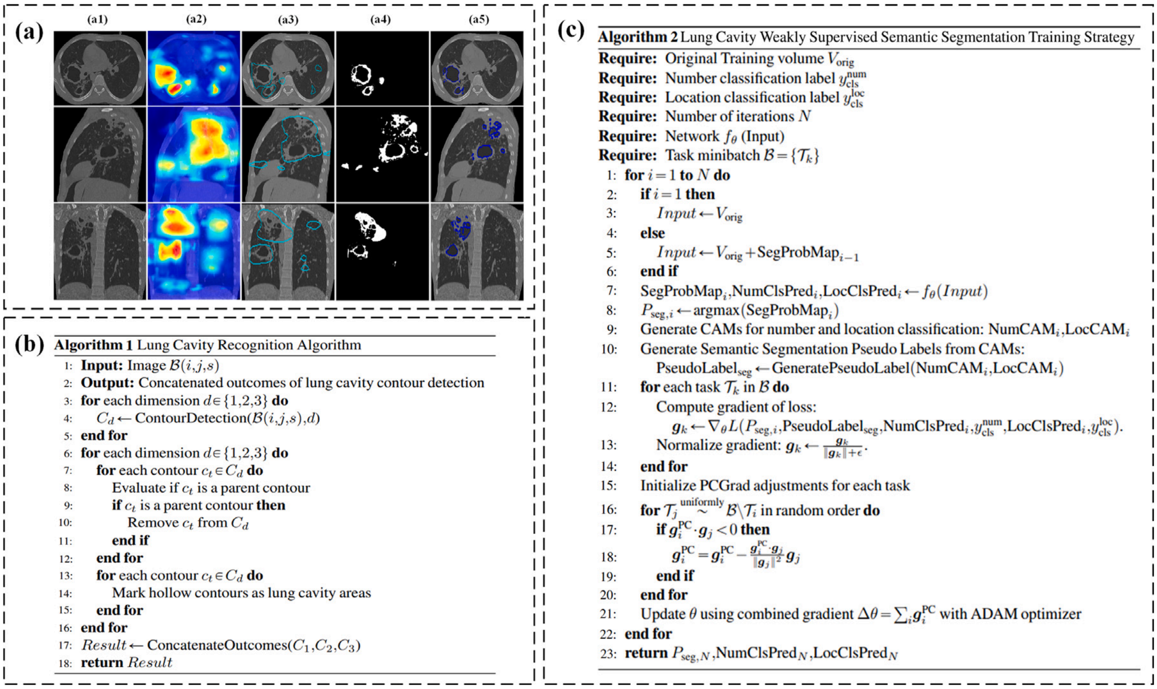
The overall loss of our approach is the weighted sum of three distinct loss functions: $\mathcal{L}_{Num}$, $\mathcal{L}_{Loc}$, and $\mathcal{L}_{Seg}$. These functions are combined as follows:

$$\mathcal{L}_{\mathrm{WSSS}} = \lambda_1 \mathcal{L}_{Num} + \lambda_2 \mathcal{L}_{Loc} + \lambda_3 \mathcal{L}_{Seg}, \tag{17}$$

where λ_1, λ_2, and λ_3 are the weighting coefficients balancing the contributions of the different loss functions. The function $\mathcal{L}_{Seg}$ is the binary cross-entropy loss (Ruby and Yendapalli, 2020), defined as:

$$\mathcal{L}_{\mathrm{Seg}} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha y_i (1-p_i)^{\gamma} \log(p_i) + (1-y_i) p_i^{\gamma} \log(1-p_i)], \tag{18}$$

where N represents the total number of samples, with y_i and p_i denoting the true label and the predicted probability for the ith sample, respectively. The parameter γ is set to 2, enhancing the model's focus on difficult, misclassified samples. The loss functions $\mathcal{L}_{Loc}$ and $\mathcal{L}_{Num}$ employ the multi-label soft margin loss (Liang et al., 2017) method. This method calculates the discrepancy between the predicted category



```
Algorithm 1 Lung Cavity Recognition Algorithm
Input: Image $\mathcal{B}(i,j,s)$
Output: Concatenated outcomes of lung cavity contour detection
for each dimension $d \in \{1,2,3\}$ do
    $C_d \leftarrow$ ContourDetection($\mathcal{B}(i,j,s),d$)
end for
for each dimension $d \in \{1,2,3\}$ do
    for each contour $c_t \in C_d$ do
        Evaluate if $c_t$ is a parent contour
        if $c_t$ is a parent contour then
            Remove $c_t$ from $C_d$
        end if
    end for
    for each contour $c_t \in C_d$ do
        Mark hollow contours as lung cavity areas
    end for
end for
$Result \leftarrow$ ConcatenateOutcomes($C_1,C_2,C_3$)
return $Result$
```

```
Algorithm 2 Lung Cavity Weakly Supervised Semantic Segmentation Training Strategy
Require: Original Training volume $V_{\text{orig}}$
Require: Number classification label $y_{\text{cls}}^{\text{num}}$
Require: Location classification label $y_{\text{cls}}^{\text{loc}}$
Require: Number of iterations $N$
Require: Network $f_\theta$ (Input)
Require: Task minibatch $\mathcal{B} = \{\mathcal{T}_k\}$
for $i = 1$ to $N$ do
    if $i = 1$ then
        $Input \leftarrow V_{\text{orig}}$
    else
        $Input \leftarrow V_{\text{orig}} + \text{SegProbMap}_{i-1}$
    end if
    $\text{SegProbMap}_i, \text{NumClsPred}_i, \text{LocClsPred}_i \leftarrow f_\theta(Input)$
    $P_{\text{seg},i} \leftarrow \text{argmax}(\text{SegProbMap}_i)$
    Generate CAMs for number and location classification: $\text{NumCAM}_i, \text{LocCAM}_i$
    Generate Semantic Segmentation Pseudo Labels from CAMs:
        $\text{PseudoLabel}_{\text{seg}} \leftarrow \text{GeneratePseudoLabel}(\text{NumCAM}_i, \text{LocCAM}_i)$
    for each task $\mathcal{T}_k$ in $\mathcal{B}$ do
        Compute gradient of loss:
            $\boldsymbol{g}_k \leftarrow \nabla_\theta L(P_{\text{seg},i}, \text{PseudoLabel}_{\text{seg}}, \text{NumClsPred}_i, y_{\text{cls}}^{\text{num}}, \text{LocClsPred}_i, y_{\text{cls}}^{\text{loc}})$.
        Normalize gradient: $\boldsymbol{g}_k \leftarrow \frac{\boldsymbol{g}_k}{\|\boldsymbol{g}_k\| + \epsilon}$.
    end for
    Initialize PCGrad adjustments for each task
    for $\mathcal{T}_j \overset{\text{uniformly}}{\sim} \mathcal{B} \backslash \mathcal{T}_i$ in random order do
        if $\boldsymbol{g}_i^{\text{PC}} \cdot \boldsymbol{g}_j < 0$ then
            $\boldsymbol{g}_i^{\text{PC}} = \boldsymbol{g}_i^{\text{PC}} - \frac{\boldsymbol{g}_i^{\text{PC}} \cdot \boldsymbol{g}_j}{\|\boldsymbol{g}_j\|^2} \boldsymbol{g}_j$
        end if
    end for
    Update $\theta$ using combined gradient $\Delta\theta = \sum_i \boldsymbol{g}_i^{\text{PC}}$ with ADAM optimizer
end for
return $P_{\text{seg},N}, \text{NumClsPred}_N, \text{LocClsPred}_N$
```

Fig. 5. Lung cavity weakly supervised semantic segmentation pseudo-annotation generation process and training process. (a1)–(a5) successively represent the original image, CAM output by CFFG, the region of interest for the lung cavity, binary processing of the ROI, contour detection of the lung cavity.

probability vector P_c and the true image-level labels Y_c in the LCs location and number classification tasks:

$$\mathcal{L}_\varphi = \frac{1}{K}\sum_{k=1}^{K}(Y^k \log(P_\varphi^k) + (1 - Y^k)\log(1 - P_\varphi^k)), \tag{19}$$

where φ indicates the type of task, taking the values $\mathtt{Num}$ and $\mathtt{Loc}$ for the number and location classification of LCs, respectively. The symbol k represents the total number of classes.

However, jointly training multiple tasks involves the problem of multi-objective function optimization strategies (Tian et al., 2021). This is because gradient conflict or competition (Yu et al., 2020) often occurs in the joint training process of different tasks. To solve the conflict problem between different tasks, the PCGrad algorithm (Yu et al., 2020) is a common solution. Therefore, in this paper, based on this algorithm (Yu et al., 2020), we develop a LCs WSSS training optimization strategy, as shown in Fig. 5(c). In this strategy, first, we designed an iterative input feature refinement mechanism (Zhou et al., 2021b). In this mechanism, input features are updated according to the segmentation probability map from the previous iteration, to better segment lesions in the LCs with unclear boundaries. For instance, LCs lesions such as ground-glass opacities often have indistinct borders and cannot be effectively identified in a single iteration. Secondly, to avoid the "gradient explosion" problem (Kanai et al., 2017), we normalize the gradients by dividing them by their norm to ensure that the gradient magnitude remains within a reasonable range. Additionally, to prevent division by zero errors during gradient updates, we introduce a small constant ϵ. Finally, to alleviate the gradient conflicts arising from the joint training of semantic segmentation, number classification, and location classification tasks, we employ the PCGrad algorithm to coordinate and optimize gradient updates between different tasks. PCGrad effectively alleviates the common gradient conflict problem in multi-task training by calculating the gradients of each task and adjusting them if necessary to reduce the negative influence between each other.

5. Data and experiments

In this section, to assess the effectiveness of SwinUNeLCsT, we construct a dataset that can be used for two different supervision paradigms (weak supervision and full supervision) and conduct extensive experiments on it. In this section, we first briefly present the dataset and discuss the implementation details of the experiments in this paper. Subsequently, we introduce the evaluation metrics for the experiment results.

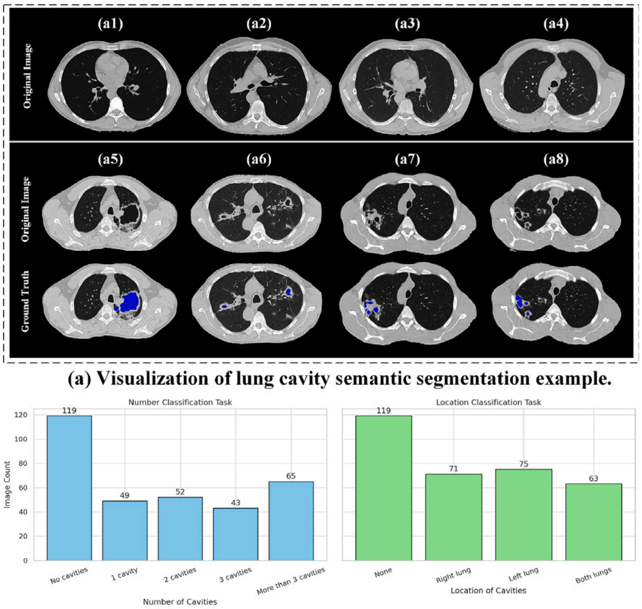
5.1. Dataset

Given that LCs semantic segmentation is an emerging research area, there are no publicly available datasets dedicated to LCs weakly-supervised and supervised semantic segmentation experiments. To fill this gap, we constructed an LCs attribute classification and semantic segmentation dataset containing 328 TB patient CT images for weakly-supervised and supervised semantic segmentation tasks. This dataset, sourced from the official ImageCLEF2022 TB cavern detection challenge,[1] is freely available for non-commercial purposes. This challenge includes annotations for the detection of LCs and their centroids, with each image measuring 512 × 512 pixels. ImageCLEF2022TB cavern detection challenge is one edition of the ImageCLEF series of events.[2] ImageCLEF is an international challenge aimed at advancing research in the field of image retrieval and image recognition. It is part of the Cross-Language Evaluation Forum (CLEF),[3] a comprehensive international challenge focusing on multilingual, multimodal, and interactive retrieval issues in information retrieval systems. However, the ImageCLEF2022TB caverns detection dataset does not contain annotations for the number classification, location classification, and semantic segmentation of LCs lesions. Therefore, to acquire the necessary annotations for number classification, location classification, and semantic segmentation of LCs lesions, we enlisted medical experts from the Department of Radiology, University of Putra Malaysia to manually annotate the TB CT image. The annotation results are shown in Fig. 6.

Moreover, to ensure the gold standard of the annotated data, we evaluated the consistency and variability among different radiologists

[1] https://www.imageclef.org/2022/medical/tuberculosis.
[2] https://www.imageclef.org/.
[3] https://www.clef-initiative.eu/.

in the segmentation and classification of LCs in pulmonary TB. The specific implementation plan is divided into the following steps: First, we selected three experienced medical experts as observers. These experts are board-certified radiologists with a specialization in thoracic imaging, each possessing over 10 years of experience in diagnosing pulmonary diseases, including TB. Next, we provided the observers with uniform guidance and standardized explanations to ensure they had the same understanding and objectives during the segmentation process. This included detailed descriptions of key details such as LCs number assessment. During the annotation process, each observer independently performed semantic segmentation and annotation of CT images from the axial, sagittal, and coronal planes. Furthermore, to observe differences in the annotation results among experienced doctors, we asked the doctors to annotate independently. For the segmentation task, the annotation software we used was ITK-SNAP (Yushkevich et al., 2016), and the annotation results were saved as Nifti format files. For the classification task, the annotation results were saved in an Excel spreadsheet. Finally, to further ensure the accuracy and reliability of the observation results, we re-evaluated the annotation data where there was inconsistency among observers and adopted the principle of the minority conforming to the majority to determine the final annotation results.



(a) Visualization of lung cavity semantic segmentation example.

(b) Lung cavity location and number classification task annotation data distribution.

Fig. 6. Annotated results of the weakly supervised semantic segmentation dataset for lung cavities. The dataset comprises 209 CT scan images featuring lung cavity lesions and 119 without lesions. (a1)–(a4) represent CT images without lung cavity lesions. (a5)–(a8) represent the CT images with lung cavity lesions, and the lesion area is marked in blue.

5.2. Experiment details

Optimizer and Learning Rate. During the training phase, the AdamW optimizer was employed to train the network. The learning rate was initially set to 6×10^{-5} and was progressively reduced following a polynomial decay schedule.

Pseudo Label Generation. For the generation of pseudo labels in LCs semantic segmentation, two thresholds were defined. The threshold for LCs extraction in Eq. (14), denoted as β_l, was set to 0.45. In binary operations for pixel extraction, the threshold in Eq. (16) β_t was determined by the lowest value between the two highest peaks in a CT scan's threshold distribution, ensuring pixels near a value of 0 were removed.

Hyperparameter Optimization. A grid search hyperparameter optimization was conducted specifically for the loss function weight parameters in Eq. (17). This process yielded the optimal weight factors λ_1, λ_2, and λ_3, which were found to be 1, 1, and 0.5, respectively.

Cross-Validation and Image Size. To mitigate experimental bias and enhance the model's generalizability, a 5-fold cross-validation method was utilized. Consistency was maintained in the size of the input images, which were uniformly set to $128 \times 128 \times 96$ pixels across all experiments.

5.3. Evaluation metrics

In our study, we evaluate the segmentation of TB lesions and lung normal regions using three quantitative metrics: the 95th percentile Hausdorff distance (95HD), the dice similarity coefficient (DSC) and the intersection over union (IoU). These metrics are calculated as follows:

95th percentile Hausdorff Distance (95HD). 95HD metric measures the distance between two sets of points that represent the surface of the segmentation predictions and the ground truth. It is defined as the maximum distance of a set percentage (in this case, 95%) of the closest point pairs between the two sets, ensuring that outliers have a less pronounced effect:

$$95\mathrm{HD} = \max\left\{ P_{95}\left(\left\{ \min_{\bar{y}' \in \bar{Y}'} \|y' - \bar{y}'\| \right\}_{y' \in Y'} \right), \right.$$
$$\left. P_{95}\left(\left\{ \min_{y' \in Y'} \|\bar{y}' - y'\| \right\}_{\bar{y}' \in \bar{Y}'} \right) \right\} \tag{20}$$

where Y' and $\bar{Y}'$ denote ground truth and prediction surface point sets. P_{95} represents the 95th percentile function, and it is applied to the set of minimum distances from each point in one set to the closest point in the other set.

Dice similarity coefficient (DSC). The DSC metric is a statistical tool used to measure the similarity between two samples. For segmentation tasks, it quantifies the overlap between the predicted segmentation and the ground truth, with values ranging from 0 (no overlap) to 1 (perfect overlap):

$$\mathrm{Dice} = \frac{2\sum_{i=1}^{I} Y_i \hat{Y}_i}{\sum_{i=1}^{I} Y_i + \sum_{i=1}^{I} \hat{Y}_i}, \tag{21}$$

where $\sum_{i=1}^{I} Y_i \hat{Y}_i$ represents the number of correctly predicted voxels. $\sum_{i=1}^{I} Y_i$ denotes the total number of voxels in the actual set. $\sum_{i=1}^{I} \hat{Y}_i$ represents the total number of voxels in the predicted set.

Intersection over Union (IoU). The IoU is another metric for quantifying the overlap between the predicted segmentation and the ground truth. It is calculated as the area of overlap between the two segmentations divided by the area of their union. Like the Dice coefficient, IoU values range from 0 to 1, where 1 indicates perfect agreement:

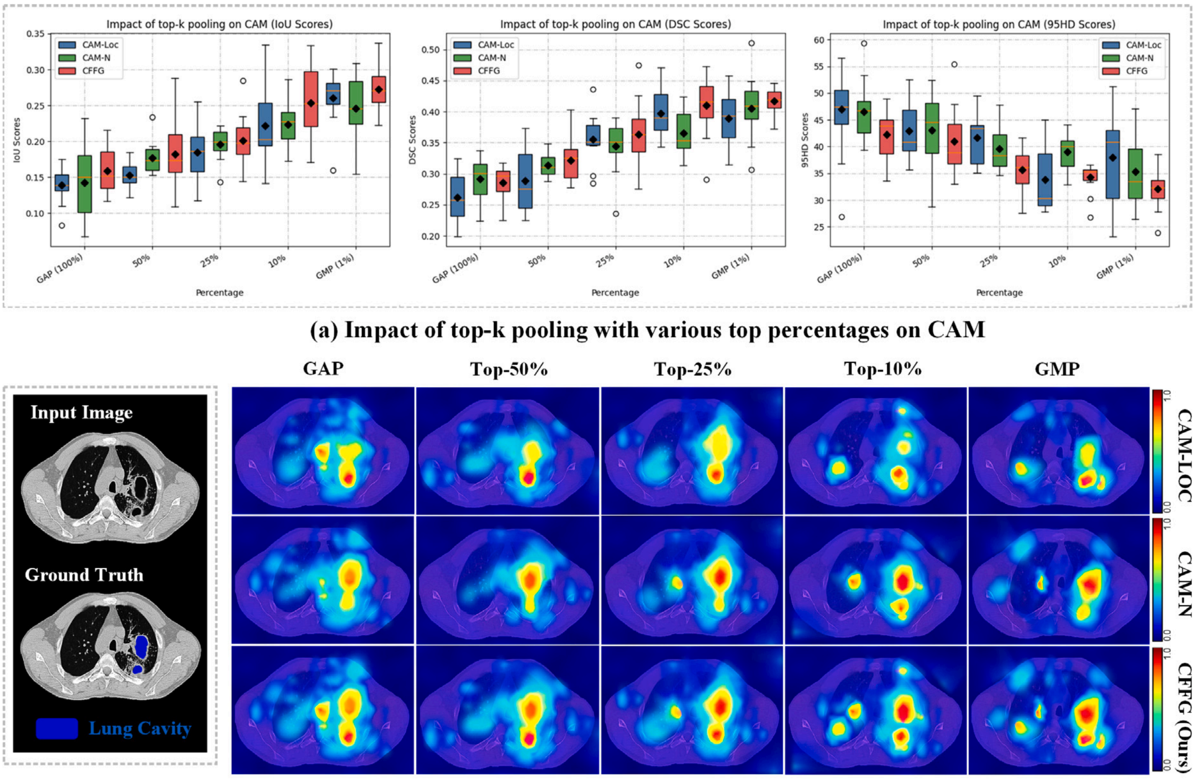
$$\mathrm{IoU} = \frac{\sum_{i=1}^{I} Y_i \cap \hat{Y}_i}{\sum_{i=1}^{I} Y_i \cup \hat{Y}_i}, \tag{22}$$

where $\sum_{i=1}^{I} Y_i \cap \hat{Y}_i$ represents the total number of voxels where both the actual value and the predicted value are 1, i.e., the intersection. $\sum_{i=1}^{I} Y_i \cup \hat{Y}_i$ denotes the total number of voxels where at least one of the actual value or predicted value is 1, i.e., the union.

For the quantitative evaluation of LCs attributes classification, we utilized receiver operating characteristic (ROC) curves, area under the ROC curve (AUC), recall (REC), precision (PRE), accuracy (ACC), false positive rate (FPR), and F1-score (F1).

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{23}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{24}$$



(a) Impact of top-k pooling with various top percentages on CAM

(b) Examples of class activation maps for top-k pooling with various top percentages

Fig. 7. Qualitative and quantitative results of various class activation mapping methods. **CAM-Loc** and **CAM-N** represent class activation mapping for lung cavity location and number classification.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{25}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{26}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}} \tag{27}$$

where FP, TP, FN, and TN represent the counts of false positives, true positives, false negatives, and true negatives, respectively.

6. Results

In this section, we first present the WSSS experimental comparison results of various LCs region of interest generation methods, as well as the ablation results for each component of SwinUNeLCsT. Finally, to better evaluate the performance of SwinUNeLCsT, we compare it with the state-of-the-art methods.

6.1. Generation of lung cavity region of interest

We use the mainstream CAM to obtain the region of interest of LCs, namely the Grad-CAM (Selvaraju et al., 2017). In practice, the choice of pooling method significantly impacts the quality of the generated CAM. Firstly, global max-pooling gmp, emphasizes the regions in feature maps with the maximum response, which is often very effective for highlighting key features that a localization model is concerned with. Secondly, global average-pooling (gap) computes the average value for each channel, resulting in information averaging across the entire feature map. This approach helps capture broader feature information. However, gmp and gap are just two special cases of “top-k%” pooling methods, namely top-100% and top-1% pooling. To further explore the optimal pooling method for classification networks suitable for swin transformer and convolutional hybrid architectures, we also considered the effects of top-k% pooling with k values other than 1% and 100%. In addition, to verify the performance of the CFFG technique, we masked the CAM of the LCs lesion number and location classification tasks respectively while keeping the training strategy of the joint classification task unchanged.

The quantitative and qualitative results analysis of the proposed and baseline methodologies is summarized in Fig. 7. From Fig. 7(a), we observed that CAM generated based on CFFG exhibits the best performance compared to CAM formed by a single task (CAM-Loc or CAM-N). In addition, Fig. 7(b) shows the qualitative recognition result of LCs for various CAM methods. From this sub-figure, we observed that the CAM output method based on gmp constructed in this paper (baseline method) can more precisely focus on LCs lesion tissues compared to other methods.

In conclusion, the experiment in this subsection explores the impact of various pooling methods on the CAM generation effect. Specifically, compared with other pooling strategies (including gap and various top-k% pooling methods), gmp demonstrates more accurate lesion feature identification capabilities. Moreover, compared to CAM outputs from a single task, the CFFG technology proposed in this study offers powerful robustness in identifying LCs by integrating CAM from both lesion location and number classification tasks. Therefore, the results of the aforementioned ablation experiments demonstrate that, in the context of classification networks employing transformer and convolution hybrid architectures, gmp is better at capturing LCs lesion features in CT scan imaging and is more effective in generating high-performance Grad-CAM for feature aggregation.

6.2. Ablation studies

6.2.1. Component ablation experiment

To analyze the contributions of different components in our proposed LCs WSSS framework, we conducted a series of ablation experiments, as shown in Tables 1–2.

Table 1
Impact of various ablation modules in SwinUNeLCsT on weakly supervised semantic segmentation for lung cavities. W-MSA stands for multi-head self-attention module. MLP represents the original MLP component in Swin UNetR. W-GL-MSA stands for window-based global–local multi-head self-attention.

ID	MLP	W-MSA	DX-blocks	MDFFN	W-GL-MSA	Segmentation (±SD)			Number classification					Location classification				
						95HD	IoU	DSC	ACC	REC	PRE	FPR	F1	ACC	REC	PRE	FPR	F1
M-1	×	✓	✓	✓	×	34.78 ± 8.29	0.247 ± 0.133	0.391 ± 0.129	0.711	0.792	0.801	0.462	0.796	0.729	0.812	0.818	0.471	0.815
M-2	×	×	×	✓	✓	34.12 ± 7.95	0.251 ± 0.142	0.395 ± 0.147	0.714	0.802	0.799	0.458	0.800	0.731	0.814	0.823	0.458	0.818
M-3	✓	×	✓	×	✓	32.51 ± 5.23	0.260 ± 0.136	0.411 ± 0.141	0.720	0.798	0.810	0.443	0.804	0.737	**0.822**	0.819	0.445	0.821
M-4	×	×	✓	✓	✓	**31.43 ± 7.46**	**0.267 ± 0.126**	**0.418 ± 0.134**	**0.723**	**0.806**	**0.811**	**0.421**	**0.809**	**0.742**	0.821	**0.835**	**0.398**	**0.828**

Table 2
The impact of various ablation modules in W-GL-MSA on weakly supervised semantic segmentation for lung cavities. W-SSC-SA stands for window-based spectral scaled cosine self-attention. LSI represents lightweight simplified inception.

ID	LSI	W-SSC-SA	Segmentation (± SD)			Number Classification					Location Classification				
			95HD	IoU	DSC	ACC	REC	PRE	FPR	F1	ACC	REC	PRE	FPR	F1
M-4	✓	✓	**31.43 ± 7.46**	**0.267 ± 0.126**	**0.418 ± 0.134**	**0.723**	**0.806**	**0.811**	**0.421**	**0.809**	**0.742**	**0.821**	**0.835**	**0.398**	**0.828**
M-5	✓	×	35.56 ± 8.27	0.235 ± 0.131	0.386 ± 0.129	0.694	0.783	0.798	0.471	0.790	0.708	0.804	0.817	0.465	0.810
M-6	×	✓	33.12 ± 6.83	0.258 ± 0.116	0.405 ± 0.123	0.711	0.794	0.805	0.452	0.799	0.726	0.812	0.824	0.442	0.818

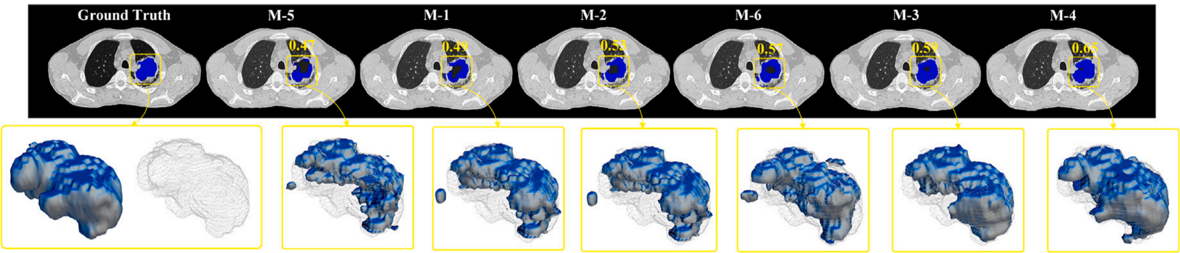


Fig. 8. Qualitative results of lung cavity recognition by various ablation methods. The average DSC is marked on each image.

Table 1 shows the ablation experimental results of DX blocks, MDFFN, and W-GL-MSA modules in the SwinUNeLCsT model. We used the method that includes all the basic modules of SwinUNeLCsT as our baseline method (M-4). Firstly, to validate the effectiveness of our proposed MDFFN in reducing model complexity and enhancing the capability of LCs feature extraction, we replaced the MDFFN approach with the original MLP module of Swin UNetR (M-3). From the control experiment between the baseline method and the M-3 method, we observed that the method based on the MDFFN module outperforms the method using the original MLP module of Swin UNetR in all metrics of LCs recognition (except the REC metric of location classification). Second, from the control group of the baseline method versus the M-2 method, we observed that removing the DX block component from SwinUNeLCsT results in a slight decrease in the overall LCs identification performance. This result confirms the effectiveness of the DX module in capturing spatial features and long-distance dependencies of lung CT scan data. Finally, to validate the effectiveness of the W-GL-MSA, which integrates global and local features in LCs identification, we replaced W-GL-MSA in SwinUNeLCsT with the original W-MSA in SwinUNetR, denoted as method M-1. From the table, we observed that the M-1 method shows a large drop in all metrics compared to the baseline method. This indicates that compared to standard W-MSA, W-GL-MSA, which integrates both global and local features, achieves better recognition of the LCs.

We also performed ablation experiments on the W-GL-MSA module to further verify the effectiveness of W-GL-MSA for WSSS in the LCs, and the results of the experiments are shown in Table 2. First, from the table, we observed that after removing the LSI (lightweight simplified inception) module that extracts local features (M-6), SwinUNeLCsT has a decrease in the overall segmentation and classification performance of LCs. This result shows that the local features of CT images extracted by the LSI module play an important role in enhancing the recognition performance of the LCs. Moreover, after removing the W-SSC-SA (window-based spectral scaled cosine self-attention) module that extracts global features (M-5), SwinUNeLCsT also shows a significant decline in the overall segmentation and classification performance of LCs. This result verifies that the global features of CT images extracted by the W-GL-MSA module also play a key role in enhancing LCs recognition. However, we observed that the SwinUNeLCsT, integrating both LSI and W-SSC-SA module (M-4), achieved the best performance across all metrics. This result corroborates the effectiveness of the W-GL-MSA module, which amalgamates global and local features, in enhancing LCs recognition. Finally, to visually show the recognition effect of each ablation method on LCs features, we visualized the recognition effect of these methods (M-1 to M-6) in Fig. 8. The figure shows that the baseline method (M-4) also achieves the best LCs recognition result compared to other methods.

To further verify the performance of our method in LCs attribute classification, we visualized the ROC curves of each ablation model (M-1 to M-6) in the LCs number and location classification task, as shown in Fig. 9. From the figure, we observed that the method incorporating all SwinUNeLCsT components (M-4) achieved the best AUC in the classification tasks of LCs number and location, with values of 0.737 and 0.762, respectively. Moreover, the W-GL-MSA module method (M-4), which merges global and local features, demonstrated superior classification performance compared to methods containing only global or local features (M-5 or M-6).

Fig. 10 presents the confusion matrices of each ablation method. In these matrices, higher values on the diagonal indicate a greater number of correctly classified lung CT instances. Higher values of the diagonal indicate a greater number of misclassified lung CT instances. From these confusion matrices, we observed that for LCs attribute features with higher similarity, the probability of misclassification is higher. For example, the misclassification rate was higher between “3 cavities” and “more than 3 cavities” and between “both lungs” and “left lung” or “right lung”. Among all ablation methods, the baseline method (M-4) demonstrated the best overall classification performance in the confusion matrix. These experimental results further illustrate the superiority of SwinUNeLCsT in the classification of LCs attributes.

In conclusion, in this ablation study, we first explored the contributions of the MDFFN, DX block, and W-GL-MSA module in the SwinUNeLCsT model. Experimental results show that the MDFFN plays

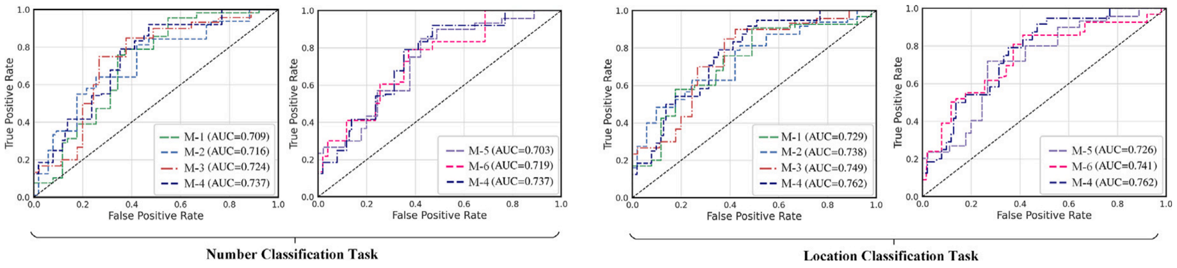


Fig. 9. ROC curves of various ablation methods for lung cavity attribute classification.

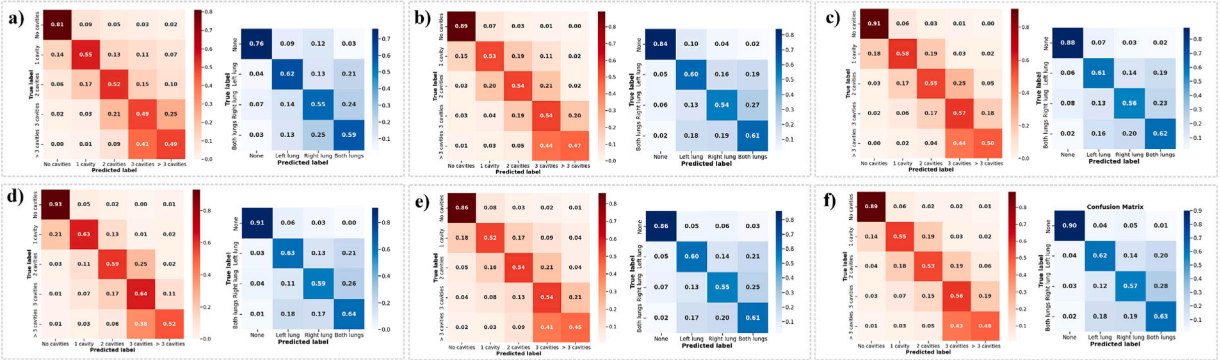


Fig. 10. Confusion matrices for various ablation methods. **(a)–(f)** represent M-1 to M-6 methods in turn.

Table 3
The impact of different classification layer combinations on lung cavity recognition performance.

ID	Num-CLS sub-layer		Loc-CLS sub-layer		Segmentation (± SD)			Number classification					Location classification				
	A	B	A	B	95HD	IoU	DSC	ACC	REC	PRE	FPR	F1	ACC	REC	PRE	FPR	F1
M-7	✓	✓	×	×	34.86 ± 8.67	0.246 ± 0.124	0.388 ± 0.118	0.709	0.789	0.798	0.482	0.793	–	–	–	–	–
M-8	×	×	✓	✓	34.19 ± 10.14	0.249 ± 0.137	0.391 ± 0.143	–	–	–	–	–	0.727	0.814	0.819	0.458	0.816
M-9	×	✓	✓	✓	34.08 ± 8.73	0.257 ± 0.128	0.406 ± 0.138	0.718	0.793	**0.812**	0.448	0.801	0.735	0.818	0.821	0.434	0.820
M-10	✓	×	✓	✓	32.04 ± 6.26	0.259 ± 0.131	0.409 ± 0.124	0.721	0.801	0.809	0.439	0.807	0.738	**0.821**	0.827	0.409	0.824
M-11	✓	✓	×	✓	33.92 ± 7.63	0.255 ± 0.117	0.405 ± 0.121	0.717	0.794	0.805	0.451	0.799	0.733	0.815	0.822	0.436	0.819
M-12	✓	✓	✓	×	32.18 ± 8.58	0.258 ± 0.142	0.408 ± 0.133	0.719	0.797	0.806	0.443	0.801	0.736	0.817	0.824	0.418	0.821
M-4	✓	✓	✓	✓	**31.43 ± 7.46**	**0.267 ± 0.126**	**0.418 ± 0.134**	**0.723**	**0.806**	0.811	**0.421**	**0.809**	**0.742**	**0.821**	**0.835**	**0.398**	**0.828**

Table 4
Quantitative results of various 3D medical weakly supervised semantic segmentation methods.

Methods	95HD	IoU	DSC
Laradji et al. (2021)	36.84 ± 5.94	0.224 ± 0.135	0.368 ± 0.142
Ye et al. (2022)	35.87 ± 6.72	0.231 ± 0.147	0.379 ± 0.143
Lu et al. (2023)	32.65 ± 6.82	0.255 ± 0.129	0.409 ± 0.125
Sun et al. (2023)	35.16 ± 8.76	0.238 ± 0.117	0.387 ± 0.113
SwinUNeLCsT	**31.43 ± 7.46**	**0.267 ± 0.126**	**0.418 ± 0.134**

a critical role in enhancing the model's efficiency in extracting lung features. The introduction of the DX block effectively enhances the model's ability to capture key spatial features and process long-distance dependencies. The W-GL-MSA module enhances the feature extraction capability for LCs of various sizes by integrating global and local features. Secondly, we also explored the role of the W-GL-MSA module in the WSSS of the LCs. Experimental results show that the W-GL-MSA module, which integrates the LSI and W-SSC-SA modules, achieves the best results in LCs recognition. This result verifies the importance of integrating local and global features to improve LCs feature extraction capabilities.

6.2.2. Multi-classification task ablation experiment

To explore the impact of the combination of different classification layers in the LCs WSSS framework proposed in this paper on the LCs recognition performance, we conducted a series of ablation experiments, as shown in Table 3. Firstly, we constructed two ablation experiment settings that only included a single classification task to investigate the impact of LCs location and quantity classification tasks on the performance of WSSS. These are the M-7 method, which represents quantity classification, and the M-8 method, which stands for quantity classification, respectively. From the table, we observed that the method that only includes the number classification task (M-7) has slightly lower segmentation performance than the method (M-8) that only includes the position classification task. This result indicates that compared to quantity classification tasks, the spatial information provided by LCs location classification tasks is more efficient for achieving segmentation of the LCs. Furthermore, compared with training on a single classification task, when jointly training on number and location classification tasks (M-9 to M-12), all LCs recognition indicators have larger improvement. This result shows that jointly training LCs location and number classification tasks is more beneficial for WSSS of LCs than training individual tasks separately. We also observed that the methods (M-4) that include all classification sub-layers also achieve the best performance in all other evaluation indicators except the recall rate (REC) indicator of position classification.

Table 5
Quantitative results of various 3D medical semantic segmentation methods. The number of parameters and GFLOPs (with a single input volume of 128 × 128 × 96 for the fully supervised segmentation tasks) are shown. SwinUNeLCsT† represents a variant of SwinUNeLCsT based on a supervised learning setting.

Methods	GFLOPs	Parameter	95HD	IoU	DSC
nnUnet (Isensee et al., 2021)	637.5	30.8M	26.93 ± 7.17	0.334 ± 0.135	0.482 ± 0.128
TransBTS (Wang et al., 2021)	198.9	33.1M	25.31 ± 8.92	0.341 ± 0.118	0.493 ± 0.130
nnFormer (Zhou et al., 2021a)	1635.6	159.0M	24.92 ± 6.35	0.347 ± 0.129	0.499 ± 0.141
UnetR (Hatamizadeh et al., 2022)	476.4	92.6M	24.53 ± 8.73	0.352 ± 0.125	0.508 ± 0.134
SwinUnetR (Tang et al., 2022)	595.3	62.2M	24.02 ± 6.73	0.355 ± 0.131	0.514 ± 0.137
UnesT (Yu et al., 2023b)	465.2	87.3M	23.97 ± 7.69	0.359 ± 0.129	0.518 ± 0.132
SwinUNeLCsT†	461.6	84.5M	**23.54 ± 6.86**	**0.365 ± 0.113**	**0.526 ± 0.121**

Overall, in this ablation study, we explored the impact of features of LCs location and number classification tasks on the performance of WSSS of LCs. The experimental result verifies that integrating features from LCs location and quantity classification tasks can enhance the performance of WSSS of the LCs.

6.3. Comparison to state-of-the-art

To further evaluate the performance of SwinUNeLCsT, we trained existing WSSS methods on the dataset constructed in this study, with the results shown in Table 4. From the table, we observed that SwinUNeLCsT achieved the best performance in WSSS of LCs, with a 95HD of 31.43, an IoU of 0.267, and a DSC of 0.418.

Moreover, to validate the superiority of the SwinUNeLCsT architecture, we compared it under a fully supervised learning setting with other state-of-the-art medical 3D semantic segmentation models, including nnUnet, TransBTS, nnFormer, UnetR, UnesT, and SwinUNetR, with the results presented in Table 5. Architecturally, under a fully supervised paradigm, SwinUNeLCsT transformed into a multi-task learning model capable of performing both segmentation and classification of LCs. Unlike WSSS, SwinUNeLCsT under full supervision does not require the generation of semantic segmentation pseudo-labels. Therefore, we removed all steps from the CAM output to pseudo-label generation and replaced global max pooling with global average pooling. To distinctly differentiate SwinUNeLCsT under different supervision paradigms, we named the fully-supervised model as SwinUNeLCsT†. From Table 5, we observed that the SwinUNeLCsT† model exhibits superior performance across the 95HD, IoU, and DSC metrics with scores of 23.54, 0.365, and 0.526, respectively. We also observed that SwinUNeLCsT† balances computational efficiency and model complexity well. Firstly, regarding computational efficiency, SwinUNeLCsT† has a lower GFLOPs value at 461.6, indicating fewer floating-point operations per second. This is less computationally intensive than nnUNet, which has a significantly higher GFLOPs value of 637.5. Furthermore, despite its efficiency, SwinUNeLCsT† maintains a moderate parameter count of 84.5M, which is higher than some models, such as TransBTS and nnFormer, but lower than UnetR and SwinUNetR. Secondly, in terms of model complexity, SwinUNeLCsT† finds a middle ground, presenting fewer parameters than the most complex model, UnetR, but more than the less complex TransBTS and nnFormer. Finally, when considering performance, SwinUNeLCsT† leads in all three metrics evaluated.

In conclusion, this ablation experiment shows that the SwinUNeLCsT model performs best in both weakly supervised and fully supervised paradigms. Meanwhile, its excellent performance also highlights the superiority of the SwinUNeLCsT model based on the CNN–transformer hybrid architecture in identifying LCs.

7. Discussion

7.1. The application of weakly supervised semantic segmentation in tuberculosis diagnosis

WSSS is particularly important in medical image analysis, as acquiring a large volume of precisely annotated medical images is often both expensive and time-consuming. In this study, we developed a WSSS framework for the LCs, aimed at improving the accuracy and efficiency of medical diagnoses. The core of this framework is to use low-cost classification annotations to train a deep-learning model capable of identifying and segmenting LCs in lung CT. Moreover, we conducted a series of experiments to demonstrate the effectiveness of this framework on various lung image datasets and compared it with traditional fully supervised learning methods. The experimental results show that, even with limited annotated data, our weakly supervised framework can achieve satisfactory segmentation accuracy, demonstrating its great potential in addressing the high-cost problem of annotating LCs images in TB diagnosis.

7.1.1. Enhanced diagnostic efficiency

The application of our WSSS framework can significantly expedite the process of LCs analysis in lung CT scans. By automating the segmentation process, radiologists can quickly identify areas of interest, leading to faster diagnosis and treatment planning. This not only improves workflow efficiency but also enhances patient throughput in medical facilities.

7.1.2. Integration with clinical workflows

Implementing our WSSS framework within clinical workflows could facilitate assistance during diagnostic procedures. This integration could provide radiologists with immediate insights, enhancing decision-making processes and potentially leading to more accurate diagnoses.

7.2. Why do we need an efficient transformer-based LCs segmentation model?

Unlike other WSSS methods designed for various pulmonary CT diseases, SwinUNeLCsT focuses on the segmentation of the LCs, avoiding overly broad areas of interest, and thus can more efficiently focus on key features. On the other hand, pure transformer architecture networks sometimes miss key local features due to the lack of inductive bias of CNN in processing image data, such as the nnFormer model. In contrast, although pure CNN architectures (such as nnUNet) can effectively capture local features through convolution operations, they have limitations in capturing long-range feature dependencies in images. To solve these limitations, some medical imaging models adopt CNN–transformer hybrid architecture, such as TransBTS, UnetR, UnesT, and SwinUNetR. One thing these hybrid models have in common is that the way the transformer interacts with the CNN module is by concatenating features sequentially. However, in the SwinUNeLCsT model constructed in this paper, the transformer and CNN modules mainly interact through element-wise addition at the channel level. Moreover, a large number of experimental results show that this interactive method can more closely and efficiently combine local and global features, thereby more accurately utilizing the spatial information in CT data.

7.3. Principal findings and their significance

In our research, to enhance the recognition capabilities of the pulmonary LCs, we introduced a novel window-based hybrid global–local multi-head self-attention mechanism. This mechanism effectively combines the local feature recognition capabilities of CNN with the global information extraction process of transformer architecture networks, achieving efficient integration of global context and local detail information. On the other hand, to address the high computational complexity challenge posed by the self-attention mechanism within transformer models, we designed an innovative spectral scaling cosine multi-head self-attention mechanism. This mechanism initially reduces the dimensionality of the feature space through point convolutions (using $1 \times 1 \times 1$ kernels), followed by capturing spatial information with fewer parameters through depthwise convolutions (using $3 \times 3 \times 3$ kernels). By generating lower-dimensional query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) vectors, this approach alleviates the computational burden of the attention mechanism. Moreover, by incorporating cosine similarity, we reduce the need for large-scale matrix multiplication operations. The resulting transposed attention map A, sized $\mathbb{R}^{C \times C \times C}$, is based on channel dimensions rather than the traditional spatial dimensions. Compared to the attention maps generated by standard self-attention mechanisms, it features a smaller size and lower spatial complexity.

7.4. Limitations and future work

Although SwinUNeLCsT has achieved state-of-the-art performance in the field of LCs recognition, its hybrid architecture based on CNN and transformer is not as good as the baseline model SwinUnetR in terms of computational efficiency and model complexity. However, this challenge reduces its usefulness in clinical diagnostic applications as it may result in longer identification times. To address this issue, our future research will focus on finding strategies to further improve computational efficiency and reduce model complexity. For example, we will explore sparse attention techniques suitable for LCs recognition to further reduce the computational complexity of the self-attention mechanism.

8. Conclusion

In this study, we present an innovative model for 3D TB image analysis (SwinUNeLCsT). This model employs a hybrid architecture integrating transformer and CNN, effectively consolidating global and local features in TB CT imaging to improve recognition of LCs of various sizes. To demonstrate SwinUNeLCsT's effectiveness, we collaborated with radiologists to develop the first WSSS CT dataset for TB LCs. Using this dataset, we trained SwinUNeLCsT under both weakly-supervised and supervised semantic segmentation paradigms. In the supervised semantic segmentation paradigm, SwinUNeLCsT demonstrated superior performance, surpassing current popular medical 3D supervised semantic segmentation methods. Similarly, in the WSSS paradigm, SwinUNeLCsT also achieved the best performance among the state-of-the-art medical 3D WSSS methods. This innovative WSSS approach not only reduces the labeling burden but also sets a new benchmark for WSSS methods in medical imaging, especially in the field of TB LCs CT medical analysis. In addition, we also introduce a novel optimization strategy for multi-task training of the SwinUNeLCsT model under weak supervision settings. Experimental results show that this strategy effectively alleviates gradient conflict and enhances the identification efficiency of LCs. Overall, these achievements not only verify the effectiveness of SwinUNeLCsT but also highlight its forward-looking and practicality in handling complex medical image analysis tasks.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hizmawati Madzin reports financial support was provided by the Malaysia Ministry of Higher Education and Universiti Putra Malaysia. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Malaysia Ministry of Higher Education under the Fundamental Research Grant Scheme [Project Code: FRGS/1/2022/ICT01/UPM/02/1]; and in part by Universiti Putra Malaysia under the Geran Inisiatif Putra Siswazah (GP-IPS) [Project Code: GP-IPS/2023/9773600].

References

Alebiosu, D.O., Dharmaratne, A., Lim, C.H., 2023. Improving tuberculosis severity assessment in computed tomography images using novel davou-net segmentation and deep learning framework. Expert Syst. Appl. 213, 119287.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al., 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In: International Conference on Machine Learning. PMLR, pp. 173–182.

Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2021. Deep semantic segmentation of natural and medical images: a review. Artif. Intell. Rev. 54, 137–178.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning. Adv. Neural Inf. Process. Syst. 32.

Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S., 2022a. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11676–11685.

Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q., 2022b. Class re-activation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 969–978.

Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Dartois, V.A., Rubin, E.J., 2022. Anti-tuberculosis treatment strategies and drug development: challenges and priorities. Nat. Rev. Microbiol. 20, 685–701.

Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L., 2022. Davit: Dual attention vision transformers. In: European Conference on Computer Vision. Springer, pp. 74–92.

Dong, Y., Gao, D., Qiu, T., Li, Y., Yang, M., Shi, G., 2023. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22262–22271.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gordaliza, P., Vaquero, J.J., Sharpe, S., Gleeson, F., Muñoz-Barrutia, A., 2019. Tuberculosis lesions in ct images inferred using 3d-cnn and multi-task learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 294–297.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Huang, Z.A., Hu, Y., Liu, R., Xue, X., Zhu, Z., Song, L., Tan, K.C., 2022. Federated multi-task learning for joint diagnosis of multiple mental disorders on mri scans. IEEE Trans. Biomed. Eng. 70, 1137–1149.

Huang, G., Laradji, I., Vazquez, D., Lacoste-Julien, S., Rodriguez, P., 2021. A survey of self-supervised and few-shot object detection. arXiv preprint arXiv:2110.14711.

Iqbal, A., Usman, M., Ahmed, Z., 2023. Tuberculosis chest x-ray detection using cnn-based hybrid segmentation and classification approach. Biomed. Signal Process. Control 84, 104667.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211.

Kanai, S., Fujiwara, Y., Iwamura, S., 2017. Preventing gradient explosions in gated recurrent units. Adv. Neural Inf. Process. Syst. 30.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 876–885.

Kirsch, A., Van Amersfoort, J., Gal, Y., 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Adv. Neural Inf. Process. Syst. 32.

Laradji, I., Rodriguez, P., Manas, O., Lensink, K., Law, M., Kurzman, L., Parker, W., Vazquez, D., Nowrouzezahrai, D., 2021. A weakly supervised consistency-based learning method for covid-19 segmentation in ct images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2453–2462.

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348.

Li, R., Mai, Z., Zhang, Z., Jang, J., Sanner, S., 2023b. Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. J. Vis. Commun. Image Represent. 92, 103800.

Li, K., Qian, Z., Han, Y., Eric, I., Chang, C., Wei, B., Lai, M., Liao, J., Fan, Y., Xu, Y., 2023a. Weakly supervised histopathology image segmentation with self-attention. Med. Image Anal. 86, 102791.

Liang, X., Wang, X., Lei, Z., Liao, S., Li, S.Z., 2017. Soft-margin softmax for deep classification. In: International Conference on Neural Information Processing. Springer, pp. 413–421.

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, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022a. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q., 2021a. Conflict-averse gradient descent for multi-task learning. Adv. Neural Inf. Process. Syst. 34, 18878–18890.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Liu, Y., Wu, Y.H., Wen, P., Shi, Y., Qiu, Y., Cheng, M.M., 2020. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 44, 1415–1428.

Lu, F., Zhang, Z., Liu, T., Tang, C., Bai, H., Zhai, G., Chen, J., Wu, X., 2023. A weakly supervised inpainting-based learning method for lung ct image segmentation. Pattern Recognit. 144, 109861.

Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., Cheng, J.Z., 2020. Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. IEEE Trans. Med. Imaging 40, 2698–2710.

Ru, L., Zhan, Y., Yu, B., Du, B., 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16846–16855.

Ru, L., Zheng, H., Zhan, Y., Du, B., 2023. Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102.

Ruby, U., Yendapalli, V., 2020. Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng. 9.

Sainz de Cea, M.V., Diedrich, K., Bakalo, R., Ness, L., Richmond, D., 2020. Multi-task learning for detection and classification of cancer in screening mammography. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8 2020, Proceedings, Part VI 23. Springer, pp. 241–250.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3136–3145.

Su, Y., Cheng, M., Yuan, Z., Liu, W., Zeng, W., Wang, C., 2023. Multi-stage scene-level constraints for large-scale point cloud weakly supervised semantic segmentation. IEEE Trans. Geosci. Remote Sens..

Sun, W., Feng, X., Liu, J., Ma, H., 2023. Weakly supervised segmentation of covid-19 infection with local lesion coherence on ct images. Biomed. Signal Process. Control 79, 104099.

Sun, G., Wang, W., Dai, J., Van Gool, L., 2020. Mining cross-image semantics for weakly supervised semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28 2020, Proceedings, Part II 16. Springer, pp. 347–365.

Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Muhammad, N., 2020. A survey of end-to-end driving: Architectures and training methods. IEEE Trans. Neural Netw. Learn. Syst. 33, 1364–1384.

Tan, Z., Madzin, H., Ding, Z., 2022. Semi-supervised semantic segmentation methods for uw-octa diabetic retinopathy grade assessment. In: MICCAI Challenge on Mitosis Domain Generalization. Springer, pp. 97–117.

Tan, Z., Madzin, H., Ding, Z., 2023a. Image quality assessment based on multi-model ensemble class-imbalance repair algorithm for diabetic retinopathy uw-octa images. In: Mitosis Domain Generalization and Diabetic Retinopathy Analysis: MICCAI Challenges MIDOG 2022 and DRAC 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18–22 2022, Proceedings. Springer, pp. 118–126.

Tan, Z., Madzin, H., Ding, Z., 2023b. Semi-supervised semantic segmentation methods for uw-octa diabetic retinopathy grade assessment. In: Mitosis Domain Generalization and Diabetic Retinopathy Analysis: MICCAI Challenges MIDOG 2022 and DRAC 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18–22 2022, Proceedings. Springer, pp. 97–117.

Tan, Z., Madzin, H., Norafida, B., ChongShuang, Y., Sun, W., Nie, T., Cai, F., 2024. Deeppulmotb: A benchmark dataset for multi-task learning of tuberculosis lesions in lung computerized tomography (ct). Heliyon.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Tian, Y., Si, L., Zhang, X., Cheng, R., He, C., Tan, K.C., Jin, Y., 2021. Evolutionary large-scale multi-objective optimization: A survey. ACM Comput. Surv. 54, 1–34.

Ullah, U., Saleem, S., Farooq, M., Yameen, B., Cheema, M.I., 2024. Lipoarabinomannan-based tuberculosis diagnosis using a fiber cavity ring down biosensor. Biomed. Opt. Express 15, 1428–1436.

Unal, O., Dai, D., Van Gool, L., 2022. Scribble-supervised lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2697–2707.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, J., Bhalerao, A., Yin, T., See, S., He, Y., 2024. Camanet: class activation map guided attention network for radiology report generation. IEEE J. Biomed. Health Inf..

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021. Transbts: Multimodal brain tumor segmentation using transformer. 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. 109–119.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284.

Wen, S., Liu, J., Xu, W., 2021. A novel lesion segmentation algorithm based on u-net network for tuberculosis ct image. In: 2021 International Conference on Control, Automation and Information Sciences. ICCAIS, IEEE, pp. 909–914.

Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H., 2021. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16765–16774.

Xu, R., Yu, Y., Ho, J., Yang, C., 2023. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2501–2505.

Ye, Q., Gao, Y., Ding, W., Niu, Z., Wang, C., Jiang, Y., Wang, M., Fang, E.F., Menpes-Smith, W., Xia, J., et al., 2022. Robust weakly supervised learning for covid-19 recognition using multi-center ct images. Appl. Soft Comput. 116, 108291.

Yoo, D., Kweon, I.S., 2019. Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 93–102.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C., 2020. Gradient surgery for multi-task learning. Adv. Neural Inf. Process. Syst. 33, 5824–5836.

Yu, L., Xiang, W., Fang, J., Chen, Y.P.P., Chi, L., 2023a. Ex-vit: A novel explainable vision transformer for weakly supervised semantic segmentation. Pattern Recognit. 142, 109666.

Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al., 2023b. Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Med. Image Anal. 90, 102939.

Yushkevich, P.A., Gao, Y., Gerig, G., 2016. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 3342–3345.

Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., 2022. Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739.

Zhang, R., Liu, S., Yu, Y., Li, G., 2021. Self-supervised correction learning for semi-supervised biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. Springer, pp. 134–144.

Zhang, M., Zhou, Y., Zhao, J., Man, Y., Liu, B., Yao, R., 2020. A survey of semi- and weakly supervised semantic segmentation of images. Artif. Intell. Rev. 53, 4259–4288.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C., 2022. Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481.

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.T., Shen, D., 2021b. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. Med. Image Anal. 70, 101918.

Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2021a. Nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.