



**MULTI-TASK LEARNING AND WEAKLY SUPERVISED SEMANTIC
SEGMENTATION FOR CLASSIFICATION AND SEGMENTATION OF
PULMONARY TUBERCULOSIS LESIONS IN CT IMAGES**

By

TAN ZHUOYI

**Thesis Submitted to the School of Graduate Studies, Universiti Putra
Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of
Philosophy**

July 2024

FSKTM 2024 24

All material contained within the thesis, including without limitation text, logos, icons, photographs, and all other artwork, is copyright material of Universiti Putra Malaysia unless otherwise stated. Use may be made of any material contained within the thesis for non-commercial purposes from the copyright holder. Commercial use of material may only be made with the express, prior, written permission of Universiti Putra Malaysia.

Copyright © Universiti Putra Malaysia



Abstract of thesis presented to the Senate of Universiti Putra Malaysia in
fulfilment of the requirement for the degree of Doctor of Philosophy

**MULTI-TASK LEARNING AND WEAKLY SUPERVISED SEMANTIC
SEGMENTATION FOR CLASSIFICATION AND SEGMENTATION OF
PULMONARY TUBERCULOSIS LESIONS IN CT IMAGES**

By

TAN ZHUOYI

July 2024

Chairman : Hizmawati binti Madzin, PhD
School : Computer Science and Information Technology

Tuberculosis (TB) remains a significant global health challenge, characterized by high incidence and mortality rates on a global scale. With the rapid advancement of computer-aided diagnosis (CAD) tools in recent years, CAD has assumed an increasingly crucial role in supporting TB diagnosis. Nonetheless, developing such tools for TB diagnosis heavily relies on well-annotated computerized tomography (CT) datasets. However, currently, the available annotations in TB CT datasets are still limited. For instance, the types of annotated lesions are relatively singular, which to some extent constrains the further development of CAD tools for TB diagnosis. To address this limitation, this research introduces DeepPulmoTB, a CT multi-task learning dataset explicitly designed for pulmonary TB comprehensive radiological feature analysis. This dataset contains three key diagnostic indicators for TB, which are: i) lung cavity, ii) lung areas, and iii) C-LCW (consolidation or LC wall). To demonstrate the advantages of DeepPulmoTB, this research

proposes a novel comprehensive radiological feature analysis framework for TB imaging based on multi-task learning. This framework includes a multi-task learning model, namely DPTBNet, for the joint segmentation and classification of lesion tissues in CT images. The architecture of DPTBNet comprises two subnets, which are: i) SwinUnetR for the segmentation task, and ii) a lightweight multi-scale network for the classification task. In addition, to address the problem that traditional deep learning methods based on supervised learning paradigms need to rely on a large amount of accurate pixel-level data for the segmentation of lung cavity, this research constructs a weakly supervised semantic segmentation framework for lung cavity segmentation. Through this framework, radiologists need only to classify the number and the approximate location (e.g., left lung, right lung, or both) of the lung cavity in the CT scan to achieve efficient segmentation of this kind of lesion. This process eliminates the need for meticulously drawing boundaries, greatly reducing the cost of annotation. In this framework, to effectively improve the ability to identify lung cavity lesions, this research designs a novel CT medical imaging analysis model for TB, named SwinUNeLCsT. SwinUNeLCsT is a hybrid model comprising a convolutional neural network and transformer. This model aims to improve the ability to obtain lung cavity imaging features by efficiently integrating the local texture information extracted by the convolutional network and the high-level semantic information captured by the transformer. This research validates the effectiveness of the DPTBNet model and the practical value of the DeepPulmoTB dataset in TB diagnosis tasks through extensive experiments. In fully supervised segmentation tasks, SwinUNeLCsT outperforms current mainstream 3D

medical segmentation models, achieving a 95% Hausdorff Distance (95HD) of 23.54 ± 6.86 , Intersection over Union (IoU) of 0.365 ± 0.113 , and Dice Similarity Coefficient (DSC) of 0.526 ± 0.121 . Additionally, in the weakly supervised learning framework, SwinUNeLCsT demonstrates superior performance, achieving a 95HD of 31.43 ± 7.46 , IoU of 0.267 ± 0.126 , and DSC of 0.418 ± 0.134 , outperforming existing 3D medical weakly supervised learning methods. These experimental results fully demonstrate the excellent applicability and developmental potential of the methodologies proposed in this study in the field of intelligent analysis of TB imaging.

Keywords: Class Activation Mapping, Healthcare, Multi-Task Learning, Tuberculosis, Weakly Supervised Semantic Segmentation

SDG: GOAL 3: Good Health and Well-Being

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**PEMBELAJARAN PELBAGAI TUGAS DAN SEGMENTASI SEMANTIK
TERSELIA LEMAH UNTUK KLASIFIKASI DAN SEGMENTASI LESI
TUBERKULOSIS PARU-PARU DALAM IMEJ CT**

Oleh

TAN ZHUOYI

Julai 2024

Pengerusi : Hizmawati binti Madzin, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Tuberkulosis (TB) kekal menjadi cabaran kesihatan global yang besar, dicirikan oleh kadar kejadian dan kematian yang tinggi di seluruh dunia. Dengan kemajuan pesat alat bantuan diagnosis berkomputer (CAD) dalam beberapa tahun kebelakangan ini, CAD memainkan peranan yang semakin penting dalam menyokong diagnosis TB. Walau bagaimanapun, pembangunan alat-alat sebegini untuk diagnosis TB sangat bergantung kepada dataset tomografi berkomputer (CT) yang diberi anotasi dengan baik. Pada masa ini, anotasi yang tersedia dalam dataset CT TB masih terhad. Sebagai contoh, jenis lesi yang dianotasi adalah agak terbatas, yang sedikit sebanyak mengehadkan pembangunan lanjut alat CAD untuk diagnosis TB. Bagi menangani kekurangan ini, kajian ini memperkenalkan DeepPulmoTB, sebuah dataset pembelajaran pelbagai tugas CT yang direka khas untuk analisis ciri radiologi TB paru-paru secara menyeluruh. Dataset ini mengandungi tiga penunjuk diagnostik utama untuk TB, iaitu: i) kaviti paru-

paru, ii) kawasan paru-paru, dan iii) C-LCW (konsolidasi atau dinding LC).

Untuk menunjukkan kelebihan DeepPulmoTB, kajian ini mencadangkan satu rangka kerja analisis ciri radiologi komprehensif baru bagi pencitraan TB berdasarkan pembelajaran pelbagai tugas. Rangka kerja ini termasuk model pembelajaran pelbagai tugas, iaitu DPTBNet, untuk segmentasi dan klasifikasi bersama bagi tisu lesi dalam imej CT. Reka bentuk DPTBNet terdiri daripada dua subnet, iaitu: i) SwinUnetR untuk tugas segmentasi, dan ii) rangkaian pelbagai skala ringan untuk tugas klasifikasi. Selain itu, bagi menangani masalah kaedah pembelajaran mendalam tradisional yang berdasarkan paradigma pembelajaran terselia yang memerlukan sejumlah besar data piksel tepat untuk segmentasi kaviti paru-paru, kajian ini membina satu rangka kerja segmentasi semantik terselia lemah untuk segmentasi kaviti paru-paru. Melalui rangka kerja ini, pakar radiologi hanya perlu mengklasifikasikan bilangan dan lokasi anggaran (contohnya, paru-paru kiri, paru-paru kanan, atau kedua-duanya) bagi kaviti paru-paru dalam imbasan CT untuk mencapai segmentasi jenis lesi ini dengan cekap. Proses ini menghapuskan keperluan untuk melukis sempadan dengan teliti, sekaligus mengurangkan kos anotasi dengan ketara. Dalam rangka kerja ini, bagi meningkatkan keupayaan untuk mengenal pasti lesi kaviti paru-paru, kajian ini mereka bentuk model analisis pencitraan perubatan CT baru untuk TB, yang dinamakan SwinUNeLCsT. SwinUNeLCsT ialah model hibrid yang menggabungkan rangkaian neural konvolusi dan transformer. Model ini bertujuan untuk meningkatkan keupayaan mendapatkan ciri pencitraan kaviti paru-paru dengan mengintegrasikan maklumat tekstur tempatan yang diekstrak oleh rangkaian konvolusi dan maklumat semantik tahap tinggi yang ditangkap oleh

transformer. Kajian ini mengesahkan keberkesanan model DPTBNet dan nilai praktikal dataset DeepPulmoTB dalam tugas diagnosis TB melalui eksperimen yang meluas. Dalam tugas segmentasi terselia sepenuhnya, SwinUNeLCsT mengatasi model segmentasi perubatan 3D arus perdana, mencapai Jarak Hausdorff 95% (95HD) sebanyak 23.54 ± 6.86 , Intersection over Union (IoU) sebanyak 0.365 ± 0.113 , dan Dice Similarity Coefficient (DSC) sebanyak 0.526 ± 0.121 . Selain itu, dalam rangka kerja pembelajaran terselia lemah, SwinUNeLCsT menunjukkan prestasi unggul, mencapai 95HD sebanyak 31.43 ± 7.46 , IoU sebanyak 0.267 ± 0.126 , dan DSC sebanyak 0.418 ± 0.134 , mengatasi kaedah pembelajaran terselia lemah perubatan 3D sedia ada. Keputusan eksperimen ini sepenuhnya menunjukkan kebolegunaan yang cemerlang dan potensi pembangunan metodologi yang dicadangkan dalam kajian ini dalam bidang analisis pintar bagi pencitraan TB.

Kata Kunci: Pembelajaran Pelbagai Tugas, Pemetaan Pengaktifan Kelas, Penjagaan Kesihatan, Segmentasi Semantik Berpandu Lemah, Tuberkulosis

SDG: MATLAMAT 3: Kesihatan dan Kesejahteraan yang Baik

ACKNOWLEDGEMENTS

Time flies, and in the blink of an eye, three years have passed since I embarked on my doctoral journey. Looking back to when I first started my Ph.D., the confusion and anxiety I experienced are still vivid in my mind. However, now, I have accumulated valuable skills and knowledge in my field of expertise. Throughout these challenging three years of study, I have received warm care and selfless help from many teachers, classmates, and colleagues. I would like to express my heartfelt gratitude and sincere thanks to them!

First and foremost, I would like to extend my deepest gratitude to my supervisor, Dr. Hizmawati Binti Madzin! Her rigorous scholarship, profound knowledge, humility, and broad-mindedness have set a shining example for me. Over the past three years, under her strict guidance and patient mentorship, she has provided meticulous advice on my research and patiently offered suggestions for my paper writing, helping to refine the theme, content, and structure of my thesis. Her patience in guiding my revisions and her meticulous care has left an impressive mark on me.

Lastly, I also want to thank my parents and all the relatives and friends who have supported me throughout these three years of studies! Their selfless warmth and care have given me the courage to face one challenge after another in both my academic and personal life. I will forever cherish their support. I am committed to continuing my efforts, striving diligently, and

continuously improving myself to make meaningful contributions to society and my country.



This thesis was submitted to the Senate of the Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Hizmawati binti Madzin, PhD

Associate Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Chairman)

Fatimah binti Khalid, PhD

Associate Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

Puteri Suhaiza binti Sulaiman, PhD

Associate Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

Rahmita Wirza binti O.K. Rahmat, PhD

Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 13 March 2025

TABLE OF CONTENTS

	Page
ABSTRACT	i
ABSTRAK	iv
ACKNOWLEDGEMENTS	vii
APPROVAL	ix
DECLARATION	xi
LIST OF TABLES	xvi
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xix

CHAPTER

1	INTRODUCTION	1
	1.1 Background and Problem Statement	1
	1.2 Research Questions	4
	1.2.1 Challenges in Pulmonary Tuberculosis CT Dataset Annotation	4
	1.2.2 Limitations of Current Approaches in Pulmonary Tuberculosis Lesion Identification in CT Images	5
	1.2.3 Limitations of Conventional Single-Task Learning Models in Analyzing Pulmonary Tuberculosis Lesion Information	6
	1.2.4 Time-Consuming Annotation of Pixel-Level Pulmonary Tuberculosis Lung Cavity	8
	1.3 Research Motivation	8
	1.4 Research Questions	9
	1.5 Research Objectives	10
	1.6 Research Scope	12
	1.7 Research Contributions	12
	1.8 Thesis Organization	15
2	LITERATURE REVIEW	19
	2.1 Introduction	19
	2.2 Overview of Pulmonary Tuberculosis CT Dataset	19
	2.2.1 Pulmonary Tuberculosis Lesions Recognition Dataset	20
	2.2.2 Pulmonary Tuberculosis Lesions Descriptive Analysis Dataset	21
	2.3 Medical Image Analysis Projects and Software	22
	2.3.1 Medical Open Network for AI	22
	2.3.2 ITK-SNAP	23
	2.3.3 ParaView	23
	2.4 Deep Learning Networks for Medical CT Imaging	24
	2.4.1 Convolutional Neural Network (CNN)	25
	2.4.2 Transformer-Based Network	36

2.4.3	Application in Medical CT Imaging Analysis	42
2.4.4	Deep Learning Networks for Pulmonary Tuberculosis CT Imaging Analysis	49
2.5	Multi-Task Learning	53
2.5.1	Gradient Conflict Problem	56
2.5.5	Multi-Task Learning Methods in Medical CT Imaging Analysis	58
2.6	Weakly Supervised Semantic Segmentation	61
2.6.1	Class Activation Mapping	64
2.6.1.2	Weakly Supervised Semantic Segmentation for Medical CT Imaging	69
2.7	Summary	73
3	METHODOLOGY	76
3.1	Introduction	76
3.2	Research Design and Setting	78
3.2.1	Stage 1: Formulating the Research Problems	78
3.2.2	Stage 2: Acquisition of Pulmonary Tuberculosis CT Imaging Data	80
3.2.3	Stage 3: Construction of Pulmonary Tuberculosis Analysis Datasets	82
3.2.4	Stage 4: Model and Algorithm Development	91
3.2.5	Stage 5: Experiments, Model Training, and Evaluation	94
3.4	Summary	104
4	ANNOTATION PROCESS AND GUIDELINES	106
4.1	Introduction	106
4.2	Pulmonary Tuberculosis Lesion Multi-Task Learning Dataset (DeepPulmoTB)	107
4.2.1	Data Sources	107
4.2.2	Data Annotation	110
4.2.3	DeepPulmoTB Data Visualization	120
4.3	Pulmonary Tuberculosis Lung Cavity Weakly Supervised Semantic Segmentation Dataset (DeepLungCavi-WSSS)	122
4.3.1	Data sources	122
4.3.2	Data annotation	124
4.4	Summary	125
5	COMPREHENSIVE RADIOLOGICAL FEATURE ANALYSIS FRAMEWORK FOR PULMONARY TUBERCULOSIS CT BASED ON MULTI-TASK LEARNING	126
5.1	Introduction	126
5.2	Description of the Proposed Multi-Task Learning Network (DPTBNet)	128
5.2.1	Multi-Scale Feature Extraction	136
5.3	Description of the Proposed Comprehensive Radiological Feature Analysis Framework for Pulmonary Tuberculosis	138

5.3.1	Data Preprocessing	138
5.3.2	Improved Version of Iterative Algorithm	139
5.3.3	Network Training	144
5.4	Experiment Details	150
5.5	Result and Discussion	152
5.5.1	Pulmonary Tuberculosis Lesions and Lung Areas Segmentation	152
5.5.2	Classification of Lung Cavity Attributes	154
5.5.3	Multi-Task Learning	162
5.6	Summary	166
6	LUNG CAVITY WEAKLY SUPERVISED SEMANTIC SEGMENTATION FRAMEWORK	167
6.1	Introduction	167
6.2	Description of the Proposed Hybrid CNN-Transformer Network	168
6.2.1	SwinUNeLCsT Encoder	170
6.2.2	Multi-DConv Feed-Forward Network	171
6.2.3	Window-Based Global-Local Multi-Head Self-Attention	172
6.2.4	SwinUNeLCsT Decoder	177
6.3	Description of the Lung Cavity Weakly Supervised Semantic Segmentation Framework	179
6.3.1	CFFG	181
6.3.2	Semantic Segmentation Pseudo Label Generation	182
6.3.3	Network training	188
6.4	Experiment Details	192
6.5	Experiment Results	193
6.5.1	Generation of Lung Cavity Region of Interest	194
6.5.2	Ablation studies	200
6.5.2.3	Comparison to State-of-the-Art	213
6.6	Summary	215
7	CONCLUSION AND FUTURE WORKS	217
7.1	Conclusions	217
7.2	Limitations and Recommendations for Future Research	223
	REFERENCES	227
	APPENDICES	244
	BIODATA OF STUDENT	245
	LIST OF PUBLICATIONS	246

LIST OF TABLES

Table	Page
2.1 Existing Pulmonary TB Lesions Recognition Datasets	20
2.2 Comparison of Methods in Medical CT Imaging Analysis	43
2.3 Summary of Deep Learning Methods for Pulmonary Tuberculosis CT Imaging Analysis	51
2.4 Comparison of Multi-Task Learning Methods in Medical CT Imaging Analysis	62
2.5 Comparison of Weakly Supervised Semantic Segmentation Methods in Medical CT Imaging Analysis	72
3.1 Hardware Configuration	94
3.2 Software Configuration	95
3.3 Structure of Confusion Matrix	102
4.1 Annotated Classification of the Lung Cavity Size and Number in a Patient's CT scan	111
5.1 Design Schemes of the Ablation Model	148
5.2 Performance of DPTBNet on the 95HD Metric with Respect to Weight Parameters and Number of Iterations	151
5.3 IoU Performance of DPTBNet Across Different Weight Parameters and Number of Iterations	151
5.4 DSC Performance of DPTBNet Across Different Weight Parameters and Number of Iterations	153
5.5 Segmentation and Classification Task Performance of Ablation Models	155
5.6 Evaluation of Classification Performance Using Different Multi-Scale Features	155
5.7 The Effects of Different Values of σ on Multi-Task Learning	156
5.8 The Effects of Different Values of μ on Multi-Task Learning	159
5.9 Comparison with Existing Multi-Task Learning Methods	159
6.1 Design of the Ablation Study for Different Modules in the SwinUNeLCsT Model	203

6.2	Impact of Various Module Ablation Methods on Weakly Supervised Semantic Segmentation for Lung Cavity	207
6.3	Design of the Ablation Study for Different Classification Layer Combinations in the SwinUNeLCsT Model	210
6.4	Impact of Various Module Ablation Methods on Weakly Supervised Semantic Segmentation for Lung Cavity	211
6.5	Quantitative Results of Various 3D Medical Weakly Supervised Semantic Segmentation Methods	211
6.6	Quantitative Results of Various 3D Medical Fully Supervised Semantic Segmentation Methods	212



LIST OF FIGURES

Figure		Page
2.1	Structure of Convolutional Neural Network	25
2.2	Implementation Principle of Standard 3D Convolution	26
2.3	Implementation Principle of 3D Depthwise Convolution	29
2.4	Implementation Principle of Global Max Pooling and Global Average Pooling	31
2.5	The Principle of Self-Attention Mechanism	37
2.6	Structure of the Standard Transformer Block	38
2.7	Depiction of Hard and Soft Parameter Sharing Mechanisms (Source: Deng et al., 2020)	53
2.8	Conflicting Gradients and PCGrad	58
2.9	Implementation Principle of Class Activation Mapping	65
3.1	Quantitative Research Methodology	77
3.2	Comparison Between the ImageCLEF TB 2022 Dataset and the DeepPulmoTB Dataset	83
3.3	Annotated Information Included in the DeepPulmoTB Dataset	84
3.4	Annotation Information Included in the DeepLungCavi-WSSS Dataset	87
3.5	Visualization of DeepLungCavi-WSSS Dataset Samples	88
3.6	Implementation Process of the Comprehensive Radiological Feature Analysis Framework of Tuberculosis CT	91
3.7	Implementation Process of the Lung Cavity Weakly Supervised Semantic Segmentation Framework	92

LIST OF ABBREVIATIONS

95HD	95th Percentile Hausdorff Distance
ACC	Accuracy
AI	Artificial Intelligence
AUC	Area Under the ROC Curve
BN	Batch Normalization
C-LCW	Consolidation or Lung Cavity Wall
CAD	Computer-Aided Diagnosis
CAM	Class Activation Mapping
CAM-Loc	CAM for Lung Cavity Location Classification
CAM-N	CAM for Lung Cavity Number Classification
CFFG	CAM Feature Fusion Gate
CLEF	Cross-Language Evaluation Forum
CNN	Convolutional Neural Network
CT	Computerized Tomography
DConv	Depthwise Separable Convolution
DSC	Dice Similarity Coefficient
F1	F1-score
FC	Fully Connected
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GAP	Global Average Pooling
GELU	Gaussian Error Linear Unit
GMP	Global Max Pooling
Grad-CAM	Gradient-Weighted Class Activation Mapping

IoU	Intersection over Union
ITK	Insight Segmentation and Registration Toolkit
LC	Lung Cavity
LSI	Lightweight Simplified Inception
MDFFN	Multi-DConv Feed-Forward Network
MLP	Multi-Layered Perceptron
MONAI	Medical Open Network for AI
MSA	Multi-head Self-Attention
MSE	Mean Squared Error
MTL	Multi-Task Learning
NIH	National Institutes of Health
NLP	Natural Language Processing
NTM-PD	Nontuberculous Mycobacterial Pulmonary Disease
PRE	Precision
REC	Recall
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic Curve
SSC-MSA	Spectral Scaled Cosine Multi-Head Self-Attention
STL	Single-Task Learning
SW-MSA	Sliding Window Partitioned Multi-Head Self-Attention
TB	Tuberculosis
TN	True Negatives
TP	True Positives
W-GL-MSA	Window-based Global-Local Multi-Head Self-Attention
W-SSC-SA	Window-based Spectral Scaled Cosine Self-Attention
WSMIS	Weakly Supervised Medical Image Segmentation





CHAPTER 1

INTRODUCTION

1.1 Background and Problem Statement

Over the past 130 years, pulmonary TB (Tuberculosis) has persisted as a formidable global health challenge, capturing extensive attention in the field of public health due to its high incidence and mortality rates. Currently, early disease diagnosis plays a crucial role in addressing this challenge (Dartois & Rubin, 2022; Ullah et al., 2024). Therefore, timely accurate identification of LC is critical to confirm diagnosis, treatment, and monitor pulmonary TB disease progression. However, due to the complexity and diversity of the pulmonary TB lesions, it is not easy to annotate and recognize these lesions at the pixel level in CT images. To address these challenges, innovative approaches and tools are needed to enhance diagnostic accuracy and efficiency. CAD (Computer-aided diagnosis) systems are a promising solution.

In recent years, with the development of deep learning, CAD tools have played an increasingly important role in improving the efficiency of disease diagnosis. Deep learning models, a key component of CAD tools, have gained significant popularity because of their remarkable ability to recognize complex lesion tissues (Afify et al., 2023; Alebiosu et al., 2023; Nijati et al., 2023; Wang et al., 2023). Particularly in medical image analysis, multi-task deep learning models (Graham et al., 2023; Hu et al., 2024; Pan et al., 2024; Wu et al., 2024) can simultaneously perform multiple tasks such as lesion detection, segmentation, and classification, obtaining radiological diagnostic information about the

location, shape, and type of lesions from multiple dimensions, providing a more comprehensive basis for clinical decision-making.

However, a notable drawback of the application of deep-learning models in medical imaging is that they rely on extensive training data to achieve optimal performance (Chen et al., 2022; Hao et al., 2021; Huang et al., 2024; Zhang et al., 2024). This limitation hinders their application in domains characterized by limited training data availability, particularly in automatically and accurately identifying pulmonary TB lesion tissue in lung CT.

To reduce the cost of obtaining pixel-level annotations, some recent methods have sought to use inexpensive labels (such as classification labels) to overcome the challenge of annotating radiological images, including self-supervised learning (Ben-Shaul et al., 2024; Fini et al., 2023; Schaeffer et al., 2024; Zhang et al., 2024; Zhang et al., 2024), active learning (Bi et al., 2024; Bruna et al., 2024; Yuan et al., 2023), semi-supervised learning (Fini et al., 2023; S. Huang et al., 2023; Jiao et al., 2023; Wei & Gan, 2023), and weakly-supervised learning (Kuang et al., 2024; Laradji et al., 2021; Li et al., 2023; Lu et al., 2023; Ru et al., 2022; Sun et al., 2023; Ye et al., 2022).

Weakly supervised learning allows models to be trained using data with coarser-grained annotations. Unlike fully supervised learning which requires a large amount of finely annotated data, weakly supervised learning only needs a small amount of supervision signals, such as image-level labels (Graham et al., 2023; Kuang et al., 2024; Zhou et al., 2023), partial annotations, etc., to enable efficient model learning. This greatly reduces the cost of data

annotation and expands the scale of datasets that can be used for training. Moreover, WSSS (weakly supervised semantic segmentation) is a specific application of weakly supervised learning in the field of computer vision, aiming to train semantic segmentation models using image-level labels or other weak supervision signals to obtain pixel-level predictions.

In the field of WSSS (Kuang et al., 2024; Laradji et al., 2021; Li et al., 2023; Lu et al., 2023; Ru et al., 2022; Sun et al., 2023; Ye et al., 2022), CAM (Class Activation Mapping) (Rao et al., 2023; Wang et al., 2024; Yin et al., 2023) are the most common solution, and it is also the most challenging one among all WSSS scenarios. Specifically, CAM, as a visual attention technique, can be used to explain the decision-making process of deep learning models in image classification tasks. Moreover, CAM can help understand the image regions that the model focuses on when classifying a specific category. 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) (de Andrade et al., 2023), and thus solve the problem of the high cost of pixel-level annotation in radiological images to some extent.

Overall, while deep learning methods, especially those based on the MTL paradigm, have played a pivotal role in efficiently obtaining radiological diagnostic information for pulmonary TB, deep learning models still face the

challenge of requiring extensive annotated data. WSSS, by leveraging relatively accessible image-level annotations to train semantic segmentation models, provides a promising solution to alleviate the costly pixel-level annotation burden, warranting further exploration and research.

1.2 Research Questions

This study is devoted to classifying lung TB CT images. However, during the implementation of the research, we encountered some challenges and obstacles. This subsection will explain the specific issues that need to be addressed in the study.

1.2.1 Challenges in Pulmonary Tuberculosis CT Dataset Annotation

The use of MTL (Gordaliza et al., 2019; Graham et al., 2023; Hu et al., 2024; Wu et al., 2024) and WSSS (Bi et al., 2024; Hatamizadeh et al., 2022; Zhou et al., 2023) techniques in the analysis of TB CT imaging is a promising area of research that holds potential for improving the accuracy and efficiency of pulmonary TB diagnosis. However, existing TB CT datasets do not adequately support these approaches (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020), as they often lack the detailed and multi-faceted annotations required for MTL and WSSS tasks. This limitation constrains the development and application of CAD models for pulmonary TB (Bi et al., 2024; Hatamizadeh et al., 2022; Zhou et al., 2023).

For effective implementation of MTL and WSSS, a CT dataset for pulmonary TB must satisfy certain annotation criteria. Specifically, it should include

precise semantic segmentation of lesion areas, along with classification information such as the number, size, and specific lung location of the lesions. Such comprehensive annotations are essential for training robust CAD models that can perform nuanced analyses of TB lesions (Gordaliza et al., 2019; Graham et al., 2023; Hu et al., 2024; Wu et al., 2024). Moreover, these annotations must be created by experienced and skilled medical professionals to ensure data accuracy and consistency. The expertise required for this level of annotation highlights the significant resources involved, making high-quality, expertly annotated datasets necessary yet highly challenging to acquire. These requirements underscore the need for further research and development in dataset preparation for TB analysis and the critical role of reliable, detailed annotations in advancing CAD capabilities for pulmonary TB. Overall, such efforts ensure the precision of data and enhance the robustness of machine learning models in detecting TB.

1.2.2 Limitations of Current Approaches in Pulmonary Tuberculosis Lesion Identification in CT Images

Pulmonary TB lesions present significant challenges for accurate identification in CT imaging due to their complex and variable characteristics. Unlike other pulmonary diseases, TB lesions often display considerable variation in size, shape, and appearance, complicating the task of precise delineation. For example, ground-glass opacities—a common feature of TB—typically have indistinct borders caused by surrounding inflammation and exudates. This variability and subtlety in lesion presentation demands sophisticated identification techniques that current standard medical 3D models struggle to

fulfill. Existing 3D models are typically trained on datasets containing specific image features and common patterns associated with more typical diseases, which limits their ability to accurately capture the unique characteristics of TB lesions (Afify et al., 2023; Alebiosu et al., 2023; Nijati et al., 2023; Wang et al., 2023). One significant issue with these models is their susceptibility to high false positive rates (Chen et al., 2022; Hao et al., 2021; Huang et al., 2024; Zhang et al., 2024). Standard models, lacking TB-specific adaptations, may incorrectly identify other lung abnormalities—such as scars from past infections or non-TB granulomas—as TB lesions. This misidentification leads to a higher rate of false positives, which not only requires additional manual review but may also prompt unnecessary clinical interventions. Additionally, these models tend to perform poorly in capturing TB’s diverse manifestations, resulting in a high rate of missed detections (Graham et al., 2023; Hu et al., 2024; Pan et al., 2024; Wu et al., 2024). Such limitations compromise the reliability of TB diagnosis, potentially impacting treatment outcomes. Addressing these gaps in current models is essential to developing more accurate, reliable diagnostic tools for pulmonary TB in clinical settings, ensuring that the unique and varied presentations of TB lesions are effectively accounted for in automated analysis.

1.2.3 Limitations of Conventional Single-Task Learning Models in Analyzing Pulmonary Tuberculosis Lesion Information

In CAD of pulmonary TB, capturing a comprehensive set of radiographic features is essential for distinguishing TB from other pulmonary conditions and enhancing diagnostic accuracy. Such detailed feature extraction supports

more timely and precise treatment plans. However, conventional Single-Task Learning (STL) approaches (Chen et al., 2022; Hao et al., 2021; Huang et al., 2024; Zhang et al., 2024), which train independent models for each specific task, are limited in efficiency and comprehensiveness—particularly under resource constraints. STL models are traditionally designed to perform single tasks, such as image classification or target detection, but this isolated approach can miss valuable connections between different diagnostic features. In pulmonary TB analysis, STL-based models typically fail to capture the intrinsic relationships among features, potentially limiting the depth and diversity of the extracted information. For instance, a model focused solely on classifying the number of TB lesions may overlook key aspects such as lesion size, anatomical location, and interaction with surrounding structures (Chen et al., 2022; Hao et al., 2021; Huang et al., 2024; Zhang et al., 2024). Additionally, STL models treat each feature in isolation, leading to the inability to capture interdependencies between features. This segmentation can result in the loss of crucial information about correlations between lesion characteristics, such as the tendency of certain lesion types to exhibit specific size or growth patterns. However, these interdependencies are often essential for accurate disease characterization and prediction of disease trajectory, and their absence can reduce the model's effectiveness in clinical applications. Multi-task learning (MTL) (Graham et al., 2023; Kuang et al., 2024; M. Zhou et al., 2023) has been explored as an alternative to address these challenges, especially in CT scan analysis for other conditions. MTL enables a single model to perform multiple tasks simultaneously, capturing interrelated features within a unified framework. By leveraging MTL, it is possible to build models

that provide a richer, more interconnected representation of pulmonary TB lesions, improving diagnostic insights and operational efficiency.

1.2.4 Time-Consuming Annotation of Pixel-Level Pulmonary Tuberculosis Lung Cavity

Pixel-level semantic segmentation annotation of LC (Lung Cavity) plays a crucial role in the development of CAD for pulmonary TB CT imaging. Because through these annotated data, the CAD methods can automatically learn robust feature representations related to LC, thereby efficiently identifying LC lesion tissues, and providing valuable support for clinical diagnosis. However, in the task of identifying pulmonary TB tissue, due to the complex shape and outline of pulmonary TB lesions, semantic segmentation models based on the full supervision learning paradigm typically require a large amount of labor-intensive pixel-level annotated data (de Andrade et al., 2023; Miao et al., 2023; Wu et al., 2023). Moreover, the acquisition of LC pixel-level annotation usually requires medical experts to classify each pixel in the image, which is an extremely time-consuming task. Especially in pulmonary TB CT imaging, a complete CT scan may contain hundreds of slices, and medical experts need to accurately mark the LC lesion areas in all CT slices, including edges and small lesions. This workload is huge.

1.3 Research Motivation

TB is a disease that seriously threatens global public health. Classification of pulmonary TB in lung CT scan is a crucial step. Moreover, the manifestations of pulmonary TB lesions are diverse, and manual annotation is expensive, which poses a huge challenge to lesion identification. Given the limitations of

existing medical analysis methods, this study focuses on utilizing MTL and weakly supervised learning methods to improve the efficiency of pulmonary TB disease diagnosis. Specifically, first, this study aims to address the lack of pulmonary TB datasets by constructing a dedicated dataset for identifying pulmonary TB lesion tissues. Secondly, this study aims to obtain comprehensive radiological diagnostic information on pulmonary TB lesions by using MTL technology. In addition, this study aims to reduce the need for large-scale manual labeling data by using advanced techniques such as weakly supervised learning and CAM and improve the efficiency and accuracy of identifying pulmonary TB lesions in CT images. Finally, this research is significant for improving the capacity for early diagnosis and treatment of pulmonary TB, and ultimately for controlling and reducing the impact of this global public health problem.

1.4 Research Questions

The research questions for this study are as follows:

Question 1: How can we develop a comprehensive CT dataset specifically focused on pulmonary TB lesions in medical image processing for early pulmonary TB diagnosis and treatment efficacy?

Question 2: Given the complex characteristics of pulmonary TB lesions in CT imaging, such as varying sizes, shapes, appearances, and often unclear boundaries due to inflammation and exudates, what model architectures can be designed to accurately recognize and segment these lesions, overcoming the limitations of standard medical 3D models?

Question 3: Several limitations exist in acquiring PTB features using conventional STL models. How can these shortcomings be effectively addressed to enhance the efficiency of PTB diagnosis?

Question 4: What efficient methods or algorithms can be developed to accurately generate pixel-level LC information, thereby reducing costs and saving time?

Question 5: What strategies can be employed to enhance the efficiency and precision of CAM techniques for WSSS tasks, allowing them to capture more granular semantic details within the LC region while reducing the need for extensive manual annotations?

1.5 Research Objectives

The primary objectives of this research are to develop deep learning methods that can efficiently and accurately classify TB lesion tissues in lung CT scan images. To achieve these research aims, this thesis sets the following specific objectives:

Objective 1: Develop and curate specialized datasets that provide comprehensive information about TB lesions, including detailed annotations for accurate lesion analysis. These datasets will support advanced research and clinical applications by enhancing the quality and depth of TB lesion data.

Objective 2: Design and implement deep learning models specifically tailored for the segmentation and classification of TB lesions in lung CT images. These

models will address the challenges of accurately identifying lesion contours and characteristics within the complex structure of lung tissues.

Objective 3: Establish a MTL framework for TB radiological feature analysis in CT images. This framework aims to provide a holistic diagnostic tool that offers detailed, clinically relevant information by simultaneously handling multiple related tasks (e.g., segmentation and classification).

Objective 4: Develop a WSSS method for fine-grained segmentation of LC in TB lesions. This approach will improve pixel-level segmentation accuracy using minimal annotations, targeting the efficient identification of complex LC structures.

Clarifying the Distinctions Between Objective 2 and Objective 3.

Objective 2 and Objective 3, while related to TB lesion analysis in lung CT images, have distinct focuses and aims: i) Objective 2 is specifically centered on the development of deep learning models capable of segmenting and classifying TB lesions. The focus here is on designing models that can accurately identify and delineate individual TB lesions in lung CT images. The goal is to achieve precise segmentation and classification of lesions, contributing to a deeper understanding of TB lesion characteristics, such as shape and size. ii) Objective 3, on the other hand, is focused on building a comprehensive multi-task learning (MTL) framework for the broader diagnostic analysis of TB CT radiological features. This objective goes beyond simple segmentation and classification by aiming to provide a more integrated diagnostic tool that can handle multiple tasks simultaneously within a single

framework. The MTL framework not only performs segmentation and classification but also facilitates connections and information sharing across tasks, creating a more holistic and detailed clinical profile of TB. In summary, Objective 2 is dedicated to developing specific models for the accurate segmentation and classification of TB lesions, whereas Objective 3 aims to construct a multi-task framework that synthesizes information from various tasks to enhance the overall diagnostic capability and provide more comprehensive clinical insights.

1.6 Research Scope

The scope of this study is outlined as follows: i) This study investigates the application of MTL and WSSS techniques for identifying TB lesions in lung CT images. ii) This research is dedicated to developing new TB CT datasets. iii) Patients with endobronchial pulmonary TB and tuberculous pleurisy combined with lesions in the lung parenchyma are included in the study. iv) This research focuses on the segmentation of LC and pulmonary structures in CT scans of the lungs. v) A rigorous evaluation and validation of the proposed method will be conducted based on comprehensive quantitative performance metrics.

1.7 Research Contributions

This work makes the following original contributions, aligned with the objectives outlined above:

Contribution 1: Creation of two curated datasets for pulmonary TB analysis in CT images:

- DeepPulmoTB: A dataset focused on MTL applications, containing pixel-level segmentations and detailed classification labels for LC attributes (size and count) across 354 CT scans.
- DeepLungCavi-WSSS: A dataset for WSSS of lung cavities, providing classification and segmentation information specifically for LC lesions across 328 CT scans.

Contribution 2: Development of two novel deep learning models for TB diagnosis:

- DPTBNet: A multi-scale feature fusion model that captures and integrates features across different scales, enhancing the detection and classification of TB lesions in an MTL context.
- SwinUNeLCsT: A hybrid Transformer-CNN architecture optimized for detecting lung cavities. This model integrates local and global image features to enhance LC detection, achieving efficiency and accuracy with reduced computational complexity compared to standard self-attention modules.

Contribution 3: Design of a comprehensive MTL-based framework for TB radiological feature analysis in CT images. This framework enables simultaneous segmentation and classification of TB lesions, generating detailed diagnostic outputs that capture multiple disease aspects. Through

feature sharing across tasks, this framework strengthens the robustness of feature representations, facilitating more accurate and reliable clinical insights.

Contribution 4: Introduction of an innovative WSSS framework for segmenting lung cavities using only classification-level annotations. This framework achieves high performance in LC segmentation, surpassing existing weakly supervised methods in medical imaging, and demonstrates the feasibility of accurate segmentation with minimal annotation resources.

The outcomes of this research have been published in two different Science Citation Index Expanded (SCIE) journals, as detailed below:

First, a paper titled "SwinUNeLCsT: Global-local spatial representation learning with hybrid CNN-transformer for efficient tuberculosis lung cavity weakly supervised semantic segmentation" has been accepted by the Journal of King Saud University-Computer and Information Sciences, which has an impact factor of 6.9. This work proposes an innovative framework that integrates convolutional neural networks and transformers for efficient weakly supervised semantic segmentation of tuberculosis lung cavities.

Secondly, another paper titled "DeepPulmoTB: A benchmark dataset for multi-task learning of tuberculosis lesions in lung computerized tomography (CT)" has been published in Heliyon, a journal with an impact factor of 4.0. This work introduces a benchmark dataset for multi-task learning to identify and analyze tuberculosis lesions in lung CT images.

The publication of these two papers demonstrates the innovative contributions of this research in the field of tuberculosis image analysis. Furthermore, we believe that these contributions hold significant importance in driving the advancement of related technologies and their clinical applications.

1.8 Thesis Organization

This work is divided into six chapters. Chapter 1 serves as the foundational section of the entire thesis, encompassing key elements such as research background, problem statement, research objectives and purposes, research questions, research contributions, and significance. These contents lay the groundwork for the subsequent chapters.

Chapter 2 provides a thorough analysis and review of relevant literature. Firstly, it reviews the current research status of pulmonary TB CT datasets, discussing the research achievements and challenges in this field as documented in existing literature. Secondly, it introduces the development tools and software used in the experiments of this study, laying the groundwork for subsequent experiments. Additionally, it provides a brief overview of deep learning-based methods for medical CT image analysis, emphasizing the significance and broad prospects of deep learning in this field. Finally, it explores the current applications of MTL and WSSS techniques in the field of medical imaging, providing theoretical foundations for the method selection and design of this study.

Chapter 3 discusses the quantitative research methodology, including the overall framework of research design and environment setup, as well as the fundamental steps to achieve the research objectives. The chapter begins by elucidating the approach to identifying research questions and introducing the pulmonary TB analysis dataset constructed in this study. Subsequently, it describes the development process of models and algorithms, including the design of MTL and WSSS frameworks for detecting pulmonary TB lesions. Additionally, it provides detailed explanations of the configuration of experimental environments, model training, and evaluation steps. Finally, the concluding section summarizes the entire quantitative research methodology.

Chapter 4 provides a detailed description of the pulmonary TB CT radiological comprehensive feature analysis framework proposed in this thesis, based on MTL. Firstly, it elucidates the process of constructing the pulmonary TB lesion MTL dataset. Next, it elaborates on the proposed DPTBNet learning network, including key components such as multi-scale feature extraction, improved iterative algorithms, and network training. Subsequently, specific details of the experiments are enumerated.

Finally, a thorough discussion of the results is conducted, focusing on the segmentation effectiveness of pulmonary TB lesions and pulmonary regions, as well as the classification performance of pulmonary attributes. In addition, necessary figures and tables are presented to corroborate the superior performance of this MTL model from an analytical perspective. This chapter integrates the construction, implementation, and evaluation of the model,

providing a multidimensional and comprehensive interpretation of this innovative approach.

Chapter 5 introduces a LC WSSS Framework. Firstly, it elaborates on the process of constructing the WSSS dataset for pulmonary TB LC, including data sources and annotation methods. Next, it provides a detailed description of the proposed SwinUNeLCsT model, which consists of key components such as an encoder, multi-head attention module, window-based local multi-head attention module, and decoder. Then, it introduces the WSSS framework based on this model, involving core steps such as pseudo-label generation and feature fusion gates, and lists specific details of the experiments. Finally, the chapter evaluates the model's performance in generating LC regions and ablative studies. From multiple dimensions including data construction, model architecture, framework process, and experimental validation, this chapter systematically demonstrates the feasibility and superiority of this innovative WSSS method.

Chapter 6 summarizes the main conclusions of this work and proposes suggestions for future research directions. The conclusion section succinctly summarizes the application achievements of WSSS and MTL in computer aided pulmonary TB diagnosis discussed in the paper. In the future work suggestions section, objectively evaluates the shortcomings of current methods and proposes forward-looking suggestions for areas requiring further research. By summarizing the contributions of this thesis and indicating future

directions, this chapter provides valuable insights for the continued innovation and progress in this field.



CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides a detailed review of the relevant literature and background research. It reviews the current research status of pulmonary TB CT datasets in Section 2.2, discussing the research achievements and challenges in this field as documented in the existing literature. Section 2.3 briefly introduces the development tools and software used in the experiments of this study. This is followed by Section 2.4 on deep learning-based methods for medical CT image analysis. Sections 2.5 and 2.6 outline the current applications and major challenges of cutting-edge techniques such as WSSS and MTL in medical CT image analysis, respectively. Finally, Section 2.7 summarizes the overall content of the chapter and reviews the main points covered in the literature review.

2.2 Overview of Pulmonary Tuberculosis CT Dataset

Lung CT imaging is an important tool for detailed visualization of lung abnormalities associated with pulmonary TB. Therefore, constructing high-quality pulmonary TB CT data sets is of great significance to promote the development of automated pulmonary TB diagnostic tools.

2.2.1 Pulmonary Tuberculosis Lesions Recognition Dataset

TB lesion recognition datasets are typically composed of CT images with annotations marking the specific regions of pulmonary TB lesions. These lesions exhibit various morphological appearances in the images, including LC, LC walls, and lung consolidations.

Table 2.1: Existing Pulmonary TB Lesions Recognition Datasets

Datasets	C-LCW	LC	Lung Areas	MCS
Ionescu et al. (2021)	-	-	√(Seg)	-
Wen et al. (2021)	√(Seg)	-	-	-
Li et al. (2021)	√(Det)	√(Det)	-	-
Ionescu et al. (2022)	-	√(Det)	√(Seg)	-
Z. Xiao et al. (2023)	-	√(Seg)	-	-

Note: The terms (Seg) and (Det) refer to segmentation and detection tasks, respectively. MCS: Multi-Category Segmentation. C-LCW: Consolidation or Lung Cavity Wall.

Table 2.1 summarizes the available annotation categories in existing CT pulmonary TB lesion recognition datasets. From the table, we observe that the segmentation categories in the existing datasets are relatively limited, for example, only including the LC class (Ionescu et al., 2022; Li et al., 2021). Furthermore, in these datasets, some research works primarily use bounding boxes for object detection to annotate pulmonary TB lesions. However, bounding box annotations can only provide a rough estimation of the approximate location of pulmonary TB lesion tissues.

2.2.2 Pulmonary Tuberculosis Lesions Descriptive Analysis Dataset

Descriptive datasets of pulmonary TB lesions typically provide attribute descriptions related to the lesions, including their size, shape, and distribution locations in the lungs. These detailed lesion descriptions are crucial for gaining an in-depth understanding of the various manifestations and evolutionary trends of pulmonary TB lesions. Moreover, such datasets often contain metadata such as patient's demographic information, characteristics of pulmonary TB strains, and treatment outcomes. These data allow researchers to more comprehensively explore various factors influencing the characteristics of pulmonary TB lesions.

One of the most used datasets for the analysis of CT pulmonary TB lesions is the "TB Portals" (Wollenberg et al., 2022) provided by the NIH. This dataset is intended exclusively for the study of drug-resistant TB. The "TB Portals" dataset (Wollenberg et al., 2022) contains descriptive analysis information on the lesion tissues of pulmonary TB patients that includes factors such as LC size, total number of cavities, transmission status, and more. In addition, there is the ImageCLEF pulmonary TB series dataset (Ionescu et al., 2021; Ionescu et al., 2022; Kozlovski et al., 2020). This series of datasets provides image analysis and diagnostic information about pulmonary TB and its lesions. This information includes an assessment of the severity of pulmonary TB, classification of pulmonary TB lesion types, etc.

2.3 Medical Image Analysis Projects and Software

The development of medical imaging technology has greatly promoted the progress of modern medical and health fields, making the early detection, diagnosis, and treatment of diseases more efficient and accurate. With the continuous advancement of AI and computer vision technologies, medical image analysis projects and software have become indispensable tools for researchers and medical professionals. These tools not only enhance the automation of medical image processing but also augment the analysis and interpretation capabilities of images. The following will introduce several projects and software tools widely used in the field of medical image analysis, each of which has unique functions and application scenarios.

2.3.1 Medical Open Network for AI

MONAI (Medical Open Network for AI) (Cardoso et al., 2022) is an open-source project maintained by Project MONAI, focused on improving the reproducibility, standardization, and performance of medical imaging AI models. The project supports a wide range of medical imaging tasks, such as image segmentation, classification, and enhancement. In this study, the deep learning environment was built using the MONAI framework. MONAI's core strength lies in providing a set of deep learning methods and tools, enabling researchers and medical professionals to develop, train, and deploy medical imaging AI models more easily. Additionally, MONAI offers rich data preprocessing and enhancement tools to assist users in processing medical imaging data, as well as efficient data loading and processing mechanisms to meet various medical image analysis needs.

2.3.2 ITK-SNAP

ITK-SNAP (Zhou et al., 2021) is an open-source software specifically designed for medical image segmentation. It is developed based on the ITK of the NIH and provides users with a powerful and user-friendly interface to process 3D medical imaging data. This study employed ITK-SNAP to annotate pulmonary TB CT images and display cross-sectional views of the 3D data. The main strength of ITK-SNAP lies in its precise segmentation tools, enabling users to perform complex medical image segmentation tasks either manually or semi-automatically. It supports various image formats, including CT, MRI, and ultrasound, making it an ideal tool for medical image analysis, research, and education. Additionally, ITK-SNAP offers advanced analysis capabilities, such as volume measurements and morphological analysis, which aid in clinical decision-making and research endeavors.

2.3.3 ParaView

ParaView (Shen et al., 2023) is an open-source, multi-platform data analysis and visualization application widely employed for data analysis and visualization needs in scientific research. Although ParaView was not originally designed specifically for medical image analysis, its powerful data processing and visualization capabilities make it an effective tool for handling complex medical data. ParaView supports various data formats and visualization techniques, such as volume rendering, iso-surface extraction, and vector field visualization, allowing users to gain an in-depth understanding of the data from multiple perspectives and levels. Its parallel processing capability is particularly well-suited for handling large-scale medical image

datasets, providing robust support for efficient medical image analysis and research. This study chose ParaView as the primary tool for three-dimensional visualization and processing of pulmonary TB CT data.

2.4 Deep Learning Networks for Medical CT Imaging

In previous studies, the processing of CT medical images mainly relied solely on traditional image processing techniques, including threshold selection (Wang et al., 2024; Weszka, 1978), region growing (Adams & Bischof, 1994; Tremeau & Borel, 1997; Wan & Higgins, 2003), edge detection (Elharrouss et al., 2023; Su et al., 2021) and other methods. Although these techniques can achieve satisfactory results under certain conditions, their robustness and accuracy are still limited due to the diversity in shape and size of lesion tissues in medical CT images, as well as the interference from image noise and artifacts (Wang et al., 2021).

In contrast, in recent years, deep learning technology has brought revolutionary breakthroughs in the field of medical CT image processing due to its excellent feature extraction and analysis capabilities (Chen et al., 2024; Gao et al., 2023; Yuan et al., 2023; Zhou et al., 2023). Especially CNN, networks based on Transformer architecture, and networks based on CNN and Transformer hybrid architecture. These networks have shown strong performance in processing complex medical images.

2.4.1 Convolutional Neural Network (CNN)

CNN was originally proposed by Yann LeCun and others in the 1990s to address handwritten digit recognition tasks (LeCun et al., 2015). Its design was inspired by the hierarchical structure and local receptive field characteristics of the visual cortex in biological neural systems, enabling it to automatically learn feature representations from input data. In recent years, CNN have demonstrated outstanding performance in computer vision (Deb et al., 2024), pattern recognition (Piao & Jin, 2023), and other fields, with particularly broad application prospects in medical CT image analysis (Akter et al., 2024; Bi et al., 2024; Han et al., 2024; Mikami et al., 2024; Yuan et al., 2023; Zhao et al., 2023).

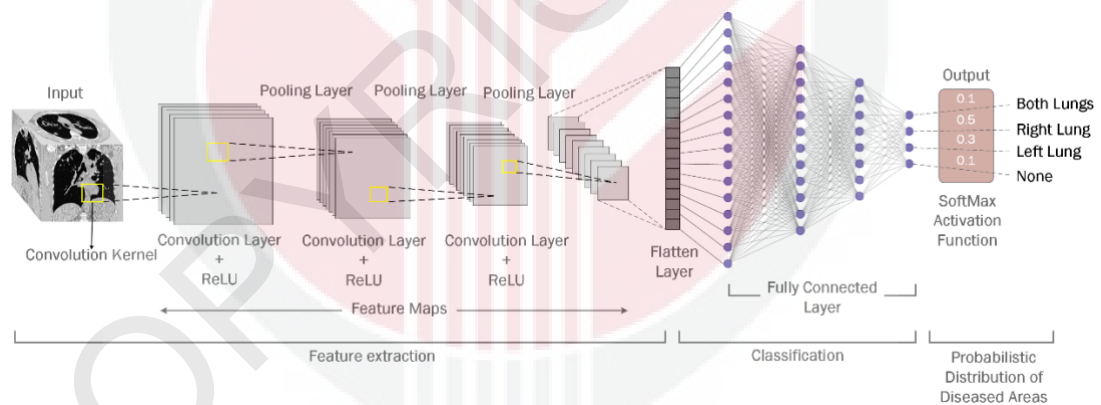


Figure 2.1: Structure of Convolutional Neural Network

Typical medical CNN model (as shown in Figure 2.1) is built by alternately stacking multiple convolutional layers and pooling layers. The convolutional layers are responsible for capturing local features in the input data by learning local filters, while the pooling layers reduce the computational cost and increase the abstraction level of the features by lowering their resolution

(Hosain et al., 2021). In the deeper parts of the network, FC layers encode the high-level features learned from the convolutional and pooling layers to perform classification or regression tasks (Babiloni et al., 2023; Reus-Muns et al., 2023). The core advantage of CNNs lies in their excellent feature learning capability and efficient modeling of image data. This makes it a key technology in fields such as medical image analysis.

2.4.1.1 Convolution Layer

The convolutional layer is the fundamental building block of CNN. It operates by sliding a filter (also known as a convolutional kernel) over the input data (such as an image) and performing convolution operations on the input data to capture local features (Basha et al., 2020; Zhou et al., 2023). Furthermore, convolutional layers also can automatically extract effective feature representations from the input data, thus reducing the need for manual feature engineering. In the field of medical CT image analysis, different types of convolutions serve different purposes and have different effects (Han et al., 2024; Zhou et al., 2023). Here are two common types of convolutions:

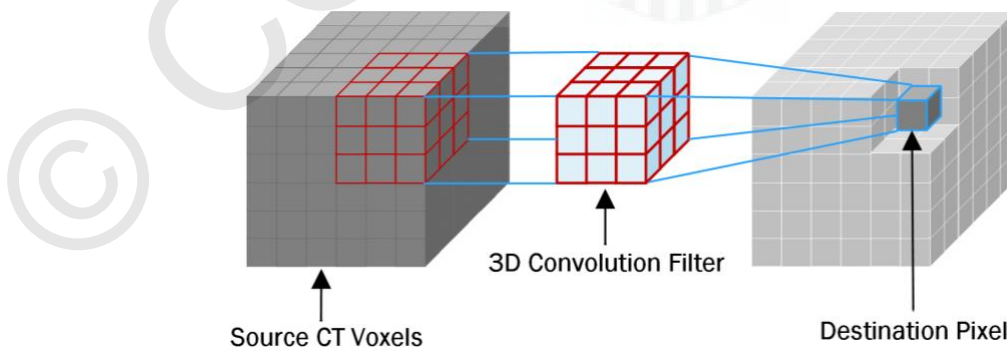


Figure 2.2: Implementation Principle of Standard 3D Convolution

The first one is standard 3D convolution operation (Khan et al., 2023; Opfer et al., 2023). This convolution operation is one of the most basic and widely used operations in the field of medical deep learning image analysis. It extracts features or performs feature mapping by applying one or more convolutional filters to the input data, as shown in Figure 2.2. This operation involves the element-wise multiplication and summation between the filter and the input data, usually with an added bias term. The implementation principle of standard 3D convolution (Khan et al., 2023; Li et al., 2022; Opfer et al., 2023) is as follows:

Let F be the input feature volume with dimensions $H \times W \times D \times C$, where H , W , and D are the height, width, and depth respectively, and C is the number of channels. Let K be a 3D convolutional kernel with dimensions $K_H \times K_W \times K_D \times C$, where K_H , K_W , and K_D are the height, width, and depth of the kernel respectively. B is the bias term for the convolution operation, typically a scalar. Let G be the output feature volume with dimensions $H' \times W' \times D' \times C'$, where H' , W' , and D' are the height, width, and depth of the output feature volume, and C' is the number of output channels. For each element $G_{h',w',d',c'}$ in the output feature volume, its value is calculated by the following formula (Khan et al., 2023; Li et al., 2022; Opfer et al., 2023):

$$G_{h',w',d',c'} = \sum_{i=0}^{K_H-1} \sum_{j=0}^{K_W-1} \sum_{k=0}^{K_D-1} \sum_{c=0}^{C-1} F_{h+i,w+j,d+k,c} \cdot K_{i,j,k,c} + B_{c'} \quad (2.1)$$

In formula (2.1), h' , w' , and d' are the height, width, and depth indices in the output feature volume, respectively, and c' is the index of the output channel. $h+i$, $w+j$, and $d+k$ correspond to the region in the input feature volume that aligns with the three-dimensional convolutional kernel K . $B_{c'}$ is the bias term corresponding to the c' -th output channel (Khan et al., 2023).

The second one is 3D DConv (Depthwise Separable Convolution) (Hou et al., 2021; Xing et al., 2024; Ye et al., 2019). This convolution operation is a variant derived from 3D standard convolution. Overall, it decomposes a 3D standard convolution into two smaller convolution operations: depthwise convolution and pointwise convolution, thereby reducing the computational cost and the number of model parameters, as shown in Figure 2.3.

In 3D Depthwise convolution, each input channel is convolved with a three-dimensional convolutional kernel, which moves across the three spatial dimensions. This step independently extracts three-dimensional spatial features for each input channel. Suppose the dimensions of the input feature volume are $H \times W \times D \times C$, where H , W , and D represent height, width, and depth, respectively, and C is the number of channels. The size of each convolutional kernel is $K_H \times K_W \times K_D \times 1$. Therefore, the output of the 3D depthwise convolution will maintain the same number of channels, with dimensions $H \times W \times D \times C$. The computation formula for the 3D depthwise convolution is as follows (Hou et al., 2021; Xing et al., 2024; Ye et al., 2019):

$$G_{h,w,d,c} = \sum_{i,j,k} F_{h+i,w+j,d+k,c} \cdot K_{i,j,k,c} \quad (2.2)$$

In formula (2.2), F represents the input feature volume, K is the convolutional kernel, G is the output feature volume, and c refers to the channel. The indices i , j , and k correspond to the height, width, and depth dimensions of the convolutional kernel, respectively (Xing et al., 2024).

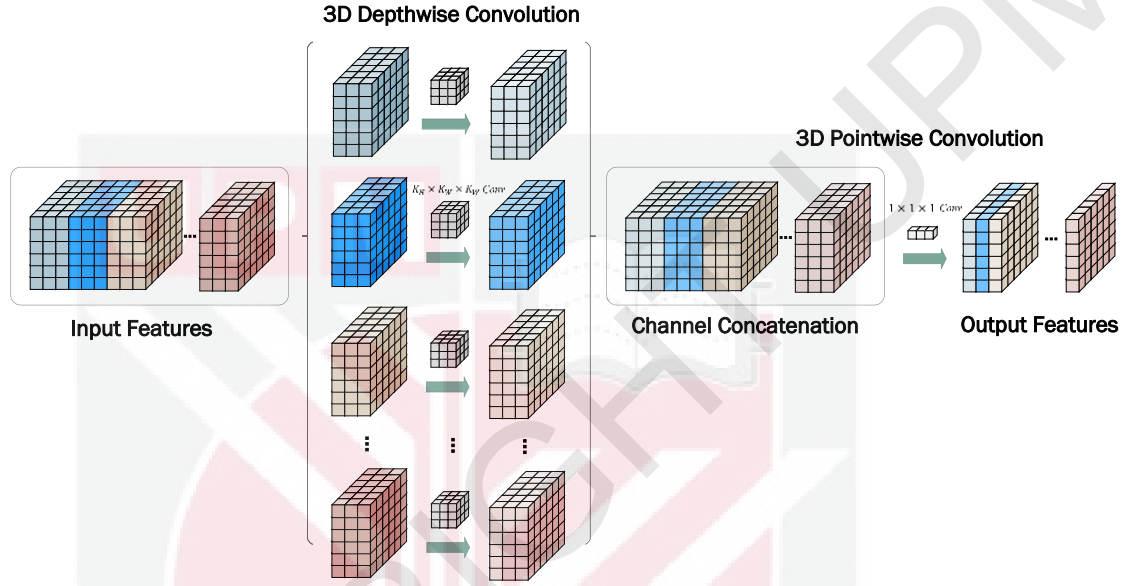


Figure 2.3: Implementation Principle of 3D Depthwise Convolution

After the 3D depthwise convolution, the 3D pointwise convolution is performed using $1 \times 1 \times 1$ convolutional kernels on the output of the 3D depthwise convolution. The purpose is to mix information from different channels without changing the spatial or depth dimensions. This is equivalent to performing a FC operation along the channel dimension. If the output of the 3D depthwise convolution is $H \times W \times D \times C$, and the intended number of output channels is C' , then the size of the convolutional kernels for the 3D pointwise convolution will be $1 \times 1 \times 1 \times C$. The computation formula for the 3D pointwise convolution (Hua et al., 2018) is as follows:

$$H_{h,w,d,c'} = \sum_c G_{h,w,d,c} \cdot K_{c,c'}' \quad (2.3)$$

In formula (2.3), H is the final output feature volume, K' is the convolutional kernel for pointwise convolution, and c' is the index of the output channel.

2.4.1.2 3D Pooling Layer

3D pooling layer (Hicham et al., 2022; Hosain et al., 2021; Li et al., 2022) is a structure commonly used in 3D deep learning model, especially widely used in CNN. The purpose of the pooling layer is to reduce the number of parameters and the amount of calculation by reducing the spatial size of the data, while also suppressing overfitting and improving the spatial invariance of the models, as shown in Figure 2.4. The specific implementation principles of 3D GMP (Dou et al., 2016) and GAP (Çiçek et al., 2016) are introduced below:

GMP (Dou et al., 2016): This pooling operation captures the salient features in the feature map by retaining the maximum value in the region while remaining invariant to small changes in neighboring pixels. Specifically, it divides the input image (or feature map) into multiple distinct regions and takes the maximum value in each region as the output for that region. Suppose an input feature map is represented by $f(x, y, z)$ at position (x, y, z) . For GMP, the output $g(i, j, k)$ is computed as:

$$g(i, j, k) = \max_{(x,y,z) \in R(i,j,k)} f(x, y, z) \quad (2.4)$$

In formula (2.4), $R(i, j, k)$ represents the area in the input feature map corresponding to the position (i, j, k) in the output feature map.

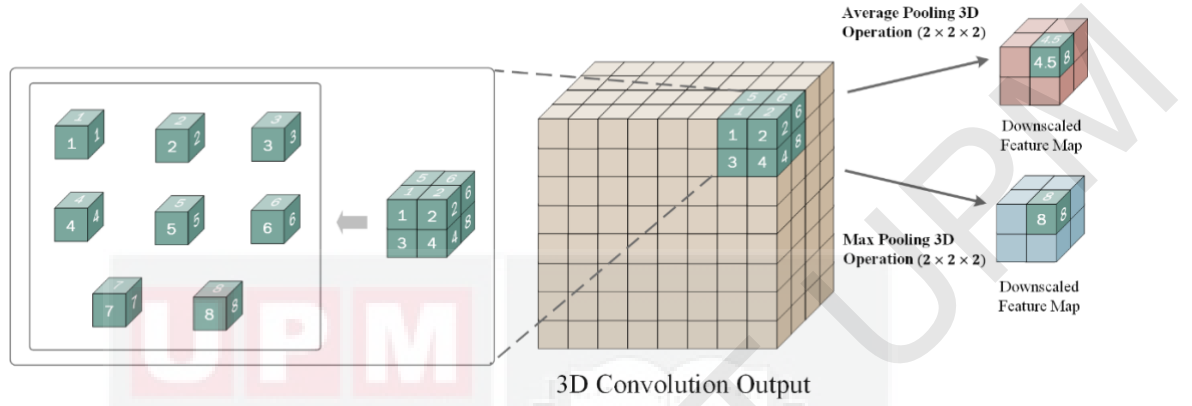


Figure 2.4: Implementation Principle of Global Max Pooling and Global Average Pooling

GAP (Çiçek et al., 2016) is another common pooling operation. It smooths the feature maps by calculating the average value of all pixel values within each region, thereby reducing the model's sensitivity to small shifts in the input. Specifically, this pooling operation first divides the input image (or feature map) into multiple independent regions, then calculates the arithmetic mean of all pixel values within each region and takes it as the output value for that region. Suppose the value of the input feature map at position (x, y, z) is $f(x, y, z)$, then the corresponding average pooling output $g(i, j, k)$ is calculated as:

$$g(i, j, k) = \frac{1}{|R(i, j, k)|} \sum_{(x, y, z) \in R(i, j, k)} f(x, y, z) \quad (2.5)$$

In formula (2.5), $|R(i, j, k)|$ represents the number of elements in the region $R(i, j, k)$.

2.4.1.3 Fully Connected layer

The FC layer (Basha et al., 2020; Reus-Muns et al., 2023) is typically located at the end of a CNN. Its main function is to integrate all the features learned from the previous convolutional and pooling layers to accomplish the final classification or regression task. In the FC layer, each input node is directly connected to all nodes in the output layer, forming a FC network structure. This structure allows the FC layer to fully integrate the multi-dimensional feature representations from the convolutional layers and perform complex pattern recognition and decision-making based on the task objective. Specifically, given an input vector x , the output y of the FC layer (Basha et al., 2020) can be calculated using the following formula:

$$y = Wx + b \quad (2.6)$$

In formula (2.6), x is the input vector; W is the weight matrix; b is the bias vector; y is the output vector. Each element y_i in the vector y can be represented as:

$$y_i = \sum_j W_{ij}x_j + b_i \quad (2.7)$$

In formula (2.7), W_{ij} represents the weight from input node j to output node i , and b_i is the bias of the output node i (Yamashita et al., 2024).

2.4.1.4 Activation Function

Activation Function plays a crucial role in CNN and other types of neural networks. They are typically introduced in the convolutional layers and FC layers, introducing non-linearity to the entire network, and enabling it to learn and model complex function mappings. Without activation functions, no matter how many hidden layers a neural network has, it essentially remains a simple variant of a linear model. Here are the working principles of some common activation functions:

1. ReLU (Gong et al., 2023): The output of the ReLU function is the positive part of the input. If the input is positive, it directly outputs that value; if the input is negative, it outputs 0. This simple non-linearity helps accelerate the training process of neural networks, while also mitigating the vanishing gradient problem (although it may cause the exploding gradient problem). ReLU expression (Gong et al., 2023) is as follows:

$$f(x) = \max(0, x) \quad (2.8)$$

The ReLU function is one of the most widely used activation functions in modern neural networks, including CNN, due to its effective non-saturating form and sparse activation properties.

2. GELU (Kang et al., 2022): GELU activation function is a function form that considers the distribution of the input data. It adjusts the final output by multiplying the input value by a weight value scaled according to the input's distribution. This mechanism simulates the "on/off" behavior of neurons to some extent, while also reserving a certain probability for small and negative signals to pass through, giving the model stronger complexity and expressiveness. Although the mathematical expression of the GELU activation function (Gao et al., 2023) is relatively complex, it can be accurately approximated by the following formula:

$$f(x) = 0.5x \left(1 + \tanh \left[\sqrt{2/\pi} \left(x + 0.044715x^3 \right) \right] \right) \quad (2.9)$$

Compared with other simple activation function forms such as ReLU, the GELU function is more adaptive to the input, and has shown superior performance in fields such as NLP and computer vision.

3. Softmax (Zhu et al., 2023): The output of the Softmax function (Zhu et al., 2023) is a set of values between 0 and 1 that sum to 1, which can be interpreted as a probability distribution. It allows the network's output to be interpreted as a probability distribution, which is particularly useful for classification tasks. The Softmax function is typically used in the output layer of neural networks and is very effective for multi-class problems. It transforms a vector or set of scores into a probability distribution. Softmax mathematical formula is as follows:

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (2.10)$$

In formula (2.10), x is the input vector, i is the index of the current element, and j is the index that sums over all elements of x . The Softmax function ensures that the sum of all outputs is 1, allowing them to be interpreted as probabilities for the different output classes.

2.4.1.5 Batch Normalization

BN (Batch Normalization) (Guo et al., 2023) is a commonly used technique in CNN to accelerate the training process and enhance the performance and stability of models. In CNN, BN is typically applied after convolutional layers and FC layers to standardize the input data for each batch. The computation of BN involves calculating the mean and variance for each feature channel, then normalizing the data, and finally learning appropriate scaling and shifting parameters for each feature channel through scaling and shifting operations. These parameters are also updated by the gradient descent algorithm during the training process. Given a mini-batch dataset $X = \{x_1, x_2, \dots, x_m\}$, where x_i is the i -th sample in the dataset, and each sample is a d -dimensional vector, the specific implementation principle of BN (Yanmeng Wang et al., 2023) is as follows:

$$\text{BN}(x_i) = \frac{x_i - \mu}{\sqrt{\sigma^2 + \delta}} \cdot \gamma + \beta \quad (2.11)$$

$$\mu = \frac{1}{m} \sum_{i=1}^m x_i \quad (2.12)$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu)^2 \quad (2.13)$$

In formula (2.11), γ and β are learnable parameters used for scaling and shifting the normalized data. δ is a very small positive number used to prevent division by zero. μ is the mean of the mini-batch dataset, and its formula is shown in (2.12). σ^2 is the variance of the mini-batch dataset, and its formula is shown in (2.13).

2.4.2 Transformer-Based Network

In recent years, Transformer-based models have achieved tremendous success in the field of NLP and have rapidly expanded to computer vision (Liu et al., 2023). In the implementation principle of Transformer networks, they first divide the image into a series of small patches (Xiao et al., 2023). At the same time, these image patches are flattened and inputted into a linear layer to be transformed into a series of embedding vectors, like converting words into embedding vectors in NLP. Subsequently, these vectors are inputted into the encoder of the Transformer. In the encoder, each encoder layer uses a self-attention mechanism to capture the relationships between image patches (Han et al., 2023).

2.4.2.1 Self-Attention Mechanism

In the self-attention mechanism (Huang et al., 2023; Li et al., 2023; Wang et al., 2023), the model is able to calculate the attention scores between any two image patches, thereby directly capturing the dependence between them

without being restricted by their positions in the image (Xu et al., 2022), as shown in Figure 2.5. In principle, this mechanism is mainly realized by comparing the embedding representations of the image patches (by mapping them to query, key, and value vectors respectively). Through this mechanism, the model can better understand the global context information of the image. Specifically, the principle of implementing the self-attention mechanism (Huang et al., 2023; Li et al., 2023; Wang et al., 2023) includes the following steps:

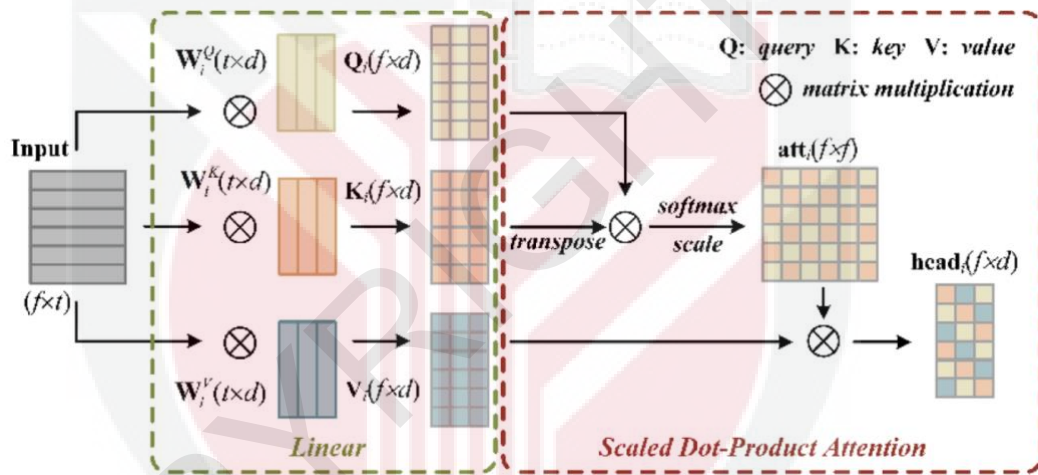


Figure 2.5: The Principle of Self-Attention Mechanism
(Source: Xu et al., 2022)

Step 1: Input representation. For a given input sequence $X = [x_1, x_2, \dots, x_n]$, where x_i represents the embedding representation of the i^{th} element in the sequence (Wang et al., 2023).

Step 2: Query, Key, Value representation. The self-attention mechanism uses three learned weight matrices W^Q , W^K , and W^V to transform each input

element x_i into query q_i , key k_i , and value v_i representations: $q_i = W^Q x_i$, $k_i = W^K x_i$, $v_i = W^V x_i$ (Xu et al., 2022).

Step 3: Attention score calculation. The attention score for each element towards other elements is obtained by calculating the similarity between the query and all keys. Dot product attention is commonly used, where the attention score a_{ij} indicates the attention level of element i towards element j :

$$a_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}} \quad (2.14)$$

In formula (2.14), d_k is the dimension of the key vectors, and this scaling factor helps with the stability of training (Wang et al., 2023).

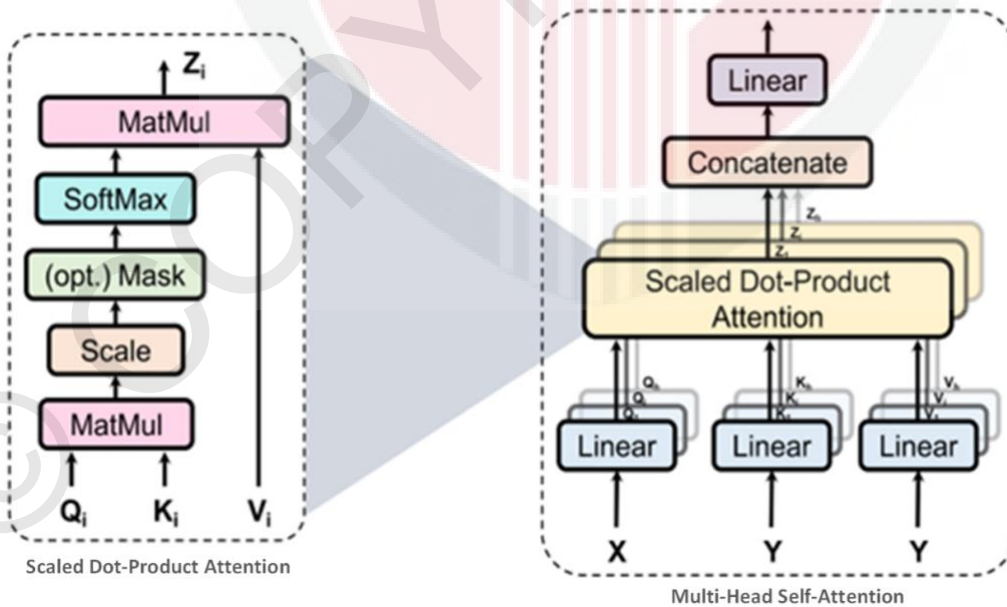


Figure 2.6: Structure of the Standard Transformer Block
(Source: Liu et al., 2023)

Step 4: Attention weights normalization. The SoftMax function is used to normalize the attention scores for all elements to ensure their sum equals 1, resulting in the weights w_{ij} (Huang et al., 2023; Zhu et al., 2023):

$$w_{ij} = \frac{\exp(a_{ij})}{\sum_{l=1}^n \exp(a_{il})} \quad (2.15)$$

Step 5: Weighted sum and output generation. For each position i , the final output y_i is obtained by accumulating the product of the values v_j of all elements in the input sequence and their corresponding normalized attention weights w_{ij} (Huang et al., 2024). Here, w_{ij} represents the attention scores normalized by the SoftMax function, reflecting the importance or relevance of the j^{th} element in the sequence relative to the i^{th} element:

$$y_i = \sum_{j=1}^n w_{ij} v_j \quad (2.16)$$

2.4.2.2 Multi-head Self-attention

A basic Transformer (Liu et al., 2023) block typically contains multiple self-attentions. These self-attentions enhance the model's ability to capture information by executing multiple self-attention operations in parallel (called "heads"). This process is called MSA (multi-head self-attention) (Bi et al., 2024; Liu et al., 2022), as shown in Figure 2.6.

In the MSA mechanism, each head learns different representations at different positions in the sequence, and then these representations are concatenated

or averaged to obtain more comprehensive information (Zhang et al., 2023). This design allows the model to capture different features of the sequence in different representational subspaces, thereby increasing the model's expressive power and complexity. Furthermore, the parallel processing mechanism of MSA can also improve the learning efficiency and performance of the model, allowing the Transformer to better handle complex sequence tasks. The implementation steps of MSA (Ru et al., 2022; Tang et al., 2022; Zhang et al., 2023) are as follows:

Step 1: Multi-head design. The model employs h sets of distinct weight matrices W_h^Q , W_h^K , W_h^V , generating different queries, keys, and values representations for each element in the input sequence. These h sets of representations form h "heads" (Liu et al., 2023).

Step 2: Independent attention calculation. For each head, the attention scores, weights, and outputs are calculated independently in the manner of self-attention:

$$y_i^h = \sum_{j=1}^n w_{ij}^h v_j^h \quad (2.17)$$

In formula (2.17), y_i^h represents the output of the h -th head, and w_{ij}^h and v_j^h respectively are the weights and values of the h -th group (Huang et al., 2024).

Step 3: Concatenation and projection of outputs: The outputs from all heads $y_i^1, y_i^2, \dots, y_i^h$ are concatenated together, and then linearly transformed using an additional learned weight matrix W^O to produce the final output y_i (Liu et al., 2023):

$$y_i = \text{Concat}(y_i^1, y_i^2, \dots, y_i^h) W^O \quad (2.18)$$

Through this approach, the MSA mechanism allows the model to learn information in parallel across multiple subspaces, enhancing the model's capability to capture complex relationships within the sequence.

2.4.2.3 Layer Normalization

Layer Normalization (LN) (Xiong et al., 2020) is a normalization technique used in deep learning networks, commonly employed in each layer of Transformer-based networks. Unlike BN, Layer Normalization normalizes each feature dimension for each sample, rather than normalizing across the entire batch of data. Given a mini-batch of data with m samples, for each sample i , its d -dimensional input $x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)})$ at a specific layer, where $x_i^{(j)}$ is the value of the j -th feature. The implementation principle of LN (Huang et al., 2023) is as follows:

$$\text{LayerNorm}(x_i) = \gamma \frac{x_i - \mu}{\sigma} + \beta \quad (2.19)$$

$$\mu = \frac{1}{d} \sum_{j=1}^d x_i^{(j)} \quad (2.20)$$

$$\sigma = \sqrt{\frac{1}{d} \sum_{j=1}^d (x_i^{(j)} - \mu)^2 + \delta} \quad (2.21)$$

In formula (2.19), γ and β are learnable parameters, used to scale and shift the normalized data. δ is a small positive number, used to prevent division by zero. μ is the mean of the sample across the specific dimension, and its formula is shown in (2.20). σ is the standard deviation of the sample across the specific dimension, and its formula is shown in (2.21) (Xiong et al., 2020).

2.4.3 Application in Medical CT Imaging Analysis

To provide a clearer overview and comparison, Table 2.2 summarizes key methods within these categories, highlighting their core techniques, advantages, and drawbacks. This table aims to facilitate a comprehensive understanding of the current landscape in medical CT imaging analysis by comparing various approaches and identifying their suitability for different applications. Specifically:

Table 2.2: Comparison of Methods in Medical CT Imaging Analysis

Method	Primary Techniques	Pros	Cons
Zhao et al. (2023)	Adaptive 3D CNN, high-resolution fusion attention	Effective in reducing false positives by using multi-level contextual information;	High computational cost due to adaptive 3D kernels
	DM-CNN with multi-scale feature fusion and dynamic uncertainty quantification	Robust feature extraction with multi-scale pooling; reduced overfitting with multi-objective loss	Complexity increases due to multiple dynamic modules, impacting training time
Huang et al. (2024)	CNN with seagull and particle swarm optimization	Enhanced cancer area identification through noise reduction and contrast enhancement	Relies heavily on pre-processing techniques, which may not generalize well across different imaging modalities
	nnFormer with volumetric self-attention and jump attention	Effective in modeling long-range dependencies	High computational burden due to extensive self-attention operations

2.4.3.1 Applications of CNN

In the field of medical image analysis, CNN can be applied to the analysis of various types of medical images, including but not limited to MRI and CT scan images (Akter et al., 2024; Zhao et al., 2023). Through training, CNN can automatically learn and analyze complex image patterns, such as tumors, nodules, and other pathological changes, thus better assisting in disease diagnosis and monitoring.

To enable early diagnosis of lung cancer, Zhao et al. (2023) developed an adaptive 3D CNN for automatic lung nodule detection. This method contains two stages: candidate nodule detection and false positive reduction. First, in the initial stage, the method extracts the spherical, spiral, and fine-grained information of nodules through a high-resolution fusion attention module. In the second stage, the method adopts an adaptive 3D CNN structure, which extracts multi-level contextual information via adaptive 3D convolutional kernels to further reduce false positives. Yu et al. (2023) developed a deep learning model called CA-BreastNet for classifying breast cancer histopathological images. To enhance its ability to analyze image textures, the model incorporated a coordinated attention mechanism into the DenseNet network architecture. Furthermore, to improve the overall accuracy, CA-BreastNet employed a convolutional decision tree method based on a specialized classification strategy to mitigate the limitation imposed by the dataset structure on model performance. To address the issues of insufficient feature representation and inaccurate attention weights in CNN for medical image diagnosis, Han et al. (2024) proposed a new CNN model for medical

imaging analysis called DM-CNN. The DM-CNN model consists of four core components: First, the dynamic multi-scale feature fusion module can select different convolutional kernels according to feature maps at different levels, extracting multi-scale feature information. Second, the hierarchical dynamic uncertainty quantification attention mechanism contains a tuning block that dynamically adjusts the attention weights based on the information of each layer. Furthermore, the multi-scale fusion pooling module can accelerate computation and prevent overfitting while preserving the main information. Finally, the multi-objective loss function enables the DM-CNN model to optimize quickly and achieve good classification performance. Huang et al. (2024) proposed a CNN based on the seagull optimization algorithm and the particle swarm optimization algorithm to detect oral cancer. Moreover, to improve the quality of input data, this method also uses some advanced image processing techniques (such as noise reduction, contrast enhancement, and data enhancement). After optimization of these methods, the CNN architecture can more accurately identify cancer areas in oral images.

2.4.3.2 Applications of Transformer-Based Networks

Nowadays, the powerful image feature capturing capability of Transformer architectures has enabled them to demonstrate outstanding performance in computer vision tasks such as image classification and segmentation (Bi et al., 2024; Huang et al., 2024; Yuan et al., 2023). Particularly when dealing with large-scale medical image datasets, Transformers' ability to capture long-range dependencies and global attention allows them to obtain global features more efficiently from images.

Zhou et al. (2023) proposed a Transformer-based 3D medical image segmentation model, dubbed nnFormer. Different from existing methods that only use Transformer as an auxiliary module to encode global context information, nnFormer comprehensively uses interleaved convolution and self-attention operations and introduces local and global volumetric self-attention mechanisms to learn volumetric representations. In addition, nnFormer also innovatively uses jump attention to replace the traditional concatenation/summation operations in the U-Net (Akter et al., 2024) architecture. Overall, the nnFormer model aims to give full play to Transformer's long-range dependency modeling capabilities and improve the performance of 3D medical image segmentation.

To address the high computational burden and fixed window partitioning issues faced by traditional Transformer models in medical image segmentation applications, Lin et al. (2023) proposed a novel boundary-aware lightweight Transformer, named BATFormer. The model effectively leverages multi-scale feature maps for low-complexity global information modeling by integrating a cross-scale global transformer module. Moreover, Lin et al. (2023) also developed a boundary-aware local Transformer module and adopted an entropy-based adaptive window partitioning technique, which optimized computational efficiency while preserving important shape features. Huang et al. (2024) proposed a self-supervised Transformer model called SSiT (Saliency-guided Self-Supervised Image Transformer) for grading diabetic retinopathy lesions. To leverage domain-specific prior knowledge to guide self-supervised pretraining, SSiT innovatively incorporates saliency

maps into the self-supervised learning process. Specifically, SSiT adopts two saliency-guided learning tasks: 1) A saliency-guided contrastive learning method based on momentum contrast. This method removes non-essential regions from the input sequence using saliency maps of retinal images, thereby guiding the query encoder to capture salient features. 2) Training the query encoder to predict saliency segmentation to preserve fine-grained information in the learned representations.

2.4.3.3 Applications of CNN-Transformer Hybrid Networks

The CNN-Transformer hybrid architecture network combines the local feature extraction capability of CNNs and the global dependency-capturing capability of Transformers, providing a powerful tool for processing medical images (Bi et al., 2024; Manzari et al., 2023). This hybrid architecture allows the model to achieve a deeper understanding of images while maintaining high efficiency. In multiple domains of medical image analysis, such as human organ segmentation and abnormal region detection, CNN-Transformer hybrid networks have demonstrated superior performance compared to models based on a single architecture (pure convolution or transformer).

Yuan et al. (2023) proposed a novel Complementary Transformer-CNN Network (CTC-Net). This model combines two encoders built from Swin Transformers and residual CNNs to generate complementary features and effectively integrates these features through cross-domain fusion blocks. In addition, the model enhances feature expressiveness by computing the correlation between features from different domains and employing a dual-

attention mechanism and uses a Swin Transformer decoder with skip connections to extract richer spatial and semantic information.

To address the reliability issue of CNNs being susceptible to adversarial attacks in medical diagnostic systems, Manzari et al. (2023) proposed a highly robust and efficient CNN-Transformer hybrid model. This hybrid model combines the local characteristics of CNNs and the global connectivity of visual Transformers. Subsequently, to mitigate the high quadratic computational complexity of the self-attention mechanism while co-attending to information in different representational subspaces, Manzari et al. (2023) also constructed an attention mechanism based on efficient convolutional operations. Finally, to alleviate the vulnerability of Transformer models to adversarial attacks, Manzari et al. (2023) attempted to learn smoother decision boundaries by augmenting shape information through batch-wise feature mean and variance substitution in the high-level feature space.

To address the limitations of existing fully convolutional networks in automatically segmenting musculoskeletal images with special features such as low contrast, uneven textures, and blurred boundaries, Bi et al. (2024) proposed a semi-automatic segmentation model that combines CNN and Transformer networks. This model leverages the long-range dependency capability of Transformer networks to complement the shortcomings of FCNs in understanding global context, overcoming the limitation of FCNs focusing only on local patterns while ignoring global context, thereby achieving effective segmentation of challenging musculoskeletal images.

2.4.4 Deep Learning Networks for Pulmonary Tuberculosis CT Imaging Analysis

In recent years, deep learning networks have emerged as a powerful tool for the analysis of pulmonary TB using CT imaging (Alebiosu et al., 2023; Nijati et al., 2023), showcasing considerable potential in diagnostics, treatment planning, and disease monitoring. These networks, particularly CNN (Akter et al., 2024; Wen et al., 2021) and hybrid models, are capable of automatically extracting key features such as lung cavities, lesion distribution, and lymph node enlargement from CT images. By training on extensive annotated datasets, these models can learn complex patterns that distinguish between healthy and diseased lungs, contributing to more accurate and sensitive TB diagnosis. In addition to improving diagnostic sensitivity and reducing clinical workload, deep learning models can assist in grading and evaluating the severity of TB lesions, thus supporting precision treatment decisions and enabling continuous monitoring for disease progression.

Given the broad scope of applications, various methods have been developed and tailored specifically for pulmonary TB CT imaging analysis. To provide a clearer comparison of these approaches, Table 2.3 summarizes key methods, highlighting their core techniques, advantages, and limitations. This table serves as a quick reference, helping readers understand the strengths and drawbacks of each method and facilitating the selection of suitable models based on specific requirements in pulmonary TB diagnosis and analysis (Li et al., 2021).

Alebiosu et al. (2023) developed a new medical image segmentation model called DAvoU-Net, which can utilize multi-scale residual blocks and dense atrous connections to generate features for pulmonary TB analysis. To reflect the severity of pulmonary TB more accurately in the images, the model employs an innovative approach: feeding the features extracted by a 3D CNN into a bidirectional long short-term memory network, enabling the network to extract higher-level discriminative features. This novel approach holds promise for providing robust support for accurate diagnosis and assessment of pulmonary TB.

Nijati et al. (2023) investigated methods for early prediction of pulmonary TB treatment outcomes, prevention of adverse consequences of drug resistant pulmonary TB, and interruption of its transmission. By analyzing 612 CT scans (including pre-treatment and follow-up scans) from 204 pulmonary TB patients who completed treatment, the research team developed three models based on radiomic features (Bagging, Random Forest, Gradient Boosting) and two deep learning models (supervised and weakly supervised). By fusing the best prediction scores from these models, Nijati et al. (2023) created a new ensemble model that improved the accuracy and robustness of predictions.

Table 2.3: Summary of Deep Learning Methods for Pulmonary Tuberculosis CT Imaging Analysis

Method	Primary Techniques	Pros	Cons
Alebiosu et al. (2023)	DAVoU-Net with multi-scale residual blocks, dense atrous connections,	Captures multi-scale features and higher-level discriminative features;	High computational cost
Nijjati et al. (2023)	Ensemble of radiomics-based (Bagging, Random Forest, Gradient Boosting)	Improved predictive accuracy and robustness by fusing radiomics and deep learning;	Complexity in ensemble model design
Zhang et al. (2023)	Integration of radiomic features, deep learning	Comprehensive model utilizing multi-source information for differentiating TB nodules from lung cancer	Requires extensive data for accurate feature extraction
Yoon et al. (2023)	nnU-Net for 3D cavity volume detection and quantification	High sensitivity in detecting cavity volumes; strong association with clinical markers	Limited generalizability outside of cavity volume analysis

Zhang et al. (2023) proposed a deep learning model that integrates deep learning features, radiomic features, and clinical information to differentiate between tuberculous nodules and lung cancer. Specifically, Zhang et al. (2023) extracted 100 radiomic features from positron emission tomography and CT images using the pyradiomics platform and obtained 2,048 deep learning features through a residual neural network. In their study, Zhang et al. (2023) developed four models: 1) a machine learning model based on traditional radiomic features; 2) a deep CNN model inputting image features alone; 3) a radiomics-DCNN model inputting both radiomic and deep learning features; 4) an integrated model inputting radiomic features, deep learning features, and clinical information.

Yoon et al. (2023) explored the clinical application of CT cavity volume in diagnosing pulmonary TB and NTM-PD. Specifically, first, Yoon et al. (2023) developed a novel 3D medical analysis model, dubbed nnU-Net. This model can be applied to automatically detect and quantify cavity volumes from CT images. Second, their study included data analysis of 392 patients. The analysis showed that the established model demonstrated high sensitivity and a relatively low false positive rate in internal validation. This proved the association between CT cavity volume and sputum smear positivity for pulmonary TB as well as treatment needs for NTM-PD. This result simultaneously demonstrated that the nnU-Net model aids in more accurately assessing pulmonary TB infectiousness and formulating treatment strategies for NTM-PD.

2.5 Multi-Task Learning

MTL (Multi-Task Learning) is a machine learning paradigm. Its core concept is to process multiple related tasks in parallel and capture potential task correlations by learning shared feature representations, thereby enhancing the generalization of the model ability (Hu et al., 2024; Yajing Liu et al., 2023; Song et al., 2020; Zhou et al., 2021). Different from traditional STL, MTL can capture more general and high-quality feature representations between tasks in a shared underlying network by utilizing the associated knowledge between tasks, thereby improving the overall learning effect (Deng et al., 2020).

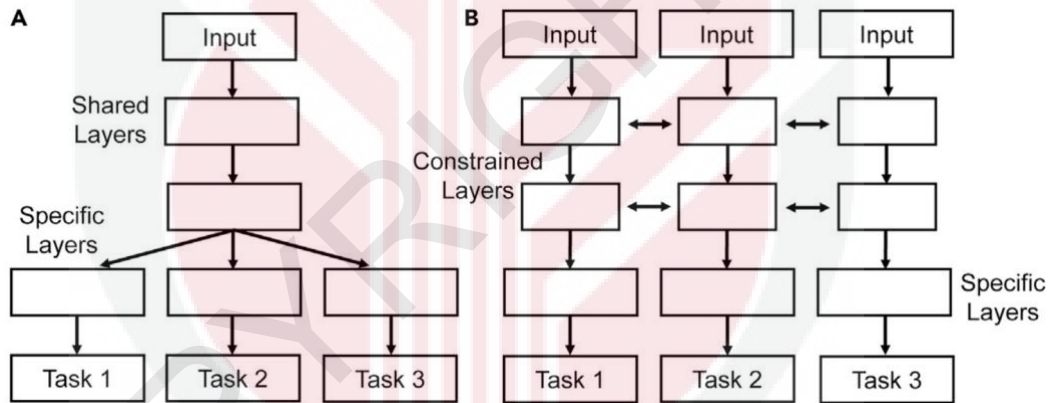


Figure 2.7: Depiction of Hard and Soft Parameter Sharing Mechanisms
(Source: Deng et al., 2020)

According to the model's communication of information between tasks and the degree of this communication during the MTL process, MTL can be divided into hard parameter sharing and soft parameter sharing (Deng et al., 2020), as shown in Figure 2.7. Among them, hard parameter sharing (HPS) (Zhang et al., 2022; Zhang et al., 2023) is the most widely adopted MTL strategy, which mainly learns multiple tasks simultaneously by sharing parameters in the hidden layers. In the specific implementation of HPS method, typically,

different tasks share the underlying structure and parameters of the model, such as convolutional layers or encoder layers, while using independent parameters only in the top output layer specific to each task (Yajing Liu et al., 2023). This setting allows the bottom network to learn a common feature representation, while the top layer captures the unique patterns of each task, thereby realizing knowledge transfer and sharing (Hu et al., 2024). The implementation steps of hard parameter sharing MTL are as follows:

Step 1: Basic form. Given K related tasks, each task k with a dataset $\left\{ \left(x_i^{(k)}, y_i^{(k)} \right) \right\}_{i=1}^{N_k}$, where $x_i^{(k)}$ represents the input features, $y_i^{(k)}$ represents the corresponding target output, and N_k is the number of samples for task k . In the MTL framework, the model attempts to learn a shared representation $h(x)$ and a set of task-specific functions $\{f_k\}_{k=1}^K$, where each f_k maps the shared representation to the output space of task k . The overall model output for task k can be represented as $f_k(h(x))$ (Wu et al., 2024).

Step 2: Loss function. The goal of model training is to minimize the total loss across all tasks, which is typically represented as the weighted sum of each task's loss (Graham et al., 2023):

$$L = \sum_{k=1}^K \alpha_k L_k \quad (2.22)$$

In formula (2.22), L_k is the loss function for task k , and α_k is the weight assigned to task k , indicating the relative importance of different tasks. For

regression tasks, the loss function L_k might be MSE, and for classification tasks, L_k might be cross-entropy loss (Bai et al., 2022).

Step 3: Parameter sharing and task-specific parameters. Suppose the model's parameters can be divided into two sets: shared parameters θ_{shared} and a set of task-specific parameters $\{\theta_k\}_{k=1}^K$, where K is the total number of tasks (Liu et al., 2023). For a given input x , the shared feature representation $h(x; \theta_{\text{shared}})$ is determined by the shared parameters, and the task-specific output y_k is calculated by the function $f_k(h(x; \theta_{\text{shared}}); \theta_k)$, where f_k is a task-specific function related to task k (Yan et al., 2024). Thus, the overall model can be represented as:

$$y_k = f_k(h(x; \theta_{\text{shared}}); \theta_k), \forall k \in \{1, 2, \dots, K\} \quad (2.23)$$

Step 4: Training optimization. Consider the total loss L as the weighted sum of all task loss functions:

$$L(\theta_{\text{shared}}, \{\theta_k\}_{k=1}^K) = \sum_{k=1}^K \alpha_k L_k(\theta_{\text{shared}}, \theta_k) \quad (2.24)$$

In formula (2.24), L_k is the loss function for task k , and α_k is the weight assigned to task k . The goal of model training is to find parameters θ_{shared} and $\{\theta_k\}_{k=1}^K$ that minimize the total loss L (Zhang et al., 2022; Zhang et al., 2023).

Using the gradient descent method, the parameter updates in each iteration t can be represented as:

$$\theta_{\text{shared}}^{(t+1)} = \theta_{\text{shared}}^{(t)} - \eta \frac{\partial L}{\partial \theta_{\text{shared}}} \quad (2.25)$$

$$\theta_k^{(t+1)} = \theta_k^{(t)} - \eta \frac{\partial L}{\partial \theta_k}, \forall k \in \{1, 2, \dots, K\} \quad (2.26)$$

In formula (2.25) and (2.26), η is the learning rate, $\frac{\partial L}{\partial \theta_{\text{shared}}}$ and $\frac{\partial L}{\partial \theta_k}$ are the gradients of the total loss L with respect to the shared and task-specific parameters, respectively.

2.5.1 Gradient Conflict Problem

In the MTL framework, the gradient conflict problem is a common challenge (Liu et al., 2023; Zhou et al., 2022), this problem can be defined as: Given K different tasks, where each task k has a specific loss function $L_k(\theta)$. Moreover, these tasks share a set of parameters θ_{shared} , aiming to find a parameter set during training that can effectively minimize the sum of the loss functions of all tasks. When using the gradient descent method for training, the update rule for the shared parameters (Gao et al., 2022; Haji & Abdulazeez, 2021) can be expressed as:

$$\theta_{\text{shared}}^{(t+1)} = \theta_{\text{shared}}^{(t)} - \eta \sum_{k=1}^K \alpha_k \nabla L_k(\theta_{\text{shared}}) \quad (2.27)$$

Algorithm 2.1: Gradient surgery Update Rule

Require: Model parameters θ , task minibatch $B = \{T_k\}$

- 1: $\mathbf{grad}_k \leftarrow \nabla_{\theta} L_k(\theta) \forall k$
 - 2: $\mathbf{grad}_k^{\text{PC}} \leftarrow \mathbf{grad}_k \forall k$
 - 3: **for** $T_i \in B$ **do**
 - 4: **for** $T_j \stackrel{\text{uniformly}}{\sim} B$, T_i in random order **do**
 - 5: **if** $\mathbf{grad}_i^{\text{PC}} \cdot \mathbf{grad}_j < 0$ **then**
 // Subtract the projection of $\mathbf{grad}_i^{\text{PC}}$ onto $\mathbf{grad}_j$
 - 6: Set $\mathbf{grad}_i^{\text{PC}} = \mathbf{grad}_i^{\text{PC}} - \frac{\mathbf{grad}_i^{\text{PC}} \cdot \mathbf{grad}_j}{\mathbf{grad}_j^2} \mathbf{grad}_j$
 - 7: **return** update $\Delta\theta = \mathbf{grad}^{\text{PC}} = \sum_i \mathbf{grad}_i^{\text{PC}}$
-

In formula (2.27), $\nabla L_k(\theta_{\text{shared}})$ represents the gradient of the loss function for the k -th task with respect to the shared parameters θ_{shared} . α_k is the weight assigned to the k -th task, used to adjust the contribution ratio of each task in the total loss, and η is the learning rate. This method allows the model to consider the losses of all tasks during the training process, by adjusting their respective contributions, to jointly optimize the shared parameters, thus achieving simultaneous learning of multiple tasks.

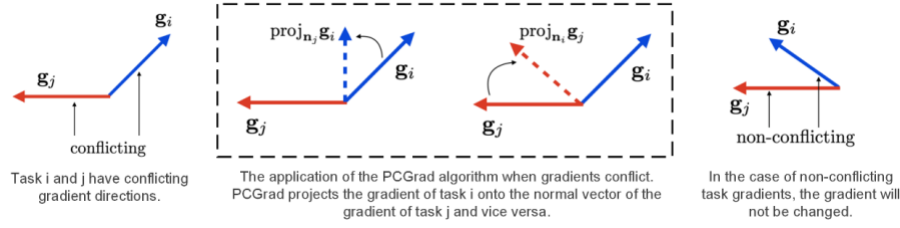


Figure 2.8: Conflicting Gradients and PCGrad

To address the gradient conflict problem (Liu et al., 2023; Zhou et al., 2022), researchers have proposed various strategies (Dou et al., 2023; Huang et al., 2023; Koloskova et al., 2023; Liu et al., 2021; Yang et al., 2023). Among them, the gradient surgery algorithm (Deng et al., 2020; Dou et al., 2023) is a widely adopted method. This method mainly mitigates gradient conflicts between tasks by modifying the gradients across tasks, as shown in Figure 2.8. Specifically, when the gradient directions of two tasks are opposite, directly summing them will cancel each other out, leading to a reduction in learning efficiency. To alleviate this issue, gradient surgery projects each task's gradient onto the orthogonal space of the gradients from other tasks, reducing the conflict between task gradients and ensuring that the gradient update directions for each task are more coordinated and consistent (Deng et al., 2020), as shown in Algorithm 2.1.

2.5.2 Multi-Task Learning Methods in Medical CT Imaging Analysis

In the field of medical image analysis, MTL has shown its unique importance (Gordaliza et al., 2019; Graham et al., 2023; Hu et al., 2024; Wu et al., 2024). This is mainly due to the inherent characteristics of medical image data: they are typically multimodal and multidimensional, containing rich information about disease indicators and anatomical structures. This makes it not only

feasible but also highly beneficial to tackle multiple related tasks simultaneously. Furthermore, by leveraging the correlations between tasks, MTL can also better exploit the shared feature representations inherent in the image data, thereby improving the performance across all tasks (Yang et al., 2023).

Therefore, in recent years, several innovative MTL methods have been developed specifically for medical imaging, particularly within CT imaging analysis. Each approach has been designed to address distinct challenges within medical diagnostics, from segmentation and classification of organs and lesions to registration and alignment. Table 2.4 below provides a comparative overview of key MTL methods, highlighting their main techniques, applications, advantages, and limitations. This summary enables a clearer understanding of each method's contributions and trade-offs, helping to identify the most suitable approaches for specific medical CT imaging applications.

For ultrasound diagnosis of thyroid nodules, Wu et al. (2024) proposed an automated MTL framework that simulates the typical diagnostic process of radiologists, designed for thyroid nodule diagnosis in ultrasound videos. This framework comprises two ordered learning tasks progressing from partial to holistic. The first task leverages prior knowledge to interpret the features of each frame image, thus obtaining the corresponding frame-level TI-RADS score and embedding representation, providing diagnostic information for radiologists while reducing computational costs. The second task models the

sequence of embedding vectors, capturing temporal context information, and selectively enhancing important information, thereby distinguishing between benign and malignant nodules, and improving the framework's efficiency and generalization ability.

To address the challenge of predicting incidental prostate cancer (iPCa) in multiparametric magnetic resonance imaging, Pan et al. (2024) proposed a Transformer-based framework called iPCa-Former. Firstly, the framework simultaneously optimizes two tasks through a mutual learning mechanism: prostate transitional zone segmentation and iPCa prediction. Secondly, the framework introduces a joint optimization function that combines focal loss and boundary-based mutual information loss, effectively addressing the issues of class imbalance in classification and small foreground region proportions in segmentation. Graham et al. (2023) proposed a MTL method based on analyzing digital pathology slide images for simultaneously performing segmentation and classification tasks of cell nuclei, glands, lumens, and different tissue regions. Moreover, the Cerberus model introduced in this method can effectively combine different tasks by leveraging data of the same tissue type and resolution from multiple independent data sources.

Furthermore, in the analysis of medical CT scan images, the gradient conflict problem in MTL is particularly prominent. This is because different tasks (such as lesion detection and type classification) need to learn different feature representations, and the gradient update directions of these features are always inconsistent. For example, regarding the problem of adjusting

hyperparameters to balance multiple loss functions (such as registration accuracy and deformation field smoothness) in deformable registration methods for medical image analysis, Dou et al. (2023) proposed an alignment model called GSMorph based on a gradient surgery mechanism. By orthogonally projecting the gradients of similarity losses, GSMorph avoids the need to introduce additional balancing hyperparameters and can be merged with any deep registration network, thereby achieving dynamic balancing between multiple losses without increasing parameters or affecting inference speed.

2.6 Weakly Supervised Semantic Segmentation

WSSS is a technique widely used in image processing and computer vision (Rong et al., 2023; Ru et al., 2022; Ru et al., 2023). This technique aims to train models using limited or incomplete annotation information to achieve categorical segmentation of each pixel in an image (Rong et al., 2023). These annotations may be bounding boxes (Wang & Xia, 2024; Xie et al., 2023), scribbles (Ying et al., 2023; Zhou et al., 2023), or merely image-level labels (Chen et al., 2022). The typical workflow of WSSS first involves training a model using weak labels, and then generating pseudo-segmentation masks through CAM (Li et al., 2023).

Table 2.4: Comparison of Multi-Task Learning Methods in Medical CT Imaging Analysis

Method	Primary Techniques	Pros	Cons
Wu et al. (2024)	Automated MTL framework with sequential learning tasks	Simulates radiologist's diagnostic process; reduces computational load by processing frame-level features first	Primarily designed for sequential data; may not be suitable for static imaging
Pan et al. (2024)	Transformer-based iPCa-Former with mutual learning mechanism and combined focal loss	Effectively addresses class imbalance and small target regions;	High computational cost due to Transformer structure;
Graham et al. (2023)	Cerberus model that integrates data from multiple sources, performing segmentation	Utilizes data from different sources to improve generalization; performs well on segmentation	Limited application to pathology slides;
Dou et al. (2023)	GS-Morph alignment model	Without additional hyperparameter tuning	Primarily focused on registration tasks;

These masks are later refined through a series of refinement strategies and are ultimately used to train the segmentation model in a standard fully supervised manner. Since this approach does not rely on precise pixel-level annotations, it is particularly useful for reducing annotation costs and handling scenarios where precise annotations are difficult to obtain. Specifically, WSSS (Rong et al., 2023; Ru et al., 2022; Ru et al., 2023) can be defined as follows: Assume there is a collection of images $X = \{X_1, X_2, \dots, X_N\}$, where each image X_i corresponds to a set of weak annotations Y_i , and Y_i can be image-level labels, bounding box annotations, or other forms of weak annotations. The goal is to learn a mapping $f: X \rightarrow S$, which maps each image X_i to its semantic segmentation map S_i , where each pixel p in S_i is assigned a category label $c \in \{1, 2, \dots, C\}$, with C being the total number of categories (Kweon et al., 2023).

Typically, this mapping f is learned by minimizing a loss function $L(f(X_i), Y_i)$, which measures the inconsistency between $f(X_i)$ (i.e., the predicted segmentation map) and the weak annotations Y_i . Since the annotations are weak, the loss function is designed to accommodate different types of weak annotation information and guide the model to effectively learn from these incomplete annotations (Xu et al., 2024).

2.6.1 Class Activation Mapping

CAM (Class Activation Mapping) (Rao et al., 2023; Wang et al., 2024; Yin et al., 2023) is a deep learning model visualization technique that can visualize the focus of attention of a deep learning network when recognizing a specific class by generating a heatmap. The principle behind CAM is as follows:

First, it operates on the output of the last convolutional layer of the network, and by applying GAP to the image dimensions of that convolutional layer, it converts it into a one-dimensional vector whose length is equal to the number of channels in that layer. This vector is then passed to a FC layer to perform the classification task (Wang et al., 2024). This working principle is illustrated in Figure 2.9.

From the figure, we observe the key steps involved in the CAM prediction process of visualizing deep learning networks for specific categories (such as "Australian dog"), including analyzing the weight parameters related to the "Australian dog" category in the last FC layer (e.g., $W_1, W_2, \dots, W_n$). Because, in this layer, these weight parameters reflect the salient features that the network focuses on when identifying images of that category. At the same time, this series of parameters will also be used to weight the output of the last convolutional layer to generate the CAM map. In addition, Figure 2.9 also reveals which areas of the image have a greater influence or contribution when the model makes the Aussie dog prediction.

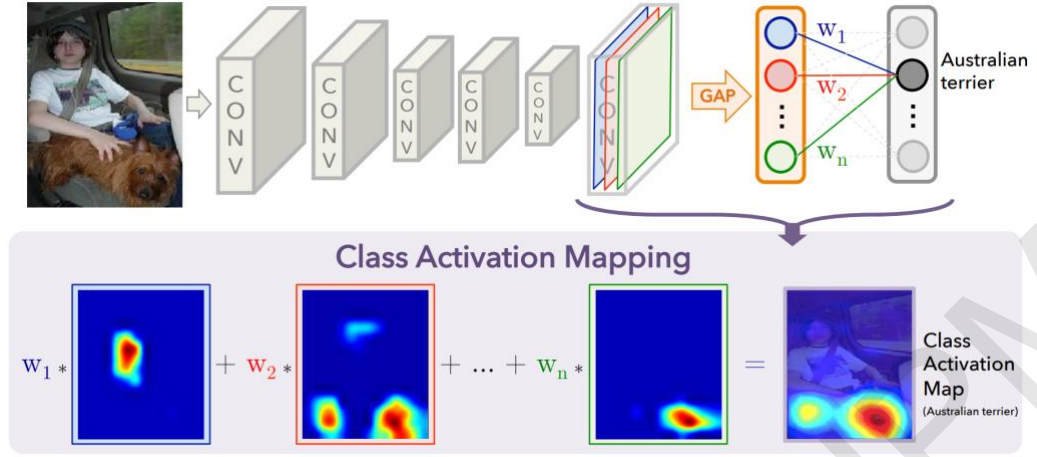


Figure 2.9: Implementation Principle of Class Activation Mapping
(Source: Zhou et al., 2016)

In the WSSS task, the CAM technique enables the model to efficiently focus on the regional features most relevant to the target category by relying solely on image-level annotations, thereby achieving the segmentation of deep semantic information. This alleviates the high cost of pixel-level annotation to a certain extent. Specifically, the implementation of CAM in the WSSS task (Li et al., 2023; Ru et al., 2022) includes the following steps:

In the first step, suppose a given deep learning network, where the last convolutional layer outputs K feature maps A^k . The weights of the FC layer are w_k^c , where c represents the category, with a total of C categories (Ru et al., 2022).

In the second step, calculate the importance weights of the feature maps. For the c^{th} class, its importance weight is:

$$M_c = \sum_k w_k^c \quad (2.28)$$

In formula (2.28), w_k^c represents the weight of the k^{th} feature map corresponding to the c^{th} class in the FC layer weights (Ying et al., 2023).

In the third step, obtain the weighted features for each position. Multiply each feature map A^k by its corresponding weight w_k^c , and sum over all feature maps:

$$L_c = \sum_k w_k^c A^k \quad (2.29)$$

In formula (2.29), L_c is a two-dimensional feature map, representing the weighted feature importance for each position for class c (Ru et al., 2022).

In the fourth step, obtain the CAM. This involves applying the ReLU operation to the weighted features L_c , i.e., keeping only the positive values and setting the negative values to zero:

$$CAM_c = \text{ReLU}(L_c) \quad (2.30)$$

In formula (2.30), CAM_c is the CAM for class c . Areas with higher values indicate greater importance for that category (R. Li et al., 2023).

Since the introduction of CAM (Zhou et al., 2016), there have been various improved and extended variant techniques emerging (Chattopadhyay et al., 2018; Selvaraju et al., 2017; Wang et al., 2020; Wang & Zhang, 2023). These

variants not only refine the original method but also extend its capabilities. Among them, Grad-CAM (Wang & Zhang, 2023) has become one of the most widely used techniques due to its ability to assign higher weights to CAM through gradient information, producing more refined visualization results.

2.6.1.1 Gradient-Weighted Class Activation Mapping

Grad-CAM (Wang & Zhang, 2023) is an improved version of CAM. Different from CAM, Grad-CAM does not require structural modification and retraining of the original classification model but adopts a more flexible method. Specifically, when making predictions on an image, Grad-CAM (Wang & Zhang, 2023) utilizes the pre-trained weights and performs backpropagation to the layer that needs to be visualized (such as the convolutional layer), obtaining a gradient matrix of the same size as the output feature maps of that layer. Then, GAP is applied across the spatial dimensions of the gradient matrix, generating a vector whose length corresponds to the number of channels in the feature layer. This vector is used to weight each channel of the feature layer, thereby generating a heatmap that visually displays the network's attention to different features. The specific implementation of Grad-CAM includes the following steps:

Step 1: Feature extraction. The input image passes through a CNN, producing a feature layer A . Assume A is the feature map from the last convolutional layer of the network (Liu et al., 2023).

Step 2: FC Layer prediction. After passing through a series of layers, the network outputs the final prediction value y through a FC layer. For a specific category c (e.g., "Tiger Cat"), its predicted value is denoted as y^c (Song et al., 2023).

Step 3: Gradient calculation. Backpropagation is performed on y^c , calculating the gradient of A with respect to y^c , denoted as $A' = \frac{\partial y^c}{\partial A}$. This gradient A' represents the importance of each element in the feature layer A for the predicted value y^c of category c (Li et al., 2023).

Step 4: Gradient weighting. For each channel in the feature map A , calculate the importance weight α_k^c of that channel by performing GAP on the gradients within A' for that channel, expressed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2.31)$$

In formula (2.31), Z is the total number of elements in A , and i and j are the spatial dimension indices of the feature map (Afify et al., 2023).

Step 5: Linear weighted combination. Using α_k^c as the weights, perform a weighted combination of all K channels of the feature map A to obtain the unactivated CAM $L_{\text{Grad-CAM}}^c$ (Wang & Zhang, 2023), expressed as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (2.32)$$

The ReLU function is applied here to only consider features that positively contribute to the category c , as negative values are considered to have a suppressive effect on the prediction outcome. Through this series of steps, Grad-CAM generates a heatmap $L_{\text{Grad-CAM}}^c$, which highlights the image regions most important for the prediction of the specific category y^c .

2.6.1.2 Weakly Supervised Semantic Segmentation for Medical CT Imaging

With the continuous increase in medical image data, efficient and accurate image segmentation methods have become increasingly important for disease diagnosis and research (Isensee et al., 2021). In traditional fully supervised medical image segmentation, models typically require a large amount of pixel-level accurate annotation data for training (Bi et al., 2024; Hatamizadeh et al., 2022; Zhou et al., 2023). However, obtaining precise pixel-level annotations for medical images is both time-consuming and expensive, especially in the medical field where professional medical knowledge is required (Wang et al., 2023). Therefore, WSMIS (weakly supervised medical image segmentation) has become an active research area (Du et al., 2023; Lei et al., 2023; Lin et al., 2023; Wang & Voiculescu, 2023). This method aims to train segmentation models using less, incomplete, or inaccurate annotation information. Among the many WSMIS methods, the method of using image-level labels to train a classification network and output a CAM is the most common solution due to

its cost-effectiveness (Li et al., 2023; Wang & Voiculescu, 2023; Ye et al., 2022). Meanwhile, this method is also among the most challenging in WSSS for medical images. CAM and its variants (e.g., Grad-CAM) offer an intuitive view of key image areas influencing model predictions, assisting medical experts in interpretation. Table 2.5 below summarizes key WSMIS methods, comparing their techniques, applications, pros, and cons, to help identify the most suitable approaches for medical CT imaging. On the one hand, CAM provides an efficient way for medical image analysis to generate pseudo-annotated data, which can then be used to train segmentation models. On the other hand, by visualizing the key regions, CAM also facilitates communication between medical researchers and machine learning models, thus better aiding in decision support and explanation during the medical diagnostic process. To address the challenges of label co-existence and location proximity under image-level labels, Kuang et al. (2024) developed a novel weakly supervised learning method for multi-class medical image segmentation. This method includes a multi-level classification network capable of generating binary predictions and class activation maps for each class, as well as a feature decomposition module based on semantic affinity.

Furthermore, the method effectively also alleviates the issues of label co-existence and region overlap through two mechanisms: the cross-guidance loss and the mutual exclusion loss. Wei et al. (2024) proposed a weakly supervised glottis segmentation method. This method is capable of end-to-end training of neural network models using only point annotations as training labels. Specifically, the method includes an automatic seed region growing

algorithm. This algorithm combines network predictions and image features to propagate sparse point annotations into high-quality pseudo-labels. Furthermore, the method introduced a new loss function based on the structural similarity index to enhance boundary segmentation performance.



Table 2.5: Comparison of Weakly Supervised Semantic Segmentation Methods in Medical CT Imaging Analysis

Method	Primary Techniques	Pros	Cons
CAM and Variants (e.g., Grad-CAM)	Classification network with CAM to generate pseudo-labels	Cost-effective; provides intuitive visualization of key regions for prediction interpretation	Limited accuracy in boundary details; may struggle with complex structures
Kuang et al. (2024)	Multi-level classification, feature decomposition with cross-guidance	Effectively addresses label co-existence and region overlap	Complex architecture; computationally intensive for training
Wei et al. (2024)	End-to-end glottis segmentation using point annotations and seed region growing	Minimizes annotation effort by using sparse point labels	Limited to small, localized structures
Li et al. (2023)	SA-MIL approach treating pixels as instances	Reduces annotation burden for pathologists	Primarily applicable to histopathology

To simplify the annotation work of pathologists and promote the development of automated analysis of pathological images, Li et al. (2023) proposed a novel weakly supervised method called SA-MIL for accurate pixel-level segmentation of histopathological images. Specifically, this method ingeniously treats pixels as instances in multiple instance learning and incorporates a self-attention mechanism, thereby enhancing global relevance among instances and improving segmentation performance. However, although WSSS has been widely used in many medical fields, it is still in its infancy in the field of pulmonary TB CT image analysis.

2.7 Summary

This chapter first reviews two common types of pulmonary TB CT datasets and points out their existing gaps. The first type is the pulmonary TB lesion identification dataset. This dataset typically consists of CT images with annotations marking specific regions of pulmonary TB lesions. The second type is the descriptive analysis dataset for pulmonary TB lesions. This dataset typically provides attribute description information related to the lesions, including their size, shape, and distribution location in the lungs. Next, this chapter introduces three commonly used medical image analysis projects and software tools: MONAI, ITK-SNAP, and ParaView. These three tools have different emphases in terms of functionality and application scenarios. When used in combination, they can effectively support various stages of medical image analysis, such as data preprocessing, model training, annotation, and result visualization.

Furthermore, this chapter discusses deep learning methods for medical CT image analysis and focuses on introducing two deep learning networks widely used in medical CT image analysis: CNN and Transformer-based network models. These two networks have widespread applications in medical CT image analysis tasks, such as lesion detection, segmentation, and classification. CNNs are typically used for extracting local features, while Transformers are capable of capturing global information. To fully leverage the advantages of both, recent research has also explored the application of hybrid CNN-Transformer network models in medical CT image analysis. Finally, this section discusses the applications of MTL and WSSS techniques in computer-aided medical disease diagnosis. On one hand, MTL methods have important applications in obtaining comprehensive medical disease diagnosis information.

However, the gradient conflict problem is a critical challenge that must be addressed when implementing MTL. By adopting appropriate strategies and techniques, such as gradient surgery, gradient conflicts can be effectively alleviated, thereby improving the performance of MTL models in pulmonary TB CT image analysis. On the other hand, WSSS techniques have important applications in reducing the dependence of CAD models on pixel-level annotations of pulmonary TB lesions in the lungs. Among them, CAM and its variants (such as Grad-CAM) are commonly used methods for generating pseudo-annotations for WSSS. They can leverage classification network scores and gradient information obtained through backpropagation to

generate heatmaps of salient regions for each class in the image, enabling localization of the target regions.

In summary, technologies such as deep learning, weakly supervised learning, and MTL have brought new opportunities for medical CT image analysis, but also face numerous challenges that need to be addressed. This review has summarized the theoretical foundations of relevant research, laying the groundwork for subsequent methodological innovations and algorithm improvements. To this end, based on the existing work, this thesis will continue to conduct further research, aiming to propose innovative methods and optimize relevant algorithms to advance the development of medical CT image analysis techniques, contributing to improving the accuracy and efficiency of disease diagnosis.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter provides a detailed introduction to the quantitative methodological framework adopted in this research and elucidates the overall structure of the research design and setup. The chapter first identifies the main research questions and objectives of this study, based on a review of the existing literature relevant to this research. It then describes the process and principles followed in constructing the TB analysis dataset used in this research. Building these datasets is crucial for ensuring the effectiveness of subsequent model development and validation. In terms of model and algorithm development, to obtain more comprehensive diagnostic information about pulmonary TB lesion tissue, this research constructs a pulmonary TB CT medical-assisted diagnosis model based on an MTL architecture; however, since this MTL model is designed based on the fully supervised learning paradigm, its training process relies on expensive LC pixel-level annotations. To alleviate this issue, this research designs a WSSS framework for LC. Within this framework, to efficiently obtain pixel-level LC annotation information for pulmonary TB lesions in CT images, this research also constructs a deep learning model composed of a hybrid of Transformer and CNN modules. Furthermore, this chapter describes the experimental environment, performance evaluation metrics, model training, and evaluation processes, to ensure the reproducibility and reliability of the experimental results. Finally, this chapter summarizes the main content discussed.

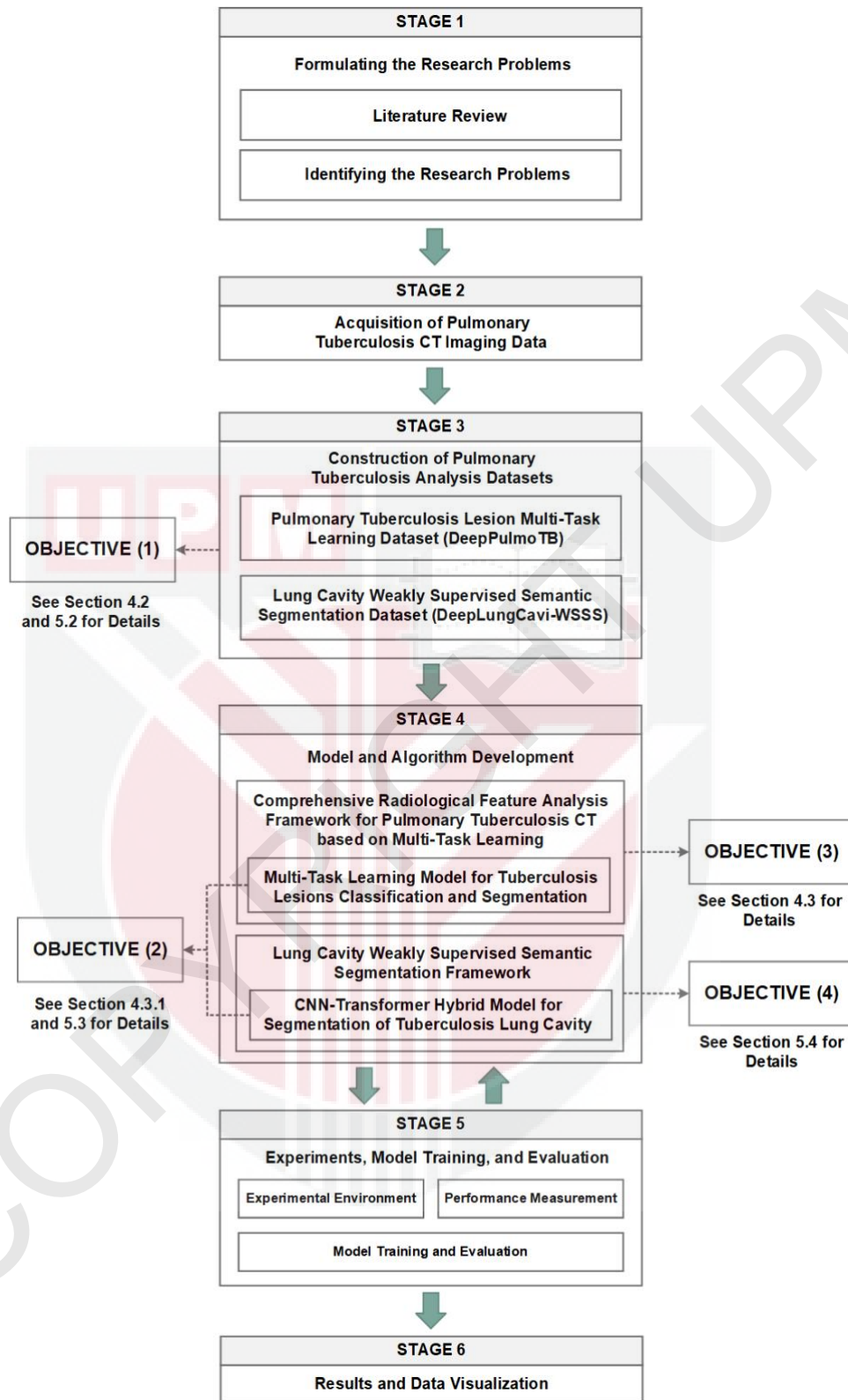


Figure 3.1: Quantitative Research Methodology

3.2 Research Design and Setting

The quantitative research framework of this thesis is shown in Figure 3.1, which includes six interrelated stages carried out in chronological order. These stages are: i) Formulating the Research Problems; ii) Acquisition of pulmonary TB CT Imaging Data; iii) Construction of pulmonary TB Analysis Datasets; iv) Model and Algorithm Development; v) Experiments, Model Training, and Evaluation. vi) Results and Data Visualization. The following section will detail the execution of each stage.

3.2.1 Stage 1: Formulating the Research Problems

The objective of the problem statement is to clarify and define the topic or issue of research, providing a clear direction and focus for the study. This involves a precise description of the research problem, as well as an explanation of why the issue is worth investigating. The problem statement should specify the specific research objectives and expected outcomes, as well as the potential impact of these outcomes on the relevant field. A clear problem statement and definition of terms are crucial for achieving the research aims and objectives because it can serve as a guide for the direction of the study. In addition, correctly identifying and clearly articulating the research problem also helps researchers to focus on key points and efficiently advance their work.

This study takes the statement of the problem as the starting point and foundation, extensively and meticulously reviewing the relevant work of

predecessors in the field of CAD of pulmonary TB CT images, and from it combs out the key issues that remain to be solved in this field.

This study reviewed 736 relevant academic papers collected from multiple academic search engines such as IEEE, Google Scholar, and Springer. These papers covered the following aspects: i) the development status of CAD tools for pulmonary tuberculosis. ii) the characteristics of existing publicly available pulmonary TB CT datasets. iii) Research on CT image comprehensive analysis methods based on the MTL framework. iv) Exploration of deep learning models applicable to medical image analysis tasks, such as CNN (convolutional neural network) and transformers, and their advantages in extracting local and high-level semantic features. v) The development of WSSS methods in the field of medical imaging, and their potential advantages in reducing annotation costs for pulmonary TB CT scans.

A detailed review of related work has revealed the following shortcomings: Firstly, there are gaps and inadequacies in the datasets available for public access for CAD of pulmonary TB. Second, there is a lack of model architectures specifically designed for the analysis of pulmonary TB CT images, and the applicability of existing general models needs to be improved. Additionally, based on literature review study it shows that existing medical image analysis models based on a single task find it difficult to comprehensively obtain comprehensive diagnostic information about pulmonary TB lesions. Finally, the cost of manually annotating pixel-level labels of LC in CT images is high, making it difficult to obtain large-scale, high-quality annotated data.

In summary, based on a comprehensive review and summary of existing related research, we have found that there are still some areas worth further exploration. These unsolved problems are the important starting points for this study. Therefore, based on the literature review, we identify MTL and WSSS as potential methods to apply in the analysis of pulmonary TB CT images, which warrants further research exploration.

3.2.2 Stage 2: Acquisition of Pulmonary Tuberculosis CT Imaging Data

Stage 1 has completed the establishment of the problem statements and research objectives. Stage 2 is pulmonary TB CT data acquisition.

Data forms the foundation for training and evaluating machine learning models. In the development process of CAD models and algorithms for pulmonary TB, obtaining pulmonary TB CT imaging data is crucial. Generally, pulmonary TB CT imaging data can be sourced from multiple channels, including publicly available medical image databases such as ImageCLEF series datasets (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020) and actual clinical case images from hospitals. To ensure the diversity and representativeness of the dataset, data collection should cover various dimensions such as different types of cases, severity of conditions, and types of imaging devices.

In this study, all pulmonary TB CT medical image data involved were collected from ImageCLEF pulmonary TB 2022 (Ionescu et al., 2022). ImageCLEF pulmonary TB 2022 (Ionescu et al., 2022) is a publicly available professional

medical image database that can be access at ImageCLEF website. This database encompasses a large number of authentic clinical cases, providing reliable data sources and broad case coverage, thereby offering high-quality and diversified foundational data support for this study. Utilizing this authoritative open dataset ensures the objectivity and reproducibility of the research results. In this research, ImageCLEF dataset is used as foundation to develop new dataset for this research.

Reason for Re-annotating the ImageCLEF Serials TB Datasets

(Motivation for Annotating the Datasets): In our study, we re-annotated the ImageCLEF serials TB datasets (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020) because its existing annotations did not meet the specific requirements of our research. The original ImageCLEF serials dataset only includes bounding box annotations, which are suitable for object detection but insufficient for our tasks, namely MTL and WSSS. These tasks demand detailed annotations, including precise lesion contours, specific locations, and measurements of TB lesion size to achieve accurate segmentation and analysis. Our re-annotation provides these critical details, allowing for a higher level of precision in the semantic segmentation of TB lesions—information that is essential for training models in MTL and WSSS tasks, which rely on fine-grained lesion boundaries rather than simple bounding boxes.

How Our Work Differs from Other Studies, Such as Alebiosu 2023: Unlike studies such as Alebiosu 2023, which may focus on TB diagnosis or detection tasks that require only general lesion localization (bounding box annotations),

our research emphasizes the segmentation and in-depth analysis of TB lesions. This focus necessitates precise annotations to address the complexities of lesion contouring and size differentiation. While the ImageCLEF dataset serves as a valuable base, our re-annotation enhances its utility by providing the detailed lesion information essential to our study's objectives.

3.2.3 Stage 3: Construction of Pulmonary Tuberculosis Analysis Datasets

During data acquisition in Stage 2, several issues were identified within the existing ImageCLEF dataset. Consequently, in Stage 3, we aim to pursue objective 1, which is the construction of a pulmonary TB dataset. This research involved the construction of two datasets, DeepPulmoTB and DeepLungCavi-WSSS, for the development of CAD methods to aid in the diagnosis of pulmonary TB, as detailed in Sections 4.2 and 5.2, respectively.

3.2.3.1 Pulmonary Tuberculosis Lesion Multi-Task Learning Dataset (DeepPulmoTB)

To address the lack of datasets containing diverse and comprehensive diagnostic annotation in the field of multi-category pulmonary TB semantic segmentation research, this study introduces DeepPulmoTB, an MTL dataset for pulmonary TB. DeepPulmoTB encompasses various aspects of CT radiographic feature information related to pulmonary TB lesions. Different from existing pulmonary TB CT diagnosis data sets (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020), the key feature of the DeepPulmoTB dataset is that it annotates a variety of key

identification information for pulmonary TB diagnosis and integrates it into a single semantic segmentation data set, as shown in Figure 3.2. From the image, we can observe that the ImageCLEF TB 2022 dataset only provides bounding boxes annotating the approximate location of lung cavities, while the DeepPulmoTB dataset provides precise pixel-level segmentation annotation information for multiple key diagnostic indicators of pulmonary TB.

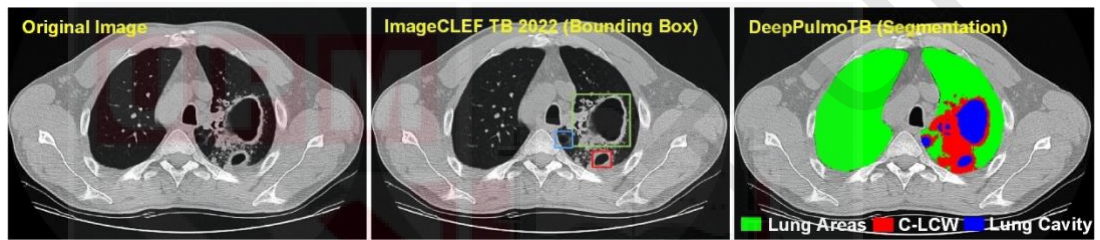


Figure 3.2: Comparison Between the ImageCLEF TB 2022 Dataset and the DeepPulmoTB Dataset

Specifically, DeepPulmoTB supports two key pulmonary TB disease analysis tasks: multi-category tissue semantic segmentation and LC attribute classification, as shown in Figure 3.3. First, the multi-category tissue semantic segmentation task includes three basic pulmonary TB diagnostic categories: LC, C-LCW (consolidation or LC wall), and lung areas. Second, the LC attribute classification task includes descriptive information about the size and number of LC.

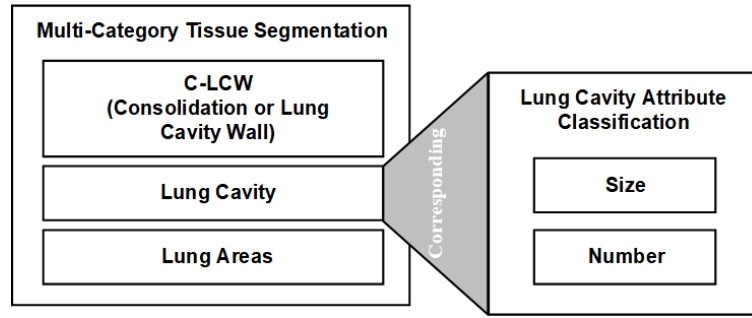


Figure 3.3: Annotated Information Included in the DeepPulmoTB Dataset

Compared to the ImageCLEF TB series dataset, a significant improvement introduced by DeepPulmoTB is the incorporation of identification indicators for lung cavities and C-LCW (consolidation or lung cavity wall) lesion tissues. Moreover, during the construction of this dataset, all ground truths were annotated by experienced medical experts. For details on the construction process of the DeepPulmoTB data set, see Section 4.2.

Dataset Annotation Process and Justification for Methodology. In designing the dataset annotation process, we aimed to achieve high standards of accuracy and clinical relevance by implementing a rigorous method that ensures both consistency and reliability in the segmentation and classification of LC in pulmonary TB cases. The choice of specific annotation methods was carefully considered based on the need for precision, consistency, and the ability to capture variability among experts. The workflow consisted of the following steps:

Step 1: Observer Selection. We chose three board-certified radiologists specializing in thoracic imaging, each with over 10 years of experience in diagnosing pulmonary diseases, including TB. This selection was based on

the rationale that radiologists with extensive expertise in thoracic imaging are best equipped to accurately identify and segment subtle TB-related lesions. Their background was intended to enhance the precision of the annotations and ensure that they would reflect clinically relevant details.

Step 2: Standardized Training and Guidelines. To maintain annotation consistency, each observer was provided with detailed and standardized guidelines. This training included specific instructions on LC segmentation protocols and classification criteria, such as identifying and differentiating the number and characteristics of LCs. Standardized guidelines were essential to reduce subjective variation and to align each radiologist's interpretation, thus enabling more consistent data across observers. Uniform training also ensured that the annotations would meet the requirements of complex analysis tasks that demand high precision.

Step 3: Independent Annotation. Each radiologist independently annotated the CT images, performing semantic segmentation across the axial, sagittal, and coronal planes. We used ITK-SNAP as the primary annotation software for its accuracy in segmentation and its compatibility with Nifti format output, which supports precise three-dimensional modeling. Independent annotation was essential to capture individual interpretations of TB lesion characteristics without observer influence. This approach allowed us to document and analyze natural variability among experts, providing valuable insights into potential inconsistencies and enhancing the robustness of the dataset.

Step 4: Review and Consensus Process. A consensus process was implemented to resolve discrepancies between observers. When annotation results differed, the conflicting annotations were re-evaluated, and a consensus rule was applied, where the minority opinion conformed to the majority. This method was chosen to ensure a reliable final annotation while balancing diverse expert perspectives. The consensus process was essential for achieving a high level of accuracy in the final dataset, as it mitigated individual biases and promoted agreement on the most clinically relevant interpretation.

Overall, these annotation methods were selected to produce a high-quality, clinically reliable dataset that captures both the consistency and variability inherent in expert interpretations. The structured methodology supports the creation of a robust dataset suited for advanced analysis and modeling in pulmonary TB research.

3.2.3.2 Lung Cavity Weakly Supervised Semantic Segmentation Dataset (DeepLungCavi-WSSS)

One of the key indicators for the radiological diagnosis of pulmonary TB is LC. In clinical applications, the annotation of ground truth data for pixel-level semantic segmentation of LC in CT images is a high-cost and labor-intensive task, as it requires experienced medical experts to meticulously annotate these three-dimensional CT data on a per-pixel, per-layer basis.

Therefore, given the critical role of LC in TB diagnosis and the complexities involved in data annotation, LC WSSS has emerged as a significant research

area. However, currently, there are no publicly available datasets in the research community dedicated to the development of CAD tools for LC WSSS. To bridge this gap, this research constructed a dataset for LC WSSS, dubbed DeepLungCavi-WSSS. This dataset primarily provides image-level LC attribute category information (for model training), while also including a small amount of pixel-level LC semantic segmentation annotations (for testing model performance), as shown in Figure 3.4. The specific data annotation sample is shown in Figure 3.5. For details on the construction process of the DeepLungCavi-WSSS dataset, see Section 5.2.

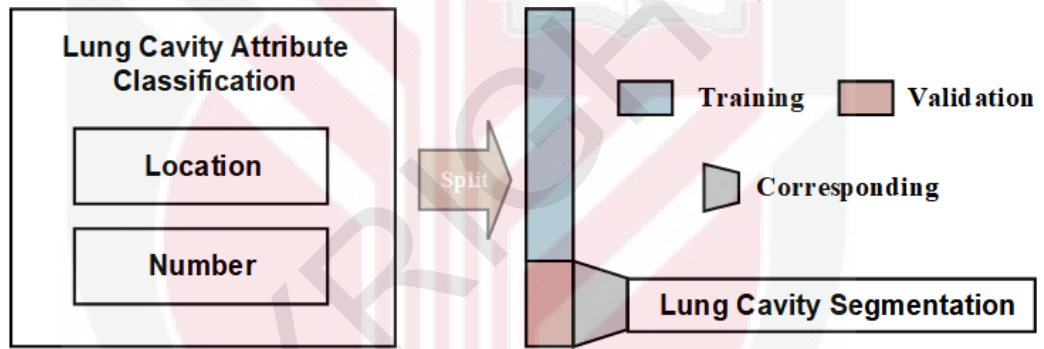


Figure 3.4: Annotation Information Included in the DeepLungCavi-WSSS Dataset

Moreover, the key feature of the DeepLungCavi-WSSS dataset is its relatively low annotation cost. This is because the dataset uses classification annotations of the quantity and location of LC for model training, rather than dense pixel-level semantic segmentation annotations as in traditional methods. In this dataset, semantic segmentation annotations are only used as samples for evaluating model performance, eliminating the need for extensive annotations, which greatly reduces the workload of manual annotation.

Compared to the ImageCLEF TB series dataset, a significant improvement introduced by DeepLungCavi-WSSS is that it annotates two key descriptive information corresponding to the semantic segmentation task, namely the number and location of LC.

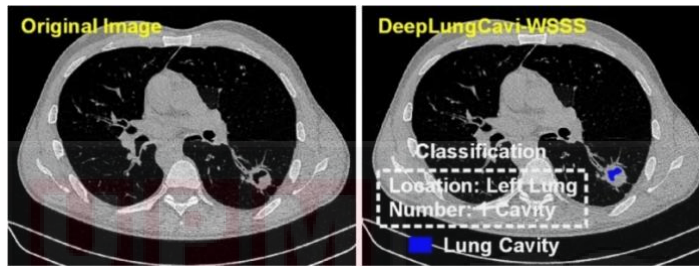


Figure 3.5: Visualization of DeepLungCavi-WSSS Dataset Samples

The DeepLungCavi-WSSS dataset is designed to support machine learning models by leveraging low-cost LC attribute annotations to generate pixel-level LC semantic segmentation annotations. This approach reduces data annotation costs while increasing the volume of training data, making it more feasible to scale machine learning applications for TB diagnosis.

Dataset Annotation Process and Justification for Methodology. In the DeepLungCavi-WSSS dataset, all data were annotated by medical professionals from the Radiology Department at Universiti Putra Malaysia, ensuring high data accuracy and reliability. To further ensure annotation reliability, we established a **Quality Assurance Process** with a standardized workflow to evaluate consistency and variability among professional observers. This process included the following steps:

- **Step 1: Observer Selection.** Three experienced medical professionals or imaging analysts were selected based on their expertise in interpreting tuberculosis lesions on CT images. Their specialized background was crucial for accurately identifying and analyzing TB-specific features, allowing the dataset to capture subtle and complex lesion characteristics that are critical for precise model training. The selection of knowledgeable observers ensures that the annotations are informed by clinical expertise, a necessary condition for developing a reliable medical dataset.
- **Step 2: Standardized Guidelines.** Three experienced medical professionals or imaging analysts were selected based on their expertise in interpreting tuberculosis lesions on CT images. Their specialized background was crucial for accurately identifying and analyzing TB-specific features, allowing the dataset to capture subtle and complex lesion characteristics that are critical for precise model training. The selection of knowledgeable observers ensures that the annotations are informed by clinical expertise, a necessary condition for developing a reliable medical dataset.
- **Step 3: Independent Segmentation Labeling.** Each observer independently performed semantic segmentation of CT images across the X, Y, and Z planes using ITK-SNAP, with the results saved in Nifti format. By conducting segmentation without communication between observers, we were able to capture natural variability in how different experts perceive and outline TB lesions. This independence is critical for

training models that need to handle subtle differences in expert annotations, helping the model generalize across varying interpretations of complex medical images.

- **Step 4: Independent Lung Cavity Attribute Classification.** Observers independently classified lung cavity attributes such as size and count, documenting these classifications in an Excel sheet for further analysis. This attribute classification was performed separately from the segmentation process to provide additional context about lesion characteristics, enabling machine learning models to make use of both high-level categorical data and pixel-level annotations. The separation of attribute classification and segmentation supports weakly-supervised learning by enriching the dataset with a range of annotation types, enhancing the model's ability to learn across multiple levels of detail.
- **Step 5: Repeated Observations.** Observers were given multiple image samples for repeated independent assessment. This step was designed to evaluate the stability and reliability of each observer's annotations over time. Repeated observations provide insights into each expert's consistency, adding another layer of quality assurance to the dataset. By assessing annotation stability, the dataset is equipped to support models that require robust, high-quality data for weakly-supervised segmentation tasks.

-

3.2.4 Stage 4: Model and Algorithm Development

In stage 3, we developed two types of datasets: DeepPulmoTB and DeepLungCavi-WSSS. Each dataset possesses distinct features. Consequently, in stage 4, we designed two separate deep learning frameworks to perform classification tasks on the DeepPulmoTB and DeepLungCavi-WSSS datasets, respectively. The following section provides a concise overview of these frameworks.

3.2.4.1 Comprehensive Radiological Feature Analysis Framework for Pulmonary Tuberculosis CT based on Multi-Task Learning

To improve the diagnostic efficiency of pulmonary TB disease, this study constructed a comprehensive radiological feature analysis framework for pulmonary TB CT based on MTL, as shown in Figure 3.6. The specific implementation details of this framework can be found in Section 4.4.

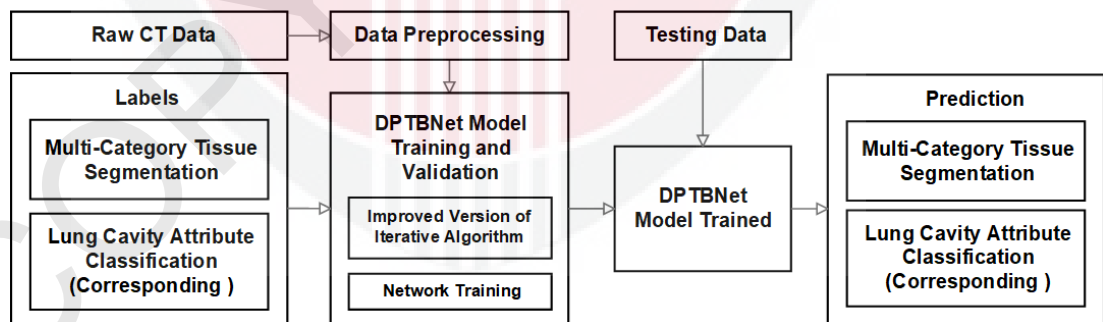


Figure 3.6: Implementation Process of the Comprehensive Radiological Feature Analysis Framework of Tuberculosis CT

This framework includes an automated diagnosis MTL model for pulmonary TB (DPTBNet). The specific implementation details of the DPTBNet model can be found in Section 4.3. This model can automatically capture CT radiographic features of pulmonary TB lesions from multiple aspects. In the design concept

of DPTBNet, to better extract high-level features from images and understand the complex spatial relationships of lesions, it contains two sub-networks: a SwinUnetR network (Tang et al., 2022) for multi-class tissue semantic segmentation, and a lightweight multi-scale network for LC attribute classification.

Specifically, these two sub-networks mainly share low-level feature representations to enable the transfer of useful feature information between different tasks. Useful feature information refers to the relevant and beneficial characteristics extracted from the CT images that are pertinent to both multi-class semantic segmentation of tissues and classification of LC attributes. Furthermore, within this framework, to enhance the model's capability in capturing the salient characteristics of pulmonary TB lesions, such as their shape, size, and pattern, the model undergoes training via an enhanced iterative optimization strategy. This strategy refines the feature maps by integrating the probability maps obtained from the previous few iterations.

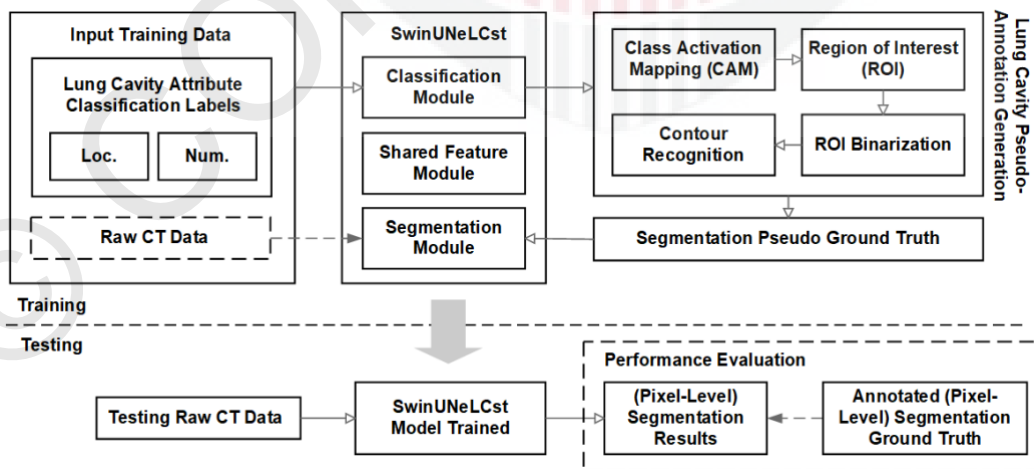


Figure 3.7: Implementation Process of the Lung Cavity Weakly Supervised Semantic Segmentation Framework

3.2.4.2 Lung Cavity Weakly Supervised Semantic Segmentation Framework

To address the difficulty in obtaining pixel-level semantic segmentation annotation data for LC in pulmonary TB CT scan, this research developed a WSSS framework for LC, as shown in Figure 3.7. The specific implementation details of this framework can be found in Section 5.4. This framework introduces a novel CNN-Transformer hybrid model (SwinUNeLCsT). The core idea of SwinUNeLCsT is to represent the features of pulmonary TB CT scan images by fusing local details and global dependencies, thereby enhancing the model's ability to capture LC features. In terms of the implementation process, the framework mainly consists of two stages: training and testing.

During the training stage, first, the image-level LC attribute category information (location and number annotations) input into the SwinUNeLCsT model is converted into a CAM. This CAM then undergoes a series of refinement operations to generate a pseudo-semantic segmentation ground truth, which is used for the segmentation training of the SwinUNeLCsT model. During the testing stage, raw CT image data is input into the trained SwinUNeLCsT model. Then, the output is the pixel-level semantic segmentation annotation of LC.

Overall, this framework can achieve effective segmentation of pulmonary TB lesions using only inexpensive coarse annotations or partial annotations. Through this framework, radiologists need only to classify the number and the approximate location (e.g., left lung, right lung, or both) of LC in the CT scan

to achieve efficient segmentation of the LC. This process eliminates the need for meticulously drawing boundaries, greatly reducing the cost of annotation.

3.2.5 Stage 5: Experiments, Model Training, and Evaluation

3.2.5.1 Experimental Environment

The experimental environment setup of this study is divided into two parts: i) hardware configuration and ii) software environment. Firstly, regarding the hardware configuration, this study utilized a high-performance desktop computer. Its configuration is shown in Table 3.1.

Table 3.1: Hardware Configuration

Device	Parameter
CPU	Intel Xeon Gold 6248R @ 3.00GHz (48 cores)
GPU	2 x NVIDIA Tesla V100 32GB
Memory	256GB DDR4
Storage	4TB NVMe SSD array

Moreover, to accelerate the training process, this study employed multi-GPU parallel acceleration technology. Secondly, in terms of the software environment, we used some common Python libraries such as scikit-learn for machine learning tools, tqdm for progress bar display, etc. Its detailed description is shown in Table 3.2. All dependency libraries were kept up to date with the latest stable versions.

Table 3.2: Software Configuration

Device	Parameter
Operating System	Ubuntu 20.04 LTS
Programming Language	Python 3.8
Deep Learning Framework	PyTorch 1.10.0 + CUDA 11.3
Data Processing Libraries	NumPy 1.21.2, Pandas 1.3.3
Image Processing Library	OpenCV 4.5.5
Visualization Library	Matplotlib 3.4.3

3.2.5.2 Performance Measurement

Performance measurements are critical for assessing the effectiveness and accuracy of models, guiding the improvement process, and facilitating comparisons with existing methods. Therefore, in this section, we outline the performance metrics used to evaluate the proposed TB comprehensive radiological feature analysis framework (detailed in Section 4) and the lung cavity WSSS framework (detailed in Section 5).

- Quantitative Evaluation Metrics for Segmentation:

This study evaluates the segmentation of pulmonary TB lesions and lung areas using three quantitative metrics: i) 95HD, ii) DSC and iii) IoU. These metrics are calculated as follows:

95HD (95th percentile hausdorff distance) metric measures the distance between two sets of points that represent the surface of the segmentation predictions and the ground truth (Kraft, 2020). 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:

$$95HD = \max \left\{ P_{95} \left(\left\{ \min_{\bar{y}' \in \bar{Y}'} y' - \bar{y}' \right\}_{y' \in Y'} \right), P_{95} \left(\left\{ \min_{y' \in Y'} \bar{y}' - y' \right\}_{\bar{y}' \in \bar{Y}'} \right) \right\} \quad (3.1)$$

In formula (3.1), 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 (Kraft, 2020). There are two main reasons for using the 95th percentile Hausdorff Distance (95HD) as the evaluation metric for the semantic segmentation task in this study: i) First, in clinical diagnosis, the precise localization of the edges of target organs or lesions is crucial. The 95HD directly reflects the maximum deviation distance at the edges between the segmentation result and the ground truth, which is of significant clinical diagnostic importance. ii) Second, for tasks such as medical image segmentation, where high accuracy of target boundaries is required, the 95HD is sensitive in detecting the maximum deviation points between the algorithm output and the ground truth annotations. This helps to better assess the quality of the segmentation.

A lower 95HD value indicates better segmentation boundary accuracy, as it means the predicted boundaries are closer to the ground truth on average. Conversely, a higher 95HD means the boundaries deviate more from the ground truth. In terms of interpreting the range of 95HD values: i) An ideal

perfect segmentation would have a 95HD of 0, meaning the predicted boundaries exactly match the ground truth boundaries everywhere. ii) For medical imaging tasks, generally a 95HD below 5-10 pixels is considered a good result, indicating the segmentation boundaries are quite accurate and clinically acceptable. iii) A 95HD in the range of 10-20 pixels may still be reasonable depending on the specific application, object size, and image resolution. But it suggests room for improvement in boundary delineation accuracy. iv) Once the 95HD exceeds 20+ pixels, it implies significant boundary errors that could impact clinical utility for precise measurement or quantification tasks.

Therefore, in summary, as the experimental 95HD values get lower (closer to 0), it demonstrates that the segmentation method is performing well in terms of generating boundary predictions very close to the ground truth across most of the data.

IoU (Intersection over Union) is another metric for quantifying the overlap between the predicted segmentation and the ground truth (He et al., 2021). 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:

$$\text{IoU} = \frac{\sum_{i=1}^I Y_i \cap \hat{Y}_i}{\sum_{i=1}^I Y_i \cup \hat{Y}_i} \quad (3.2)$$

In formula (3.2), $\sum_{i=1}^I Y_i \cap \hat{Y}_i$ 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 (He et al., 2021). The IoU is used as the evaluation metric for the semantic segmentation task in this study for two main reasons: i) First, IoU can be expressed as the ratio of true positives to (true positives + false negatives + false positives). It not only considers the correctly segmented regions (true positives) but also takes into account the missed regions (false negatives) and the over-segmented regions (false positives), providing a more comprehensive evaluation of segmentation quality. ii) Second, for the varying sizes of lesions or organs commonly seen in medical imaging, IoU can effectively distinguish the segmentation performance for different scales of targets, preventing evaluation bias caused by differences in target sizes.

The IoU ranges from 0 to 1, where: i) An IoU of 1 represents a perfect segmentation, where the predicted and ground truth masks completely overlap. ii) An IoU of 0 indicates no overlap between the predicted and ground truth masks.

In general, higher IoU values indicate better segmentation performance. Specifically: i) $\text{IoU} > 0.7$ is considered a good segmentation result for many applications. ii) IoU between 0.5 - 0.7 can be regarded as a moderate or acceptable result. iii) $\text{IoU} < 0.5$ is typically seen as a poor segmentation

quality. However, the exact IoU threshold for "good" performance can vary based on the specific task, object characteristics, and clinical requirements.

DSC (Dice Similarity Coefficient) metric is a statistical tool used to measure the similarity between two samples (Raina et al., 2023). 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):

$$DSC = \frac{2 \sum_{i=1}^I Y_i \hat{Y}_i}{\sum_{i=1}^I Y_i + \sum_{i=1}^I \hat{Y}_i} \quad (3.3)$$

In formula (3.3), $\sum_{i=1}^I Y_i \hat{Y}_i$ 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 (Raina et al., 2023). The DSC is used as the evaluation metric for the semantic segmentation task in this study for two main reasons: i) First, DSC combines true positives, false positives, and false negatives through the formula $2 * \text{true positives} / (2 * \text{true positives} + \text{false positives} + \text{false negatives})$. This allows it to evaluate the completeness and accuracy of the segmentation results in a balanced manner, avoiding the one-sidedness of focusing on a single metric. ii) Second, in medical imaging, the size of different organs or lesions can vary significantly, leading to class imbalance issues. DSC assigns the same weight to targets of different sizes, preventing the evaluation results from being dominated by larger targets.

The DSC ranges from 0 to 1, with: i) A DSC of 1 indicates a perfect segmentation where the predicted and ground truth masks completely overlap. ii) A DSC of 0 means there is no overlap between the predicted and ground truth segmentations.

In general, higher DSC values indicate better segmentation quality. The range of DSC values can be interpreted as follows: i) DSC > 0.7 is typically considered a good segmentation result suitable for many applications. ii) DSC between 0.5 - 0.7 can be regarded as a moderate or potentially acceptable result, depending on the specific use case. iii) DSC < 0.5 is often seen as a poor segmentation quality that needs improvement. However, the exact threshold for "good" DSC can vary based on the task, object characteristics, and clinical requirements.

- Quantitative Evaluation Metrics for Classification:

For the quantitative evaluation of pulmonary TB lesions attributes classification, this study utilizes REC (recall), PRE (precision), ACC (accuracy), FPR (false positive rate), F1 (F1-score), confusion matrix, ROC (receiver operating characteristic curves), and AUC (area under the ROC curve):

ACC (He et al., 2021): Defined as the proportion of true results (both true positives and true negatives) in the total population, calculated as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.4)$$

REC (Raina et al., 2023): Also known as sensitivity, measures the proportion of actual positives that are correctly identified, calculated as:

$$\text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.5)$$

PRE (Raina et al., 2023): Indicates the proportion of positive identifications that were actually correct, calculated as:

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.6)$$

FPR (Raina et al., 2023): Reflects the proportion of actual negatives that are incorrectly identified as positive, calculated as:

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (3.7)$$

F1 (Raina et al., 2023): A harmonic mean of precision and recall, providing a balance between the two, calculated as:

$$\text{F1} = 2 \times \frac{\text{PRE} \times \text{REC}}{\text{PRE} + \text{REC}} \quad (3.8)$$

These metrics are defined considering the following terms: i) True Positives (TP): Instances that were correctly predicted as positive. ii) True Negatives (TN): Instances that were correctly predicted as negative. iii) False Positives (FP): Instances that were incorrectly predicted as positive, also known as Type

I error. iv) False Negatives (FN): Instances that were incorrectly predicted as negative, also known as Type II error.

Confusion Matrix (Kraft, 2020).: This metrics is a table used to describe the performance of a classification model on a set of data for which the true values are known, as shown in Table 3.3. It is organized into four quadrants, each representing a different type of outcome between the predicted and actual classifications: TP, TN, FP, and FN.

ROC: This curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the True Positive Rate (TPR, also known as recall or sensitivity) against the False Positive Rate (FPR) at various threshold settings.

The TPR is defined as $TPR = \frac{TP}{TP + FN}$, and the FPR is defined as

$$FPR = \frac{FP}{FP + TN}.$$

Table 3.3: Structure of Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

AUC: This Curve measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). The AUC represents a probability measure for the classifier's ability to distinguish between the positive and negative classes. The higher the AUC, the better the model is at predicting 0s

as 0s and 1s as 1s. An AUC of 0.5 suggests no discrimination ability (equivalent to random guessing), and an AUC of 1.0 represents a perfect model that makes no mistakes in classification.

3.2.5.3 Model Training and Evaluation

In the model training process, firstly, for multi-task model training, this research introduced a modified version of the Sigmoid Focal Loss (Kraft, 2020) and a novel segmentation loss function (He et al., 2021) based on the combination of Dice loss (He et al., 2021) to optimize the classification and segmentation tasks respectively. Secondly, for the training of the WSSS model, this thesis adopted a combination of binary cross-entropy loss and multi-label soft margin loss to optimize the segmentation and classification tasks respectively.

To train the models more effectively, this study adopted the AdamW optimizer, which incorporates a weight decay strategy to help prevent overfitting and enhance the model's generalization capability. Regarding hyperparameter selection, this study determined the optimal learning rate, batch size, and training epochs through a series of pre-experiments (e.g., grid search) to ensure the models achieve the best performance on all tasks. Additionally, an early stopping strategy was introduced to avoid overfitting and shorten the training time.

To evaluate the model performance comprehensively and objectively, this study adopted a 5-fold cross-validation strategy. Cross-validation maximizes the use of limited data resources and ensures that the model's evaluation

results on different data subsets are statistically significant. Specifically, this study randomly divided the entire dataset into five independent subsets, alternately using four of these subsets for model training and the remaining one for testing. This strategy not only reduces the impact of randomness in the evaluation process but, more importantly, accurately reflects the model's real generalization ability on unseen data, thereby reducing the risk of overfitting. By synthesizing the evaluation results from five cross-validations, this study was able to obtain a more robust and reliable estimate of model performance.

3.4 Summary

This chapter provides a detailed exposition of the research method framework adopted by this study. Initially, upon extensively reviewing existing literature, this work identified the core issues and objectives to be addressed in this research. Through the literature review, this study revealed the current gaps in datasets used for CAD of pulmonary TB. To fill this gap, this study constructed two datasets (DeepPulmoTB and DeepLungCavi-WSSS) that can be used for different pulmonary TB disease analysis tasks. Subsequently, based on the DeepPulmoTB dataset, this study proposed an MTL-based model for CT-assisted diagnosis of pulmonary TB (DPTBNet). DPTBNet can be used to simultaneously perform multitask tissue semantic segmentation and LC attribute classification for pulmonary TB. Moreover, to reduce the annotation cost of pixel-level labeled data, this study designed a WSSS framework to efficiently capture semantic segmentation information of LC in CT images. Lastly, this chapter introduces the experimental setup, evaluation

metrics, and model training and assessment procedures to ensure the reproducibility and reliability of the experimental results.



CHAPTER 4

ANNOTATION PROCESS AND GUIDELINES

4.1 Introduction

The annotation process in this chapter serves as a crucial step in constructing two specialized datasets that will support advanced deep learning applications in pulmonary TB radiological analysis. These datasets, DeepPulmoTB and DeepLungCavi-WSSS, are designed to provide high-quality, multi-faceted annotations for different but complementary analysis objectives, leveraging both fine-grained and coarse-grained annotations to enable effective model training and evaluation.

- i) The first objective is to construct a comprehensive MTL dataset for pulmonary TB radiological analysis (DeepPulmoTB). This dataset contains multi-faceted radiological feature annotations such as semantic segmentation, size, and quantity of pulmonary TB lesions, providing multi-dimensional supervisory signals for deep learning models.
- ii) The second objective is to construct a WSSS dataset for LC radiological analysis (DeepLungCavi-WSSS). This dataset aims to provide a robust foundation for training and evaluating WSSS models by leveraging coarse-grained LC annotations (such as LC attributes classification) from CT scans.

4.2 Pulmonary Tuberculosis Lesion Multi-Task Learning Dataset (DeepPulmoTB)

As outlined in section 3.2.3.1, this study constructed a multi-task learning dataset for pulmonary TB (DeepPulmoTB), addressing Objective 1.

4.2.1 Data Sources

In DeepPulmoTB, the pulmonary CT scan images for pulmonary TB are collected from the ImageCLEF TB series datasets (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020). The ImageCLEF pulmonary TB is an international competition focused on pulmonary TB image analysis and is part of the ImageCLEF (Image Cross-Language Evaluation Forum). ImageCLEF is a competition platform aimed at promoting cross-language image retrieval and classification technologies, including various tasks ranging from medical image analysis to the retrieval of natural images. As a standard pulmonary TB CT image dataset, the ImageCLEF pulmonary TB series dataset (Ionescu et al., 2021; Ionescu et al., 2019; Ionescu et al., 2022; Kozlovski et al., 2020) covers cases that have been rigorously screened by medical experts and thus only includes data from patients diagnosed with pulmonary TB.

In the construction of the DeepPulmoTB dataset, to ensure data consistency and accuracy of research results, this study further refined data selection criteria and clearly defined inclusion and exclusion criteria. Specifically, this study focused on analyzing pulmonary TB lesions, excluding extraPTB cases from the ImageCLEF TB series datasets (Ionescu et al., 2021; Ionescu et al.,

2019; Ionescu et al., 2022; Kozlovski et al., 2020), such as endobronchial pulmonary TB and tuberculous pleurisy. As a result, this study selected a total of 354 chest CT scan images from different patients with pulmonary TB. Some samples from the ImageCLEF TB dataset as shown in Figure 4.1. Moreover, to accommodate different types of CT scan needs, DeepPulmoTB dataset includes thin and thick CT scans. Each patient's CT series consists of 125 to 512 axial projection images.

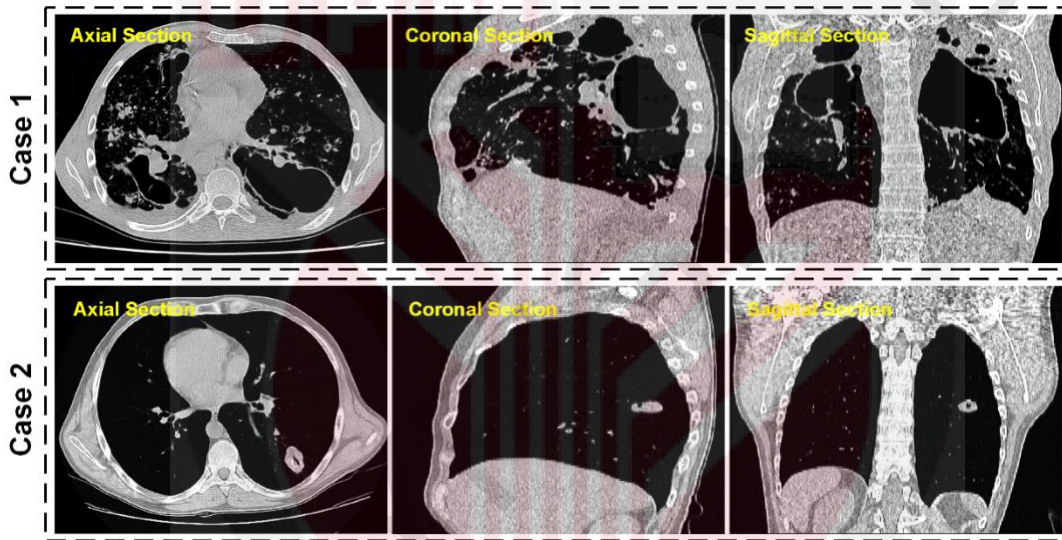


Figure 4.1: Three-dimensional Visualization of Sample Cases from the ImageCLEF TB Dataset

Role of Observers in the Annotation Process and Scope of Annotated Images. In the annotation process for the DeepLungCavi-WSSS dataset, three experienced medical professionals were responsible for annotating all 354 CT images. Each observer independently annotated every image, ensuring that the entire dataset received comprehensive, expert-reviewed segmentation and classification annotations. This complete coverage by multiple observers was crucial for achieving a robust and reliable dataset, as

it allowed for a full dataset assessment of consistency and variability across different expert interpretations.

Miliary TB: Although miliary pulmonary TB affects the lungs, its pathological process usually involves the widespread distribution of micronodules in the lungs and other organs, which is significantly different from the imaging appearance of other types of pulmonary TB (Sharma et al., 2016), as shown in Figure 4.2. Because of the systemic nature of miliary TB, diagnostic and treatment strategies need to consider the entire body and not just the pulmonary manifestations. In addition, the CT imaging characteristics of miliary TB may be confused with a variety of other lung diseases, which will increase the difficulty of automatic identification and classification and may lead to a decrease in the performance of CAD systems. Given these challenges and the limited number of miliary TB samples in the current dataset, it was decided to exclude miliary TB to ensure higher accuracy and reliability of the training model in identifying and classifying the more common forms of TB.

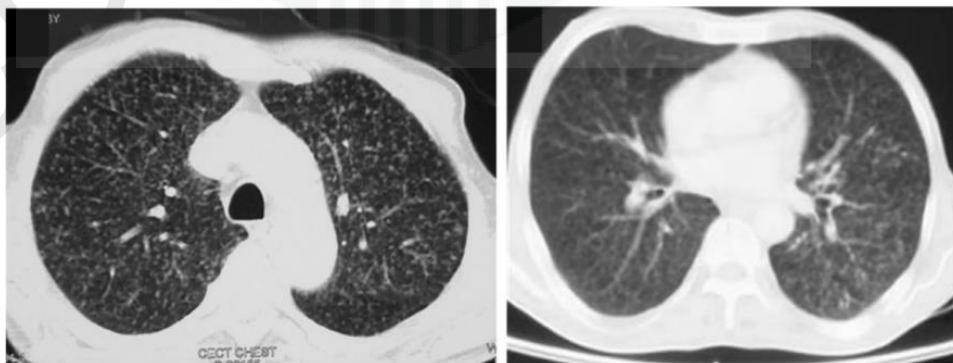


Figure 4.2: Visualization of Miliary TB Cases (Sharma et al., 2016)
(Source: Sharma et al., 2016)

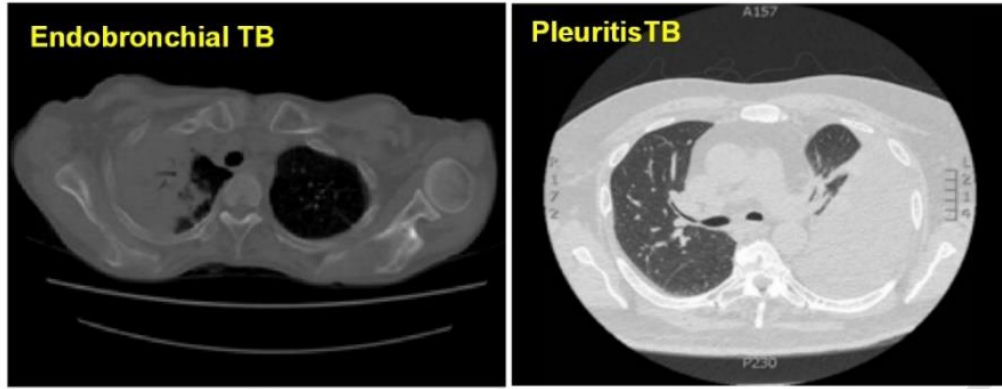


Figure 4.3: Visualization of Endobronchial TB and TB Pleuritis Cases

Endobronchial TB and TB Pleuritis: For endobronchial pulmonary TB, it is caused by the infection of *Mycobacterium TB* on the epithelial cells of the trachea and endobronchial mucosa. Its primary lesions are located in the trachea and bronchi, rather than the lung parenchyma, manifesting as thickening of the tracheal and endobronchial walls and narrowing of the lumens (Aneja et al., 2014), as shown in Figure 4.3 (Left). Secondly, tuberculous pleurisy is another form of pulmonary TB infection, mainly due to *Mycobacterium TB* infection of the pleura (the thin membrane between the lungs and the chest wall), presenting as pleural effusion and/or thickening of pleural adhesions, also not in the lung parenchyma, as shown in Figure 4.3 (Right). However, this study focuses on pulmonary TB. Therefore, for patients with solely endobronchial pulmonary TB and simple tuberculous pleurisy, since there are no lesions in the lung parenchyma, we need to exclude them.

4.2.2 Data Annotation

In the DeepPulmoTB dataset, all data are annotated by medical professionals from the Radiology Department of Universiti Putra Malaysia, ensuring the accuracy and reliability of the data. Overall, the DeepPulmoTB dataset

includes two different tasks: "LC Attribute Classification" and "TB Multi-category Tissue Segmentation". In the "LC Attribute Classification" task, medical professionals annotate the number of cavities in the patient's lungs and their respective sizes, as shown in Table 4.1.

Table 4.1: Annotated Classification of the Lung Cavity Size and Number in a Patient's CT scan

Total Cavern Num	Cavern Size	Sample size	Class
No cavities	-	125	1
1- 3 cavities	< 25mm	86	2
	> 25mm	46	3
> 3 cavities	< 25mm	52	4
	> 25mm	45	5

In this table, the "Total Caver Num" column shows statistics for the number of different cavities. "No cavities" means no cavities, "1-3 cavities" means 1 to 3 cavities, and "> 3 cavities" means more than 3 cavities. The "Cavern Size" column indicates the size grouping of the cavities. "< 25mm" indicates that the cavity size is less than 25mm, and "> 25mm" indicates that the cavity size is greater than or equal to 25mm. For the "Class" column, it represents the numerical index of the classification classes.

Clarification on Test Dataset Annotation. All images used in the testing dataset are non-annotated. No annotated data is included in the test dataset, ensuring that the evaluation of model performance is conducted on entirely unmarked images.

Clarification on the Composition of the 354 Chest CT Scans and

Inclusion Criteria of TB Patients. The "354 chest CT scans" refer to scans from 354 individual TB patients. Each patient's CT scan typically contains between 100 and 512 slices, providing comprehensive cross-sectional views of the lungs. All CT scans in the dataset are exclusively from TB patients, including those without LC. The absence of lung cavities in some scans does not indicate that these individuals are free from TB; rather, it reflects variability in the presentation of TB pathology. Patients without lung cavities still have TB, but their disease characteristics do not include cavity formation. This diversity in presentation allows the dataset to capture a broad spectrum of TB manifestations, which is valuable for training robust machine learning models.

Selection of Classes for Lung Cavity Size and Number. The classification scheme in Table 4.1 for LC size and number in a patient's CT scan was designed to capture clinically relevant variations in TB lesion characteristics, with the following justifications for each class:

Class 1 (No Cavities): Patients without cavities represent a distinct group with milder disease or early-stage TB. Classifying these cases separately allows models to learn features that differentiate normal or less severe presentations from cases with cavity formations, which are often associated with advanced TB (Wollenberg et al., 2022).

Class 2 (1–3 Cavities, <25mm): This class captures cases with a small number of small-sized cavities. A threshold of 25mm was chosen based on clinical relevance, as smaller cavities may indicate a different disease

progression or treatment response compared to larger cavities. By isolating cases with fewer and smaller cavities, this class aids in distinguishing mild cavity formation from more severe presentations (Wollenberg et al., 2022).

Class 3 (1–3 Cavities, >25mm): This class includes patients with a small number of larger cavities. The presence of larger cavities, even if few in number, can indicate more advanced or aggressive TB. Separating small and large cavity sizes helps the model recognize the impact of cavity size on disease severity, providing nuanced insights for clinical applications (Wollenberg et al., 2022).

Class 4 (>3 Cavities, <25mm): This class encompasses cases with multiple small cavities, which can signal a diffuse spread of TB across the lungs. The combination of higher cavity count with smaller sizes may represent a distinct pattern of disease progression. This class helps differentiate cases with numerous small cavities from those with fewer or larger lesions, supporting detailed lesion analysis (Wollenberg et al., 2022).

Class 5 (>3 Cavities, >25mm): Representing the most severe cases with multiple large cavities, this class characterizes advanced TB with widespread lung involvement. Such cases are typically associated with higher severity and poorer prognosis. By identifying this class, the model can better handle complex cases with extensive lung damage (Wollenberg et al., 2022).

Overall, this classification structure allows for a detailed and clinically meaningful breakdown of TB lesion characteristics, enabling the model to

capture variations in both the size and number of cavities. This granularity supports the dataset's objective to enhance model performance in distinguishing varying levels of TB severity and informs nuanced diagnostic and treatment strategies. Visualization of annotated classification examples of LC sizes and quantities (classes 1 to 5) in patient CT scan images, as shown in Figure 4.4.

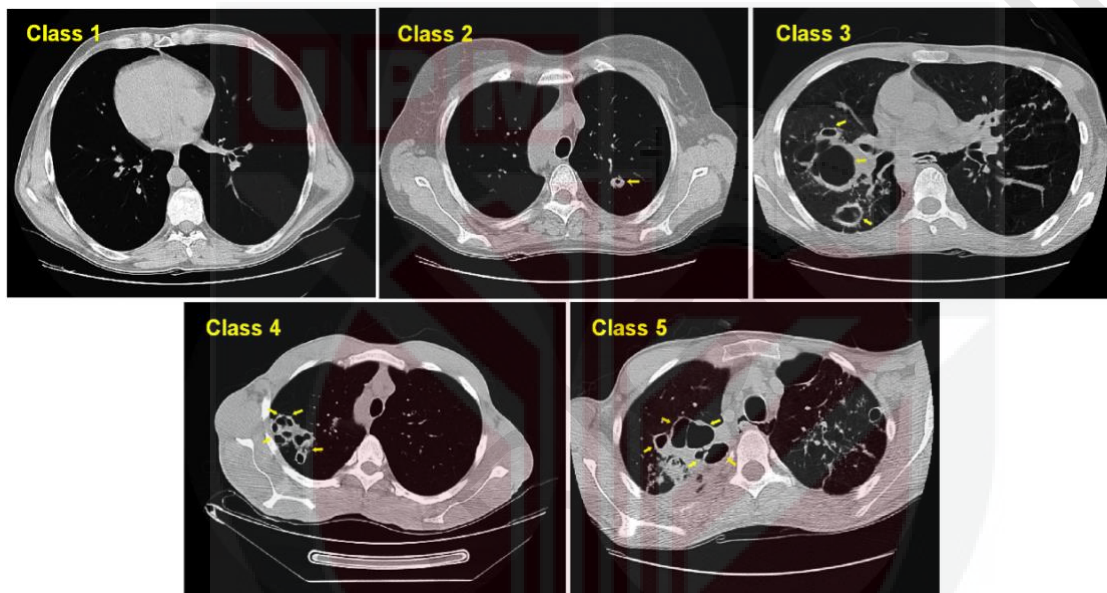


Figure 4.4: Visualization of (Classes 1 to 5) Annotated Classification Examples of Lung Cavity Sizes and Number in Patient CT Scans

In the "TB Multi-category Tissue Segmentation" task, as most of the CT images collected in the ImageCLEF series datasets are non-enhanced CT, it is difficult for medical professionals to distinguish regions of similar density, such as the LC wall and consolidation, by just observing the non-enhanced CT images, as shown in Figure 4.5.

However, if the LC wall and consolidation lesions are not annotated, much critical information for the diagnosis of pulmonary TB would be lost, which is

detrimental to the precise diagnosis of the disease. Therefore, to provide a more comprehensive diagnosis of pulmonary TB and to reduce the complexity of annotation and inconsistency in labeling, this study has categorized consolidation or LC wall into one semantic segmentation annotation category, namely Consolidation or LC Wall (C-LCW). In this segmentation task, this study has marked each lung CT slice image with semantic segmentation information for pulmonary TB diagnosis, which includes the LC, C-LCW, and the lung areas, as shown in Figure 4.6. This study believes that this kind of segmentation annotation will better aid the model in distinguishing between normal and abnormal pulmonary tissue in images.

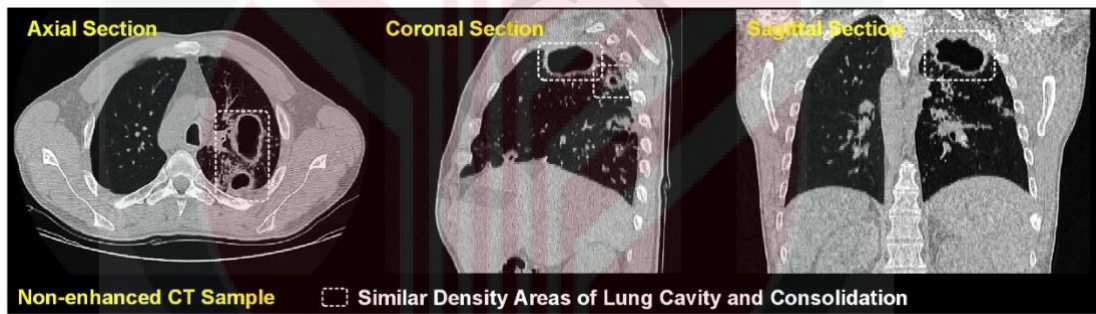


Figure 4.5: Visualization of Areas with Similar Densities in Lung Cavities and Consolidations

Justification for Including Left/Right Lung Identification to Enhance Diagnostic Relevance. The identification of lesion location (left or right lung) may appear to be a straightforward task, but it plays a crucial role in enhancing the overall diagnostic accuracy and clinical relevance of the model. In cases of tuberculosis, the distribution and progression of lesions between the left and right lungs can provide valuable insights into disease severity, spread, and response to treatment. Including left/right location identification as part of the model's capabilities allows for a more comprehensive analysis, enabling the

model to differentiate between unilateral and bilateral cases, which may impact clinical decisions. Moreover, integrating this feature does not significantly increase model complexity but rather enhances its functionality by supporting a more nuanced understanding of lung pathology. By embedding this spatial differentiation, the model is better equipped to handle complex real-world cases, ultimately improving its robustness and applicability in clinical settings.

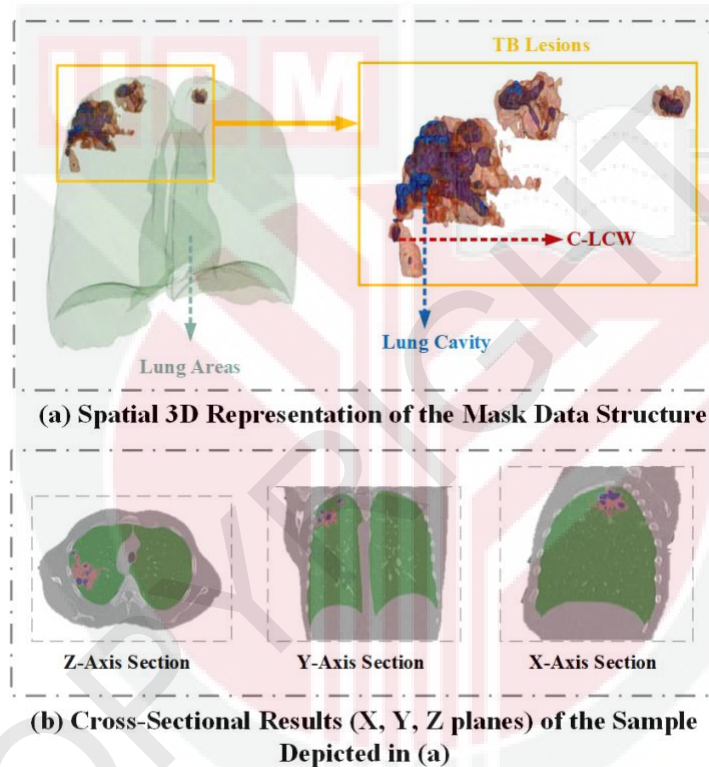


Figure 4.6: Visualization of Mask Data for Each Category in the DeepPulmoTB Dataset

Clarification on Annotation Impact and Use of Non-Annotated Images for

Evaluation. The colored regions in the images represent the annotated masks overlaid for visualization purposes; they do not alter the original pixel values of the CT images. The annotations are added as separate mask layers, ensuring that the integrity of the original image data remains unaffected. Only non-annotated, original images are used for evaluation, ensuring that the

model assessment is free from any potential biases introduced by annotated images. This approach preserves the reliability of the assessment by relying solely on raw, unaltered images for recognition tasks.

Distinguishing Features and Novel Contributions of the DeepPulmoTB

Dataset. The DeepPulmoTB dataset is a newly constructed, comprehensively annotated multi-task learning dataset specifically designed for TB diagnosis, distinguishing it from previous datasets in several keyways. Firstly, DeepPulmoTB supports both segmentation and classification tasks, making it uniquely suited for multi-task learning applications. For the segmentation task, the dataset includes three essential TB diagnostic categories: lung cavity, C-LCW (a unified category combining Consolidation and Lung Cavity Wall), and lung areas. This refined segmentation annotation provides detailed and clinically relevant labels that capture complex aspects of TB pathology, facilitating more precise model training for nuanced lesion detection and segmentation. Secondly, for the classification task, DeepPulmoTB provides specific annotations for the descriptive attributes of lung cavities, such as their size and quantity. These annotations add another layer of diagnostic value, enabling models not only to locate lesions but also to categorize them based on clinically significant features. This combination of segmentation and classification tasks within a single dataset allows for a more comprehensive approach to TB diagnosis, supporting models that require both spatial and descriptive understanding of TB lesions. These unique characteristics and enriched annotations set DeepPulmoTB apart from previous datasets, offering

a refined and targeted resource that better aligns with the complex needs of TB diagnosis in medical imaging.

Annotation Process: Number of Medical Professionals and Conflict Resolution. To ensure high-quality annotations, three experienced medical experts or imaging analysts participated in the annotation process for this study. These professionals were selected for their expertise in identifying and analyzing pulmonary TB lesions, ensuring that the annotations reflected a thorough understanding of the disease's imaging characteristics. The annotation process followed several structured steps to maintain consistency and resolve conflicts:

4.2.2.1 Step 1: Selection of Observers

Three experienced medical professionals were chosen as observers, each with relevant expertise in interpreting TB lesions. This selection ensured that the dataset annotations were carried out by individuals with the necessary skills for accurate segmentation and classification.

4.2.2.2 Step 2: Provision of Consistent Guidance

Each observer was given detailed guidance, including standardized definitions for lesion boundaries, size measurements, and other critical criteria. This consistent guidance aligned all three experts on the objectives and standards of the annotation process.

4.2.2.3 Step 3: Independent Segmentation Labeling

Each observer independently performed semantic segmentation and labeling on CT images from the X-, Y-, and Z-axes using ITK-SNAP (Y.-q. Zhou et al., 2021). Segmentation was conducted in isolation to capture natural variability without mutual influence, as each expert segmented the images without communication. Annotations were saved as Nifti files for further analysis. The ITK-SNAP operation interface is shown in Figure 4.7.

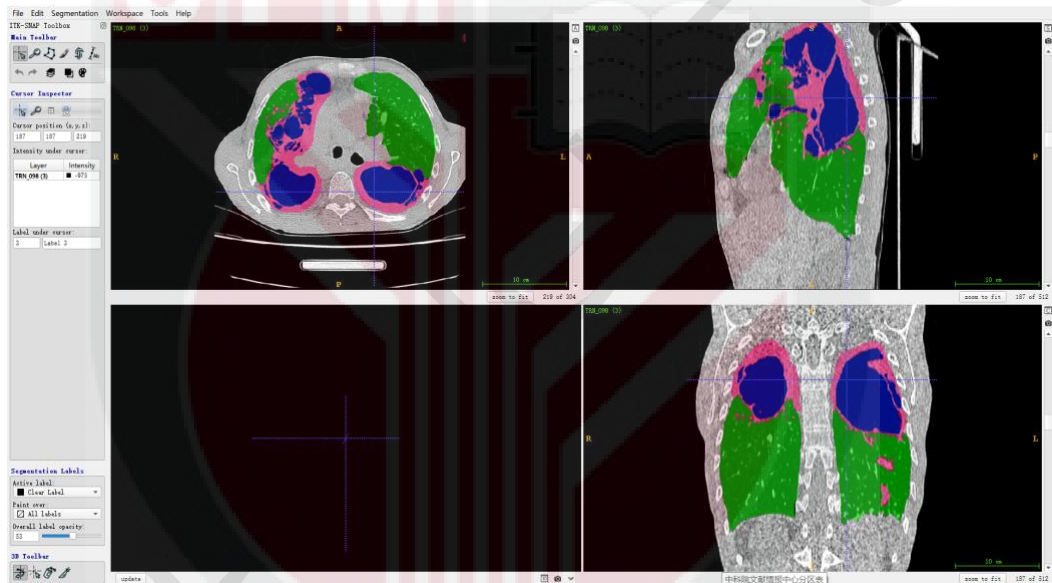


Figure 4.7: Visualization of ITK-SNAP Operation Interface

4.2.2.4 Step 4: Independent Attribute Classification Labeling

Each observer also independently classified lung cavity (LC) attributes, including size and quantity, with results recorded in an Excel table. This step enabled a comprehensive evaluation of each expert's interpretation of lesion characteristics.

4.2.2.5 Step 5: Conflict Resolution through Majority Agreement

In instances where inconsistencies arose between the observers' annotations, a majority consensus approach was used. The “minority conforms to the majority” principle was applied, whereby conflicting annotations were re-evaluated, and the final annotation was determined based on the most consistent interpretation among the three observers. This approach systematically resolved annotation disagreements, contributing to the dataset’s overall accuracy and reliability.

4.2.2.6 Step 6: Repeated Observations

Observers were provided with multiple image samples for repeated independent assessments. This step allowed for evaluation of the stability and reliability of each observer's annotations over time, further reinforcing dataset consistency.

4.2.3 DeepPulmoTB Data Visualization

This study visualizes the results of data annotation for cases TRN_495 and TRN_098 in the DeepPulmoTB dataset, as shown in Figure 4.8. For each example, this study conducts visualizations in two dimensions: three-dimensional space and two-dimensional slices. In the three-dimensional space dimension (left module), this study separately visualizes the three-dimensional structures of various segmentation masks (C-LCW, LC, and lung areas) within a single example from the DeepPulmoTB dataset.

Additionally, this study visualizes the overall spatial distribution of the segmentation masks for each category within the example and displays the positions of the spatial slices used for two-dimensional plane representation. In the two-dimensional slice dimension (right module), this study visualizes cross-sections along the Z, Y, and X axes.

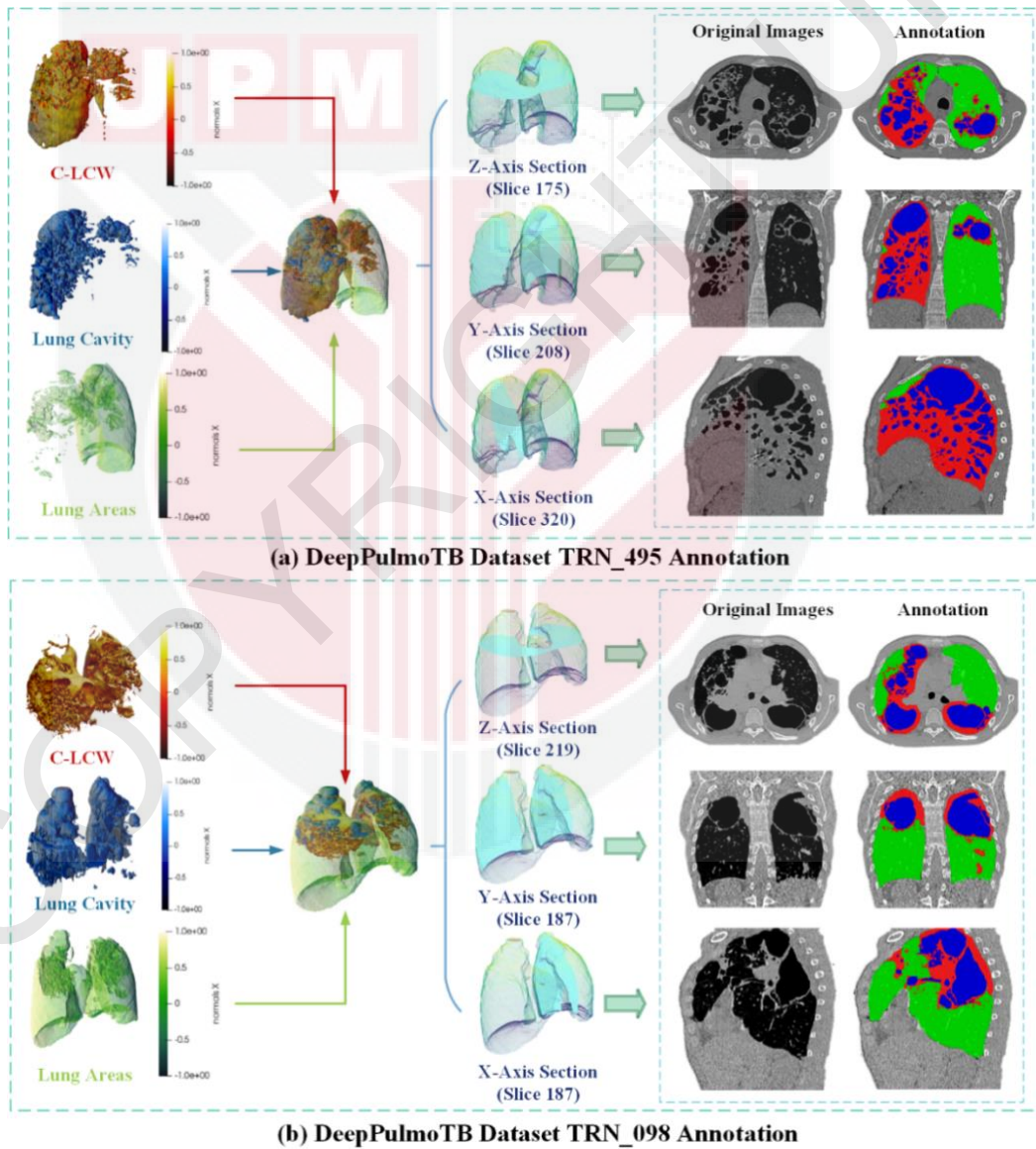


Figure 4.8: Visualization of DeepPulmoTB Example Data

Clarification on the Visualization Order of Cross-Sections (Z, Y, X Axes)

for Lung CT Scans. The cross-sections are visualized in the order of Z, Y, and X axes to align with the conventional viewing orientation in medical imaging, particularly for lung CT scans. In this context, the Z-axis (axial view) is typically presented first as it provides a comprehensive horizontal cross-section of the lungs, commonly used in diagnosing and assessing lung pathologies. The Y-axis (coronal view) follows, offering a front-to-back perspective, and finally, the X-axis (sagittal view) provides a side view of the lungs. This order is chosen to prioritize the most informative views for understanding the spatial distribution and structure of pulmonary lesions, making it easier to assess the pathology from multiple perspectives. Although it may appear unconventional, this arrangement enhances the clarity of lung structure visualization, facilitating better analysis for both radiologists and machine learning models trained on these views.

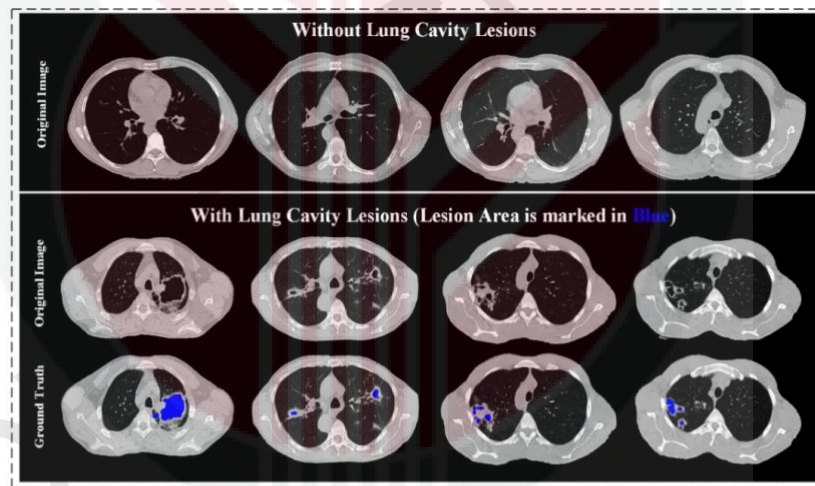
4.3 Pulmonary Tuberculosis Lung Cavity Weakly Supervised Semantic Segmentation Dataset (DeepLungCavi-WSSS)

As outlined in section 3.2.3.2, this study constructs a pulmonary TB LC WSSS dataset (DeepLungCavi-WSSS), addressing Objective 1.

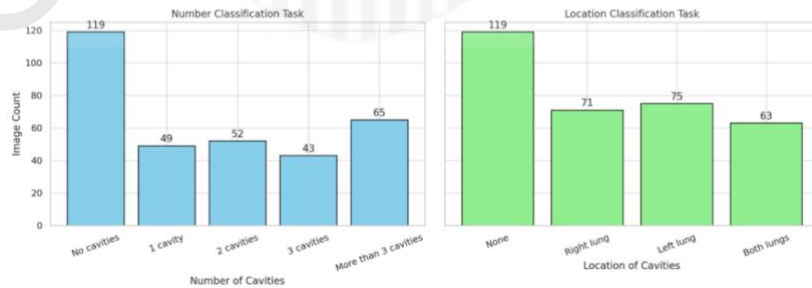
4.3.1 Data sources

The DeepLungCavi-WSSS dataset is derived from sources including the TONGREN People's Hospital and the official ImageCLEF2022 TB challenge (Ionescu et al., 2022). The ImageCLEF 2022 TB challenge dataset is freely available for non-commercial purposes and includes annotations for the

detection of lung cavities and their centroids. Each image in the dataset measures 512×512 pixels. The ImageCLEF 2022 TB challenge is part of the ImageCLEF series of events. ImageCLEF is an international challenge aimed at advancing research in the field of image retrieval and image recognition. It is part of the CLEF (Cross-Language Evaluation Forum), a comprehensive international challenge focusing on multilingual, multimodal, and interactive retrieval issues in information retrieval systems. However, the ImageCLEF2022 TB dataset does not contain annotations for the number classification, location classification, and semantic segmentation of LC lesions.



(a) Visualization of lung cavity semantic segmentation example.



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

Figure 4.9: Annotated Results of the Weakly Supervised Semantic Segmentation Dataset for Lung Cavity

Therefore, to acquire the necessary annotations for number classification, location classification, and semantic segmentation of LC lesions, we enlisted medical experts from the Department of Radiology, Universiti Putra Malaysia to manually annotate the pulmonary TB CT image. The annotation results are shown in Figure 4.9. Specifically, the DeepLungCavi-WSSS dataset comprises 209 CT scan images featuring LC lesions and 119 without lesions, for a total of 328 CT images of tuberculosis patients.

4.3.2 Data annotation

To ensure the gold standard of the annotated data, we evaluated the consistency and variability among different radiologists in the segmentation and classification of LC in pulmonary TB. The construction process of the DeepLungCavi-WSSS dataset is similar to the method used for the deepPulmoTB dataset (the specific implementation plan can be referred to in Section 4.2), including the following main steps:

i) First, this study selected two 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 pulmonary 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 LC 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 (Zhou et al., 2021), 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.

4.4 Summary

This chapter introduces two specialized datasets designed to advance deep learning applications in pulmonary TB analysis. The first, DeepPulmoTB, is an innovative MTL dataset providing comprehensive imaging feature annotations, including semantic segmentation, lesion size, and count, to support multifaceted analysis tasks in computer-aided pulmonary TB diagnosis. Unlike existing pulmonary TB datasets, DeepPulmoTB uniquely enables both multi-category tissue segmentation and pulmonary attribute classification within a single framework. Additionally, in collaboration with radiologists, this chapter presents the first WSSS CT dataset, DeepLungCavi-WSSS, focused specifically on LC analysis for pulmonary TB. This dataset offers coarse-grained annotations essential for training and evaluating WSSS models, establishing a robust foundation for future advancements in TB lesion characterization.

CHAPTER 5

COMPREHENSIVE RADIOLOGICAL FEATURE ANALYSIS FRAMEWORK FOR PULMONARY TUBERCULOSIS CT BASED ON MULTI-TASK LEARNING

5.1 Introduction

In the CAD of pulmonary TB, acquiring comprehensive radiographic characteristics of lesions can help improve the efficiency of disease diagnosis. However, obtaining pulmonary TB diagnosis data that encompasses multiple aspects (such as semantic segmentation information of lesion tissue, size, quantity, etc.) through the STL paradigm often requires training multiple independent models to complete different tasks, which is not economical under resource-limited development conditions (Graham et al., 2023; Hu et al., 2024; Pan et al., 2024; Wu et al., 2024). Moreover, methods based on STL tend to overlook the intrinsic correlations among various diagnostic features of pulmonary TB (Wu et al., 2024; Yan et al., 2024; Yang et al., 2023). This can, to some extent, limit the diversity and richness of the features learned by the model. For example, if a model only performs the task of classifying the number of pulmonary TB lesions, it often misses important information about the size of lesions, their location in the lungs, and their relationships with surrounding tissues, but this information is crucial for enhancing the overall modeling capability for lesion characteristics.

A common approach to address the above issues is MTL techniques. For example, MTL techniques allow a model to not only segment lesion tissues but also classify them (e.g., determine the type, size, or location in the lungs

of the lesions). This learning paradigm facilitates the formation of an attention-guiding mechanism across multiple tasks. Specifically, this mechanism can better help the network focus on image features that are useful for other tasks, thereby obtaining more comprehensive and intuitive clinical diagnostic information for pulmonary TB compared to methods based on STL. Furthermore, in clinical practice, using MTL techniques can complete multiple analysis tasks in a single inference, which can improve efficiency, reduce computational resources required, and save time for doctors to analyze images under resource-limited conditions. Another unique advantage of MTL methods lies in their shared feature representations. This means that during the training process, MTL models can build the inherent correlations among multi-faceted disease diagnostic features through shared features, enabling the network to learn more generalized and robust feature representations, thereby enhancing task performance. Therefore, to improve the diagnostic efficiency of pulmonary TB, this section proposed the following two main objectives centered around MTL techniques:

- i) The first objective is to propose a novel MTL model for the automated diagnosis of pulmonary TB (DPTBNet). Specifically, the DPTBNet model is composed of the SwinUnetR subnet for the segmentation task and a lightweight multi-scale subnet for the classification task.
- ii) The second objective is to develop a comprehensive radiological feature analysis framework.

5.2 Description of the Proposed Multi-Task Learning Network (DPTBNet)

As mentioned in section 3.2.4.1, this study constructed a MTL model (DPTBNet) to capture comprehensive radiological features from pulmonary TB CT images, addressing Objective 2. The reason why DPTBNet is well-suited for analyzing these TB radiological features is its adoption of a multi-scale feature fusion approach. Specifically, DPTBNet integrates feature maps from different levels through modules such as patch merging and residual blocks. This multi-scale feature fusion design facilitates better capture of pulmonary TB lesion features with varying shapes and sizes. Furthermore, the architecture incorporates two distinct task branches for semantic segmentation and classification, enabling simultaneous segmentation and classification of pulmonary TB lesions, thereby providing more comprehensive diagnostic information. The specific structure of the proposed DPTBNet network is illustrated in Figure 5.1.

Overall, in the architectural design of the DPTBNet model, it utilizes SwinUnetR (Tang et al., 2022) as the backbone network for MTL due to its great performance in 3D medical image segmentation. In the SwinUnetR backbone, the encoder part uses the Swin Transformer (Liu et al., 2021) layers to extract high-level features (such as texture patterns, edges, and contextual information) from the input image, preserving both local and global spatial relationships. The decoder part uses the UnetR layers to upsample these features back to the original resolution of the input image, ensuring that the output maintains the same dimensions as the input. It then combines these upsampled features with the corresponding features from the encoder through

skip connections, allowing the model to capture both high-level semantic information and detailed spatial information.

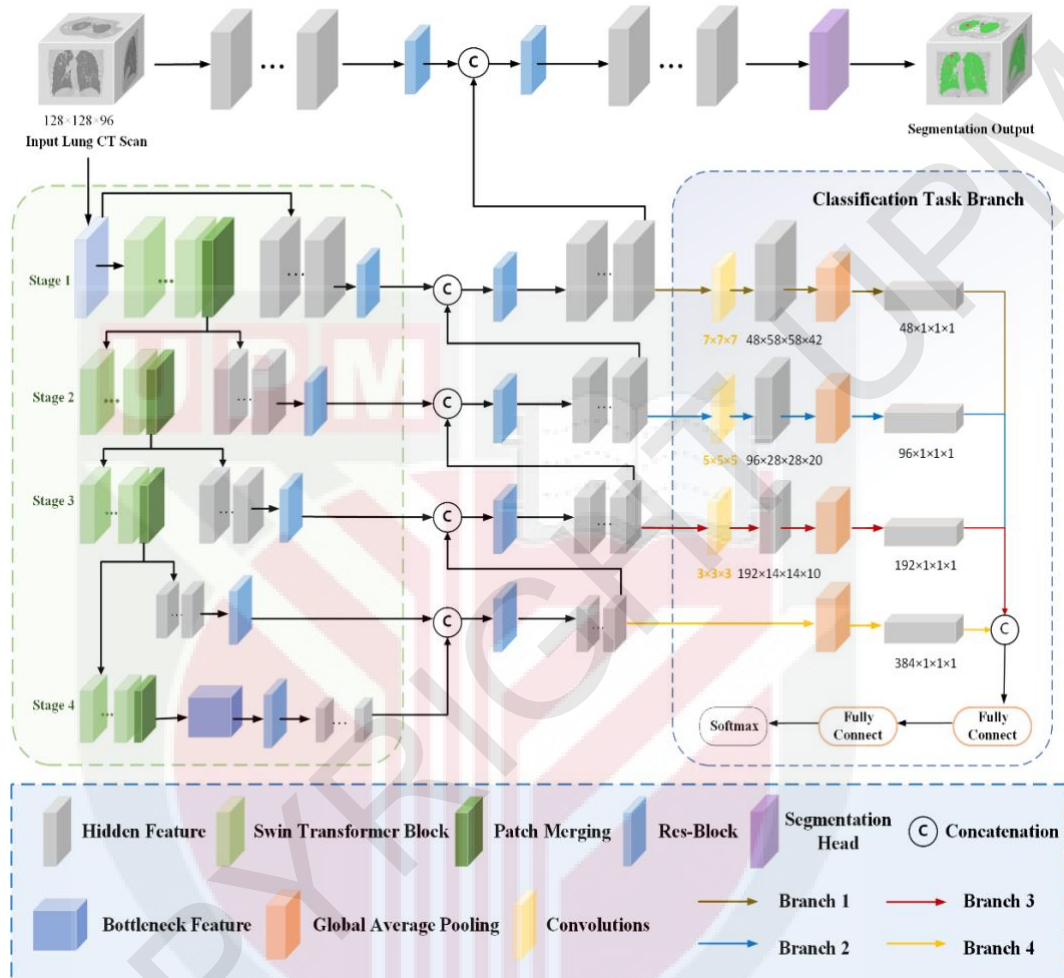


Figure 5.1: Overview of the Proposed Multi-Task Learning Network

High-level semantic information refers to the abstract, overarching features of the image that help in understanding the overall structure and context. These features include patterns, shapes, and relationships that are crucial for identifying and differentiating between various types of tissues, organs, or abnormalities within the image. For example, in the context of pulmonary TB lesions, high-level semantic information would help the model recognize the

presence and general shape of lesions, distinguishing them from healthy tissue.

Detailed spatial information, on the other hand, pertains to the finer, more precise details of the image, such as edges, textures, and small-scale variations. This information is essential for accurately delineating the boundaries of lesions, assessing their size and shape, and identifying subtle differences in tissue characteristics. By preserving detailed spatial information, our model can provide more accurate and reliable segmentation results, which are critical for effective diagnosis and treatment planning.

Overall, by combining high-level semantic information with detailed spatial information, our model can achieve a comprehensive understanding of the image, ensuring that both the overall context and the intricate details are more accurately captured and represented. Moreover, to enhance the representation of pulmonary TB lesion tissue features, this study has increased the channel number or feature dimension size of the intermediate feature map in the model from 48 to 64.

In MTL techniques, each task can extract beneficial feature information from the data of another task. This data-sharing paradigm allows the model to learn target features more effectively from limited data. In DPTBNet, this study introduced an additional supervision signal and learning objective by adding a classification task module in the decoder part for assessing the attributes of pulmonary TB lesion tissue. This enables the model to share knowledge and features of pulmonary TB CT images between the two tasks through the

shared network's intermediate feature representations. In other words, the classification task can provide additional constraints and guidance to help the model better understand and learn the features related to pulmonary TB lesions, thereby improving the performance of the semantic segmentation task. Inspired by this idea, this study designed DPTBNet as a MTL model. DPTBNet utilized SwinUnetR as the shared backbone network to extract common features for both classification and segmentation tasks. Specifically, this study added the classification branch to the decoding part of SwinUnetR, as shown in Figure 5.1 (Classification Task Branch). In this addition process, this study first input the feature maps from branches 1, 2, 3, and 4 into the classification network. Then, this study fused these shared feature maps to perform the classification task. Finally, this study input the fused features into the classification branch, which consists of two FC layers and a SoftMax layer, to classify the attributes of LC in the input pulmonary TB CT scans (Zhou et al., 2021).

Mathematical Justification for the Selection of Layer Sizes in Network Architecture. The layer sizes in the network, such as $48 \times 58 \times 58 \times 42$, are determined through a structured downsampling and upsampling approach that aligns with the input resolution and supports the network's multi-scale feature extraction goals. Beginning with an initial CT scan size of $128 \times 128 \times 96$, the network progressively reduces spatial dimensions at each stage while increasing feature depth. This structured reduction is achieved through a combination of Swin Transformer blocks and patch merging operations, which systematically halve the spatial dimensions and double the channel depth to

retain critical information at each level. Here's a detailed breakdown of this process:

1. Initial Input (Stage 0): The initial input size is $128 \times 128 \times 96$ with a base feature depth C .

2. Stage 1: The first patch merging operation halves each spatial dimension:

Width = $\frac{128}{2} = 64$, Height = $\frac{128}{2} = 64$, Depth = $\frac{96}{2} = 48$. The feature depth doubles to $2C$. Resulting feature size: $64 \times 64 \times 48 \times 2C$.

3. Stage 2: Another patch merging operation reduces spatial dimensions

further: Width = $\frac{64}{2} = 32$, Height = $\frac{64}{2} = 32$, Depth = $\frac{48}{2} = 24$. The feature depth doubles again to $4C$. Resulting feature size: $32 \times 32 \times 24 \times 4C$.

4. Stage 3: The third patch merging step halves spatial dimensions:

Width = $\frac{32}{2} = 16$, Height = $\frac{32}{2} = 16$, Depth = $\frac{24}{2} = 12$. The feature depth doubles to $8C$. Resulting feature size: $16 \times 16 \times 12 \times 8C$.

5. Stage 4: The final patch merging operation halves dimensions again:

Width = $\frac{16}{2} = 8$, Height = $\frac{16}{2} = 8$, Depth = $\frac{12}{2} = 6$. The feature depth increases to $16C$. Resulting feature size: $16 \times 16 \times 12 \times 16C$.

6. Upsampling and Multi-Scale Fusion: To reach the target feature size of $48 \times 48 \times 58 \times 42$, the network applies upsampling or deconvolutional layers in

the multi-scale classification branch. This process restores spatial resolution from the downsampled representations by selectively increasing dimensions in line with the target size requirements. The final feature depth of 42 channels is achieved through a bottleneck or final convolution layer, tailored to meet the classification task's requirements.

The layer size choices, including $48 \times 58 \times 58$, result from a calculated process of downsampling and upsampling that maintains a balance between spatial resolution and feature richness. This approach enables efficient multi-scale processing, supporting both segmentation and classification tasks, while managing computational complexity. The mathematical justification lies in the systematic halving of dimensions at each downsampling stage, followed by selective upsampling to achieve the desired final dimensions in the multi-scale classification module.

Computational Complexity Analysis of DPTBNet for Practicality Assessment. The proposed network builds on the SwinUNETR architecture by adding a multi-scale classification branch that utilizes $3 \times 3 \times 3$, $5 \times 5 \times 5$, and $7 \times 7 \times 7$ convolutions. This addition introduces a multi-scale approach to capture features at different resolutions, which enhances the model's ability to analyze TB lesions of varying sizes. Despite this modification, the computational complexity of DPTBNet remains comparable to that of the original SwinUNETR model. The added classification branch operates with manageable additional parameters due to the limited sizes of the convolutional kernels ($3 \times 3 \times 3$, $5 \times 5 \times 5$, and $7 \times 7 \times 7$), keeping the model efficient and practical for deployment without significantly increasing computational demands. This

balance ensures that DPTBNet maintains practicality while enhancing functionality.

Clarifying the Structural Enhancements and Functional Differences

Between DPTBNet and SwinUNETR. DPTBNet is a MTL model for classification and segmentation, built upon the foundation of SwinUNETR, which was originally designed as a semantic segmentation model. The primary difference between DPTBNet and SwinUNETR lies in the addition of a multi-scale fusion classification module integrated into the decoder section of DPTBNet. In SwinUNETR, the architecture is focused solely on segmentation tasks, where the decoder processes encoded features to produce segmentation outputs. In contrast, DPTBNet introduces a dedicated classification branch within the decoder to enable simultaneous classification and segmentation. This classification branch utilizes multi-scale feature extraction from different stages of the decoder, with features aggregated through global average pooling and concatenation before passing through a fully connected layer. This design allows DPTBNet to leverage both high-resolution and low-resolution features for classification, enhancing its ability to identify complex lung structures and cavity attributes in CT images. By adding this classification module, DPTBNet extends the functionality of SwinUNETR, making it suitable for MTL where both segmentation and classification tasks can be performed efficiently within a single architecture. This enhancement is essential for applications that require comprehensive analysis, such as detecting TB lesions in lung CT scans, where lesion classification and precise segmentation are equally important.

Explanation of the Complexity in Identifying TB Lesions in the Left and Right Lungs and the Need for Deep Learning Architecture.

Identifying TB lesions in the left and right lungs presents additional complexity due to the variability in lesion appearance, location, and structure between the lungs. TB lesions can vary significantly in size, shape, and density, often blending with normal anatomical structures, making manual identification challenging. Additionally, differences in lung anatomy, the proximity of lesions to lung boundaries, and overlapping structures such as blood vessels add further complexity, requiring precise localization and differentiation of TB-affected regions. A deep learning architecture is necessary for this task because it can learn complex patterns and subtle variations in lesion characteristics that are difficult to capture through traditional image processing techniques. Deep learning models, especially convolutional neural networks (CNNs), are capable of analyzing spatial hierarchies within images, allowing them to distinguish nuanced features that indicate TB lesions. By leveraging large datasets, these architectures can generalize across diverse cases, improving their ability to detect TB lesions accurately in both the left and right lungs and reducing the likelihood of missed or misidentified lesions.

5.2.1 Multi-Scale Feature Extraction

To analyze the features obtained by the network at different scales, this study visualized the CAM of branches 1 to 4 of the SwinUnetR decoder, as shown in Figure 5.1. We observed that low-level features mainly capture shape and boundary information, while high-level features summarize the attributes of different targets, which are commonly used for classification tasks. However, we noticed that the distribution of pixel values for different semantic segmentation categories is imbalanced. Specifically, pulmonary TB lesion tissues have fewer pixel values, causing the network to pay more attention to normal lung tissues and neglect large areas of pulmonary TB lesion tissues. To address this issue, this study designed a multi-scale feature fusion architecture for the classification task, as shown in Figure 5.2. This study connected and fused the feature maps from the four branches of SwinUnetR to obtain the classification features.

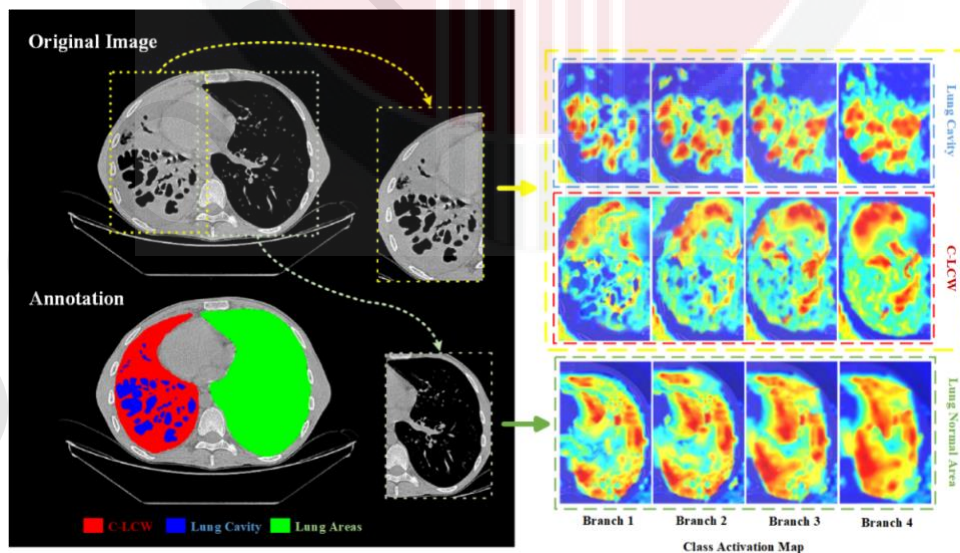


Figure 5.2: 2D Axial Views of Class Activation Maps from Different Stages of the Proposed Network

For branch 1, with a feature dimension of $48 \times 64 \times 64 \times 48$, this dimension typically focuses on the interrelation between different targets. In this branch, this study added a $7 \times 7 \times 7$ convolutional layer to help the network capture larger spatial features and global context information, as well as to better understand the relationships between objects and overall semantics in the CT images.

For branch 2, with a feature dimension of $96 \times 32 \times 32 \times 24$, this dimension focuses more on the local details and relative positions of the targets. In this branch, this study added a $5 \times 5 \times 5$ convolutional layer to enhance the network's ability to capture fine structures and local contextual information, supporting the correlation between tasks and accurate target segmentation.

For branch 3, with a feature dimension of $192 \times 16 \times 16 \times 12$, this dimension typically focuses on local contextual information. In this branch, this study added a $3 \times 3 \times 3$ convolutional layer after the feature map. By introducing smaller convolutional kernels, the network can effectively extract local features and preserve spatial details, thereby enhancing the accuracy of semantic segmentation tasks.

For branch 4, with a feature dimension of $384 \times 8 \times 8 \times 6$, this study did not add additional convolutional layers. Because this dimension is already suitable for capturing medium-sized features and local contextual information. This design efficiently captures medium-scale features without introducing additional complexity, providing rich feature representations for MTL.

As the feature maps from different branches have different sizes and cannot be directly concatenated, this study utilized GAP to transform the feature maps from different branches into channels of the same size. Finally, to help the network better learn complex feature representations and perform feature dimension transformation, we added two FC layers before the SoftMax function.

Overall, such a multi-scale design enables the network to effectively learn features at different scales and provides powerful feature representation for MTL, thereby enhancing the performance of semantic segmentation tasks.

5.3 Description of the Proposed Comprehensive Radiological Feature Analysis Framework for Pulmonary Tuberculosis

As mentioned in section 3.2.4.1; to improve the diagnostic efficiency of pulmonary TB disease, this study constructs a comprehensive radiological feature analysis framework for pulmonary TB, addressing Objective 3.

5.3.1 Data Preprocessing

In data preprocessing, this study leveraged data augmentation modules from the MONAI (Cardoso et al., 2022) library to increase data diversity and improve model generalization. The specific parameter configurations were as follows: i) For input images, this study normalized the pixel values to the range of 0 to 1, ensuring uniform data scaling. ii) To enhance data variability, this study utilized the 'Rand Crop by Pos-Neg Labeled' module for random cropping of input images and labels. The crop size was set as (128,128,96)

with one crop per sample. iii) For further augmentation, this study applied the RandFlipd module, enabling random flipping of input images and labels along the specified axes (x, y, z) . iv) To introduce additional variations, the 'Rand Rotate 90d' module was used, allowing for random rotations of input images and labels on all three axes, with up to 3 rotations of 90 degrees each.

5.3.2 Improved Version of Iterative Algorithm

TB lesions in CT imaging often have complex characteristics, such as unclear boundaries of ground-glass opacities caused by inflammation and exudates, making segmentation challenging. An iterative feature refinement strategy addresses this by progressively optimizing segmentation through multiple iterations, utilizing previous results to guide subsequent segmentation. This approach enhances boundary detection and morphological representation, improving segmentation accuracy and clarity. Taking inspiration from this concept, this study designed an improved version of the iterative feature refinement algorithm based on the work of (Zhou et al., 2021), to further enhance the target feature-capturing capabilities of DPTBNet. The original algorithm and the improved algorithm are shown in Algorithms 5.1 and 5.2, respectively. Specifically, the proposed enhanced algorithm refines input images iteratively by leveraging segmentation probabilities from prior iterations. Initially, DPTBNet processes pulmonary TB CT scans to produce a 3D probability map for TB feature segmentation and classification scores. This map, containing contextual lesion information (as shown in Figure 5.2), guides the model's focus. In later iterations, the input incorporates the previous

iteration's probability map. This approach enhances the model's capability to emphasize regions previously identified as potential pulmonary TB lesions. By introducing this iterative feedback loop, the network can progressively adapt its features based on evolving segmentation probability maps. This mechanism strengthens the model's attention to relevant regions and further refines its predictions.

Algorithm 5.1: The original iterative algorithm

1: **Require:** Original Training volume, V_{orig}

2: **Require:** Classification label, y_{cls}

3: **Require:** Segmentation label, Y_{seg}

4: **Require:** Number of iterations, N

5: **Require:** Network, $f_{\theta}(\text{Input})$

6: for $i=1$ to N do

7: if $i=1$ then

8: Input = V_{orig}

9: else

10: Input = $V_{\text{orig}} + \text{SegProMap}_{i-1}$

11: end if

12: $\text{SegProMap}_i, p_{\text{cls},i} = f_{\theta}(\text{Input})$

13: $P_{\text{seg},i} = \arg \max (\text{SegProMap}_i)$

14: Loss = $L(P_{\text{seg},i}, p_{\text{cls},i}, Y_{\text{seg}}, y_{\text{cls}})$

15: Update θ using ADAM optimizer

16: end for

17: return $P_{\text{seg},N}$, $P_{\text{cls},N}$

The segmentation probability maps generated during each iteration are combined with the original input volume scaled by the hyperparameter η . The main improvements of this algorithm can be summarized as follows:

Algorithm 5.2: The improved iterative algorithm

1: **Require:** Original Training volume, V_{orig}

2: **Require:** Classification and Segmentation label, y_{cls} , Y_{seg}

3: **Require:** Gradient clip threshold, M , Maximum threshold limit, M_{max}

4: Initialize θ with random weights;

5: Set η to control previous segmentation probabilities' influence;

6: **for** $i=1$ to n **do**

7: if $i=1$ then

8: Input = V_{orig} ;

9: else

10: Input = $V_{\text{orig}} + \eta \times \text{SegProbMap}_{i-1}$

11: $\text{SegProbMap}_i, p_{\text{cls},i} = f_{\theta}(\text{Input})$;

12: $P_{\text{seg},i} = \text{argmax}(\text{SegProbMap}_i)$;

13: $\text{Loss} = L(P_{\text{seg},i}, p_{\text{cls},i}, Y_{\text{seg}}, y_{\text{cls}})$;

-
- 14: Calculate gradients ∇Loss with respect to θ ;
 - 15: Clip gradients to ensure the maximum norm is not
 - 16: exceeded: $\nabla \text{Loss} = \text{clip}(\nabla \text{Loss}, M)$;
 - 17: Update θ using ADAM optimizer with gradients ∇Loss ;
 - 18: $\text{SegProbMap}_i = \text{SegProbMap}_i + P_{\text{seg},i}$;
 - 19: $M = \min(M + \text{growthrate} \times i, M_{\text{max}})$;
 - 20: return $\text{SegProbMap}_n, p_{\text{cls},n}$;
-

Segmentation Probability Map Scaling: Unlike the original approach, which simply adds the segmentation probability map from the previous iteration to the current input, our improved algorithm scales this map by a hyperparameter, η , before adding it to the input volume. This modification allows for more nuanced control over the influence of previous segmentation probabilities on the current iteration's input, ensuring a balanced contribution between the original volume and the feedback from previous iterations.

Enhanced Gradient Handling with Gradient Clipping: The improved algorithm introduces a gradient clipping mechanism to stabilize training by preventing excessively large gradients during backpropagation. By clipping gradients to a threshold (denoted as M), this modification mitigates gradient explosion issues, which can cause instability in training—a feature not present in the original algorithm.

Adaptive Gradient Clipping Threshold: Another enhancement in the improved algorithm is the dynamic adjustment of the gradient clipping threshold (M). Starting with an initial value, this threshold gradually increases with each iteration ($growth_rate \times i$), governed by a growth rate parameter (set to 0.001), until reaching a maximum limit (M_{max}). This adaptive threshold mechanism allows the model to optimize gradients dynamically, reducing the risk of overly aggressive clipping that could impede convergence, and improving the model's training robustness.

Iterative Adjustment Based on Evolving Segmentation Maps: In the improved version, segmentation probability maps generated in each iteration are combined with the original input volume after scaling by η . This approach incorporates evolving contextual information more effectively, guiding the model's attention to potential lesion regions with greater precision than the original algorithm, which lacks this iterative feedback adjustment.

These modifications enhance the model's ability to capture complex lesion features in pulmonary TB CT scans. By introducing gradient handling improvements, adaptive learning mechanisms, and segmentation probability influence, the improved iterative algorithm refines segmentation results and improves robustness in training. This iterative feedback loop allows DPTBNet to adapt its focus on relevant regions, continually refining its predictions with each iteration.

5.3.3 Network Training

The imbalance in training data categories is a common factor that can impact the learning effectiveness of a model. For example, in the DeepPulmoTB dataset, the amount of data for categories 3, 4, and 5 is much smaller than that for categories 1 and 2, as shown in Table 4.1. However, this data imbalance will cause the model to focus more on classes with a larger number of samples during training. To tackle this problem, this study has introduced a modified version of the Sigmoid Focal Loss (Chen & Qin, 2022) as the classification loss function:

$$L_{cls}(P_{cls}, Y_{cls}) = -\frac{1}{C} \sum_{i=1}^C \left(w_i \cdot \left(Y_{cls_i} \cdot (1 - P_{cls_i})^\varphi \cdot \log(P_{cls_i}) + (1 - Y_{cls_i}) \cdot P_{cls_i}^\varphi \cdot \log(1 - P_{cls_i}) \right) \right) \quad (5.1)$$

In formula (5.1), P_{cls} is the model's predicted probability vector, where P_{cls_i} represents the predicted probability for the i -th class by the model. Y_{cls} is the true label vector, where Y_{cls_i} represents the true label for the i -th class. C denotes the total number of classes, and its value is 5. w_i is a class weight vector, where w_i represents the weight for the i -th class. w_i is calculated as

$\frac{1}{n_i}$. n_i represents the number of data of the i -th class. φ is a hyperparameter used to adjust the weight of hard and easy samples in Focal Loss, and its value is set to 2. For segmentation tasks, the pixel values of various segmentation classes are extremely imbalanced. For example, the number of pixels belonging to "LC" and "C-LCW" is much smaller than that of "lung

areas." This imbalance can lead to segmentation biases. To overcome this challenge, this study proposes a novel segmentation loss function based on Dice loss (Ming & Xiao, 2023). The segmentation loss function is defined as follows:

$$L_{\text{seg}}(P_{\text{seg}}, Y_{\text{seg}}) = 1 - \frac{2(1 - P_{\text{seg}})^{\sigma} \cdot P_{\text{seg}} \cdot Y_{\text{seg}}}{(1 - P_{\text{seg}})^{\sigma} \cdot P_{\text{seg}} + Y_{\text{seg}}} \quad (5.2)$$

In formula (5.2), this study introduces a tuning factor $(1 - P_{\text{seg}})^{\sigma}$ to reduce the weight of easily classified samples, thus better addressing the issue of sample imbalance. This means that in semantic segmentation tasks, the loss weight for background pixels (easily classified samples) is reduced, while the loss weight for edge or minority classes (difficult-to-classify samples) is increased, facilitating better focus on challenging pixels. The introduction of $(1 - P_{\text{seg}})^{\sigma}$ also enhances the robustness of the loss function to slight changes in predicted probabilities. Moreover, when the predicted probability approaches 1 or 0, $(1 - P_{\text{seg}})^{\sigma}$ becomes very small, thereby reducing the variation of the loss function. Finally, this study utilizes a multi-task loss with a progressive trade-off factor, which can be represented as follows:

$$L_{\text{joint}} = \lambda(n) L_{\text{cls}} + (1 - \lambda(n)) L_{\text{seg}} \quad (5.3)$$

In formula (5.3), n represents the training iterations, and $\lambda(n)$ is a progressive trade-off factor that changes as the training n increases. The $\lambda(n)$ is defined as:

$$\lambda(n) = \mu + \frac{\mu}{1 + e^{-\mu(n-\gamma)}} \quad (5.4)$$

In formula (5.4), μ is a parameter that controls the maximum and minimum values. At the upper limit of the value range of $\lambda(n)$, when $n \rightarrow \infty$, the value of $\lambda(n)$ will gradually tend to μ . That is, μ represents the maximum value of $\lambda(n)$. γ is a time offset parameter, which determines the offset of the $\lambda(n)$ function on the time axis. When $n = \gamma$, the value of $\lambda(n)$ reaches 2μ . In other words, γ determines the location of the midpoint of $\lambda(n)$, and its value is half of the total epoch. $e^{-\mu(n-\gamma)}$ is an S-shaped logistic function. When n is close to γ , $e^{-\mu(n-\gamma)}$ is close to 1, and when n is far away from γ , $e^{-\mu(n-\gamma)}$ is close to 0. This function maps values into the range $[0, \mu]$ in the form

$$\frac{\mu}{1 + e^{-\mu(n-\gamma)}}.$$

Clarification and Experimental Justification of the Novelty of the Proposed Segmentation Loss. The segmentation loss function presented in this study introduces a novel approach specifically designed to address the data imbalance often encountered in medical imaging tasks. By incorporating a modified version of the Sigmoid Focal Loss for classification and an adapted Dice-based loss for segmentation, the proposed method is tailored to handle both inter-class and intra-class imbalances effectively. For classification tasks, a modified Sigmoid Focal Loss is employed, incorporating class-specific weights to compensate for classes with fewer samples. This weighting

mechanism ensures the model does not disproportionately focus on classes with a larger sample size.

In segmentation tasks, the study innovates further by introducing a tuning factor, which reduces the impact of easily classified pixels (e.g., background) while increasing the focus on challenging areas, such as edges or minority classes. This is achieved through an adjustment that dynamically reduces the weight of high-confidence predictions, promoting robustness against slight variations in probabilities. Furthermore, the loss function includes a multi-task balancing mechanism with a progressive trade-off factor.

Table 5.1: Design Schemes of the Ablation Model

Methods	IA	IIA	Branch 1	Branch 2	Branch 3	Branch 4	Multi-scale	Task
SwinUNetR	×	×	×	×	×	×	×	SEG
SwinUNetR _{CLS}	×	×	×	×	×	×	×	CLS
DPTB _(w/o IIA, w MS-B4)	×	×	×	×	×	✓	×	SEG + CLS
DPTB _(w/o IIA, w IA & MS-B4)	✓	×	×	×	×	✓	×	SEG + CLS
DPTB _(w IIA & MS-B4)	×	✓	×	×	×	✓	×	SEG + CLS
DPTB _(w/o IIA, w MS)	×	×	✓	✓	✓	✓	✓	SEG + CLS
DPTB _(w/o IA, w MS & IA)	✓	×	✓	✓	✓	✓	✓	SEG + CLS
DPTBNet	×	✓	✓	✓	✓	✓	✓	SEG + CLS

Note: CLS and SEG stand for classification and segmentation task, respectively. IA and IIA represent the original iterative algorithm and the improved iterative algorithm respectively. w and w/o represent with and without respectively. MS stands for multi-scale classification architecture. Branches 1 to 4 represent the classification branches in DPTBNet, abbreviated as B1 to B4. MS-B4 represents the exclusive use of branch 4 for classification task, with the remaining branches 1 to 3 not being utilized. This rule is also followed in the subsequent table.

This unique setup not only addresses common imbalances but also provides stability in training across tasks, thereby enhancing the overall segmentation performance on challenging datasets. Quantitative experiments have been conducted to validate the effectiveness of these novel loss functions, with comparative analyses against traditional losses. These tests demonstrate the advantages of the proposed approach in terms of precision and robustness, particularly for underrepresented classes in segmentation tasks, thereby justifying the novelty of the method.

Addressing the Limitations of Existing Models for Pulmonary TB CT Analysis and the Need for the DPTBNet Architecture. Existing medical image analysis models often lack multi-scale feature fusion capabilities, which limits their effectiveness in analyzing pulmonary TB lesions. TB lesions exhibit a wide range of shapes and sizes, from small nodules to large cavities, reflecting the complex and varied presentation of the disease. Standard models without multi-scale architecture struggle to capture these variations comprehensively, which can lead to suboptimal detection and segmentation of TB lesions. The proposed DPTBNet addresses this gap by incorporating a multi-scale architecture specifically designed to handle the diverse morphological characteristics of TB lesions. By enabling feature fusion at different spatial resolutions, DPTBNet can effectively analyze both large and small lesions within the same model, enhancing its ability to accurately segment and classify TB pathology. This architectural innovation makes DPTBNet particularly suitable for pulmonary TB CT analysis, as it aligns with the inherent variability in TB lesion presentation.

5.4 Experiment Details

To assess the efficacy of the proposed approach, this study conducted a series of ablation experiments. Table 5.1 presents the differences among the ablation models. Specifically, this study trained separate single-task models SwinUnetR_{CLS} and SwinUnetR (Tang et al., 2022) as the classification and segmentation baseline models, respectively. For SwinUnetR_{CLS}, this study utilized the encoding path from SwinUnetR, followed by a GAP layer and two FC layers for the classification task (Zhou et al., 2021).

To explore MTL, this study employed the following strategies in the experiments. Firstly, this study added a single-scale classification branch to SwinUnetR, utilizing the feature map from Branch 4 as the classification feature, denoted as $DPTB_{(w/o \text{ IIA}, w \text{ MS-B4})}$. Subsequently, we trained $DPTB_{(w/o \text{ IIA}, w \text{ MS-B4})}$ using the improved iterative algorithm (IIA) with $n = 3$ iterations denoted as $DPTB_{(w \text{ IIA} \& \text{ MS-B4})}$. Additionally, for comparison with the original iterative algorithm (IA) (Zhou et al., 2021), this study trained $DPTB_{(w/o \text{ IIA}, w \text{ MS-B4})}$ with $n = 2$ iterations, denoted as $DPTB_{(w/o \text{ IIA}, w \text{ IA} \& \text{ MS-B4})}$.

Furthermore, we explored the use of the proposed multi-scale classification branch presented in Figure 5.1 instead of the single-scale classification branch for MTL, denoted as $DPTB_{(w/o \text{ IIA}, w \text{ MS})}$. We then trained the $DPTB_{(w/o \text{ IIA}, w \text{ MS})}$ model using the improved iterative feature-refining algorithm with $n = 3$ iterations, denoted as DPTBNet. For comparison, we

also trained $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS})}$ using the original iterative feature-refining algorithm (Zhou et al., 2021) with $n = 2$ iterations, denoted as $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS \& IA})}$.

Table 5.2: Performance of DPTBNet on the 95HD Metric with Respect to Weight Parameters and Number of Iterations

N	$\eta = 0.1 (\pm \text{SD})$	$\eta = 0.4 (\pm \text{SD})$	$\eta = 0.7 (\pm \text{SD})$	$\eta = 1 (\pm \text{SD})$
0	25.19 $\pm$ 4.67	25.43 $\pm$ 8.12	22.62 $\pm$ 3.64	24.13 $\pm$ 6.67
1	22.28 $\pm$ 5.89	19.69 $\pm$ 4.14	17.13 $\pm$ 7.86	17.21 $\pm$ 5.53
2	21.79 $\pm$ 3.65	18.43 $\pm$ 5.42	16.04 $\pm$ 6.32	16.82 $\pm$ 4.78
3	18.92 $\pm$ 3.91	15.96 $\pm$ 4.83	14.24 $\pm$ 5.38	14.53 $\pm$ 5.89
4	20.61 $\pm$ 6.57	17.02 $\pm$ 3.67	14.98 $\pm$ 4.16	15.16 $\pm$ 3.78

Table 5.3: IoU Performance of DPTBNet Across Different Weight Parameters and Number of Iterations

N	$\eta = 0.1 (\pm \text{SD})$	$\eta = 0.4 (\pm \text{SD})$	$\eta = 0.7 (\pm \text{SD})$	$\eta = 1 (\pm \text{SD})$
0	39.5 $\pm$ 10.5	40.4 $\pm$ 14.2	46.4 $\pm$ 7.8	43.8 $\pm$ 10.5
1	46.9 $\pm$ 12.4	48.1 $\pm$ 9.6	51.6 $\pm$ 12.6	50.1 $\pm$ 16.2
2	47.2 $\pm$ 8.9	50.4 $\pm$ 12.9	52.6 $\pm$ 13.5	53.4 $\pm$ 10.2
3	49.4 $\pm$ 9.6	53.6 $\pm$ 11.9	55.8 $\pm$ 13.9	55.3 $\pm$ 9.2
4	48.7 $\pm$ 12.8	54.6 $\pm$ 10.7	55.8 $\pm$ 13.9	54.2 $\pm$ 6.4

To assess the effectiveness of the proposed method, this study conducted a comparative study with existing MTL approaches, including the methods proposed by Alebiosu et al. (2023), Zhou et al. (2021), Song et al. (2020), and Gordaliza et al. (2019). As some of these methods were originally designed

for 2D operations, this study extended them to 3D operations to ensure a fair comparison. To maintain consistency, this study replaced the backbone networks used in these methods with the same SwinUnetR backbone utilized in the proposed method. For the evaluation, this study performed five-fold cross-validation using the DeepPulmoTB dataset and employed the proposed loss function from Formula 5.3 in the experiments. The other parameters of the compared methods were kept consistent with the values reported in the respective works by Alebiosu et al. (2023), Zhou et al. (2021), Song et al. (2020), and Gordaliza et al. (2019).

5.5 Result and Discussion

5.5.1 Pulmonary Tuberculosis Lesions and Lung Areas Segmentation

To investigate the optimal performance of the improved iterative feature refinement algorithm under different strategies, this study trained the DPTBNet with varying numbers of iterations (n) and weight parameters (η) and summarized the results in Tables 5.2 – 5.4. From these tables, we observed that the iterative training strategy shows the best performance when $\eta = 0.7$ and $n = 3$, achieving 95HD, DSC, IoU, and values of 14.24, 69.6% and 55.8%, respectively. This outcome suggests that the model effectively uses iterative feature refinement by combining spatial information from the training volume and semantic information from the segmentation map, leading to more accurate pixel-level segmentation. The comparative analysis of different η values highlighted that tuning the weight parameter η to 0.7 in the IIA can

achieve a superior equilibrium between the original training volume and the segmentation probability map.

This, in turn, leads to improved algorithmic performance when contrasted with scenarios where this parameter remains unadjusted ($\eta = 1$). As a result, for all ensuing experiments, we have opted to consistently set $\eta = 0.7$ and $n = 3$

Table 5.4: DSC Performance of DPTBNet Across Different Weight Parameters and Number of Iterations

N	$\eta = 0.1 (\pm \text{SD})$	$\eta = 0.4 (\pm \text{SD})$	$\eta = 0.7 (\pm \text{SD})$	$\eta = 1 (\pm \text{SD})$
0	53.9 ± 11.8	53.3 ± 15.6	59.3 ± 8.4	56.9 ± 12.6
1	59.9 ± 12.3	61.9 ± 10.8	65.9 ± 12.4	65.1 ± 12.9
2	60.4 ± 9.1	63.4 ± 13.7	67.4 ± 15.4	66.5 ± 14.3
3	62.2 ± 11.7	66.2 ± 10.4	69.6 ± 15.2	68.2 ± 12.8
4	61.9 ± 14.3	65.9 ± 10.5	68.1 ± 14.2	67.3 ± 7.1

Table 5.5 presents the quantitative results of ablative experiments based on five-fold cross-validation. For the segmentation task, we observed that the DPTBNet model outperforms other methods in all metrics, with 95HD, DSC, and IoU values of 14.24, 69.6%, and 55.8%, respectively. By comparing the results of $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS-B4})}$ and SwinUnetR, we observe that $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS-B4})}$ performs better on the segmentation task, confirming the beneficial impact of incorporating classification information into the segmentation task. In contrast to $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS-B4})}$, $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS})}$ adopts multi-scale feature extraction for classification and exhibits better

performance than $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS-B4})}$ and SwinUnetR on the segmentation task.

Furthermore, the improved iterative algorithm enhances the segmentation performance compared to the original iterative algorithm. For instance, DPTBNet and $\text{DPTB}_{(w \text{ IIA} \& \text{ MS-B4})}$ achieve DSC improvements of 7.3% and 7.8%, respectively, compared to $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS})}$ and $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS-B4})}$.

Moreover, both the improved and original iterative feature refinement strategies enhance the segmentation performance. For example, DPTBNet and $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS} \& \text{ IA})}$ achieve DSC improvements of 4.6% and 2.7%, respectively, compared to $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS})}$.

Figure 5.3 displays the three cross-sectional views of pulmonary TB lesions and normal lung areas. From the figure, we observed that the results obtained from the method proposed in this research align well with ground truth in three dimensions compared to the other methods.

5.5.2 Classification of Lung Cavity Attributes

Figure 5.4 demonstrates a comparative analysis of the focus areas between a single-task classification network and a MTL network. We observed that compared to $\text{SwinUnetR}_{\text{CLS}}$, $\text{DPTB}_{(w/o \text{ IIA}, w \text{ MS})}$, which integrates multi-scale segmentation feature maps, is more effective in capturing the pulmonary TB lesion feature regions.

Table 5.5: Segmentation and Classification Task Performance of Ablation Models

Methods	Segmentation ($\pm$ SD)					Classification				
	95HD (voxel)	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1		
SwinUNetR	23.06 $\pm$ 7.27	46.9 $\pm$ 12.7	58.9 $\pm$ 11.9	-	-	-	-	-		
SwinUNetR _{CLS}	-	-	-	0.578	0.697	0.667	0.698	0.681		
DPTB _(w/o IA, w MS & IA)	15.87 $\pm$ 4.57	53.7 $\pm$ 12.6	66.9 $\pm$ 14.1	0.738	0.772	0.812	0.386	0.791		
DPTBNet	14.24 $\pm$ 5.38	55.8 $\pm$ 13.9	69.6 $\pm$ 15.2	0.757	0.804	0.817	0.408	0.810		

Table 5.6: Evaluation of Classification Performance Using Different Multi-Scale Features

Methods	IA	IIA	Branch 1	Branch 2	Branch 3	Branch 4	Multi-Scale	ACC	REC	PRE	FPR	F1
DPTB-B1	X	X	✓	X	X	X	X	0.663	0.756	0.716	0.583	0.735
DPTB-B1-2	X	X	✓	✓	X	X	✓	0.697	0.773	0.738	0.526	0.755
DPTB-B1-3	X	X	✓	✓	✓	X	✓	0.708	0.829	0.754	0.492	0.790
DPTB-B1-4	X	X	✓	✓	✓	✓	✓	0.711	0.823	0.773	0.475	0.797

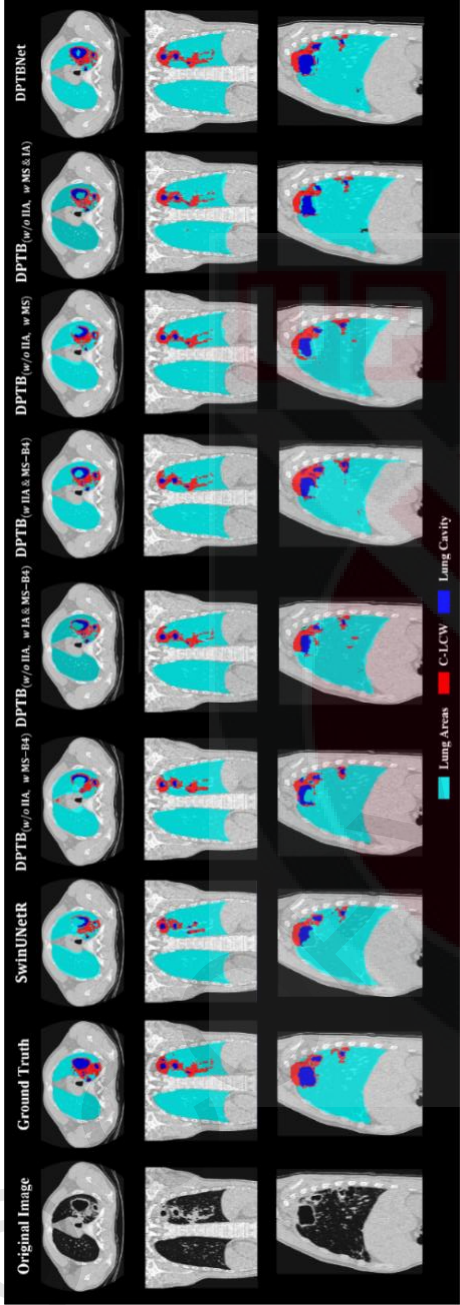


Figure 5.3: Segmentation Masks Predicted by Various Ablation Models

Table 5.7: The Effects of Different Values of σ on Multi-Task Learning

σ	Segmentation ($\pm$ SD)				Classification				
	95HD (voxel)	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1	
0	14.53 $\pm$ 4.89	54.6 $\pm$ 13.2	68.3 $\pm$ 12.3	0.739	0.785	0.820	0.462	0.802	
2	14.24 $\pm$ 5.38	55.8 $\pm$ 13.9	69.6 $\pm$ 15.2	0.757	0.804	0.817	0.408	0.810	
4	24.08 $\pm$ 4.95	45.2 $\pm$ 12.3	57.8 $\pm$ 14.5	0.641	0.762	0.732	0.617	0.746	

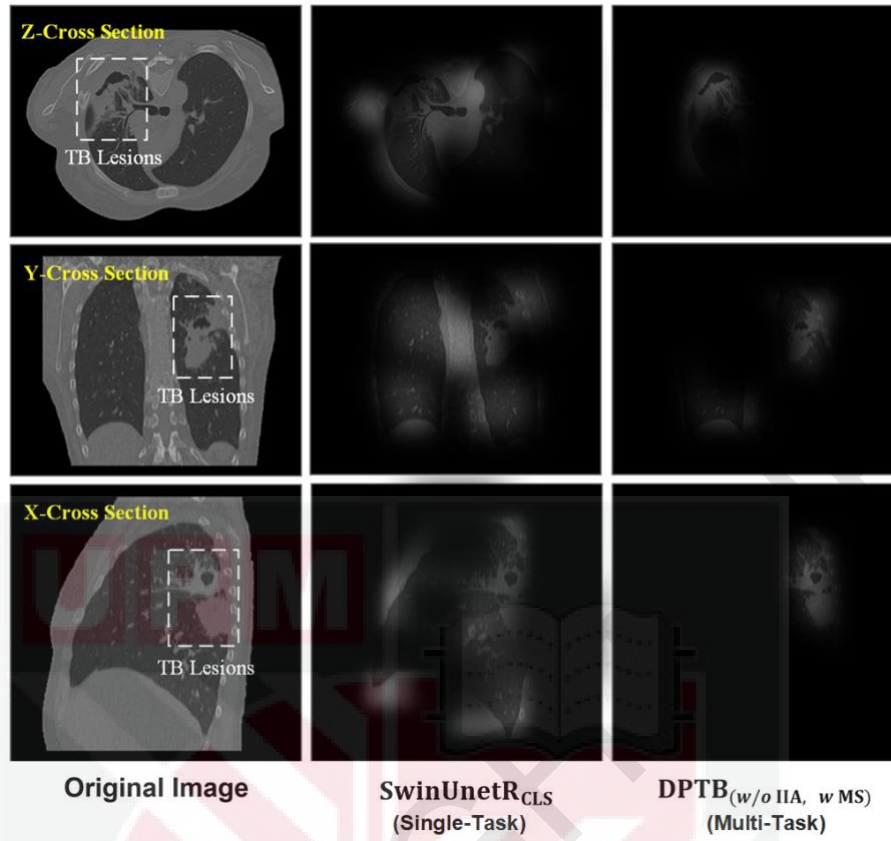


Figure 5.4: Comparison of the Focus Areas Between Single-Task Learning and Multi-Task Learning Networks for Classification Tasks

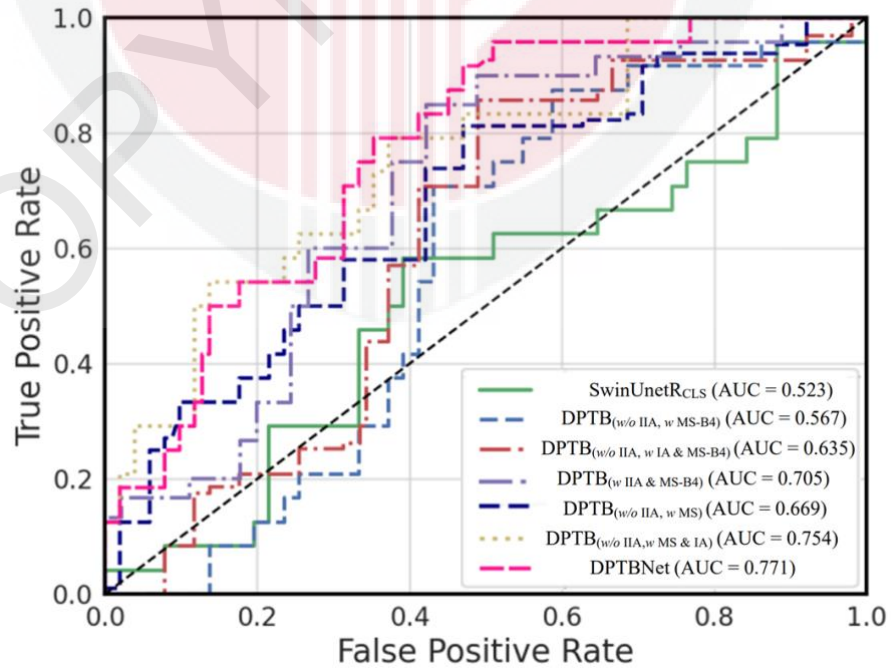


Figure 5.5: ROCs of Ablation Models

Table 5.5 also summarizes the quantitative classification results of single-classification and MTL models. $DPTB_{(w/o\ IIA,\ w\ MS-B4)}$ employs single-scale feature extraction for classification and demonstrates an improvement in classification results compared to the classification baseline model, SwinUnetR_{CLS}, confirming the enhancement brought by segmentation to classification. Moreover, from the table, we observed that $DPTB_{(w/o\ IIA,\ w\ MS)}$ achieved the best REC value of 0.823, while DPTBNet attained optimal values of 0.757, 0.817, 0.408, and 0.810 for ACC, PRE, FPR, and F1 scores, respectively. Figure 5.5 displays the ROC curves for each ablation model. In comparison to $DPTB_{(w\ IIA\ \&\ w\ MS-B4)}$, DPTBNet exhibited superior performance by elevating the AUC from 0.705 to 0.771. Moreover, we observed that the AUC values of the proposed improved iterative algorithm are better than the original iterative algorithm, whether in a single-scale or multi-scale network structure.

To further compare the impact of feature map combinations from different scales on the classification task, this study trained DPTBNet with various feature combinations, as shown in Table 5.6. DPTB-Bi-j connected the feature maps from the i -th to j -th branches of DPTBNet. For instance, DPTB-B1-3 utilized the feature maps from branches 1, 2, and 3. From the table, we observed that the fusion of features from branches 1 to 4 yielded the optimal values of 0.711, 0.773, 0.475, and 0.797 for ACC, PRE, FPR, and F1 scores, respectively.

Table 5.8: The Effects of Different Values of μ on Multi-Task Learning

μ	Segmentation ($\pm$ SD)				Classification			
	95HD (voxel)	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1
0.3	14.24 $\pm$ 5.38	55.8 $\pm$ 13.9	69.6 $\pm$ 15.2	0.757	0.804	0.817	0.408	0.810
0.5	15.03 $\pm$ 4.78	54.7 $\pm$ 11.5	68.7 $\pm$ 12.7	0.723	0.831	0.789	0.492	0.809

Table 5.9: Comparison with Existing Multi-Task Learning Methods

Methods	Segmentation ($\pm$ SD)			Classification					
	95HD (voxel)	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1	
DeepPulmoTB	Gordaliza et al. (2019)	29.36 $\pm$ 7.64	38.8 $\pm$ 15.8	52.4 $\pm$ 16.7	0.675	0.717	0.732	0.672	0.724
	Alebiosu et al. (2023)	18.76 $\pm$ 5.57	51.4 $\pm$ 12.5	64.1 $\pm$ 13.6	0.728	0.763	0.858	0.463	0.807
	DPTBNet	14.24 $\pm$ 5.38	55.8 $\pm$ 13.9	69.6 $\pm$ 15.2	0.757	0.804	0.817	0.408	0.810
DeepCervix	Gordaliza et al. (2019)	7.12 $\pm$ 1.12	65.6 $\pm$ 11.4	73.1 $\pm$ 11.3	0.821	0.853	0.791	0.211	0.822
	Alebiosu et al. (2023)	6.34 $\pm$ 0.93	72.9 $\pm$ 9.7	78.8 $\pm$ 9.8	0.873	0.895	0.841	0.148	0.859
	DPTBNet	6.08 $\pm$ 0.89	74.6 $\pm$ 9.4	80.3 $\pm$ 9.5	0.876	0.903	0.823	0.136	0.862

Class-Wise Performance Results based on cavity size: To provide a more detailed analysis, the model's performance is presented class-wise based on cavity sizes, as outlined in Table 4.1. The results below highlight the model's accuracy, precision, recall, and F1-score across different classes, giving a more granular view of its effectiveness in detecting various LC characteristics.

- Class 1 (No Cavities). Accuracy: 84.0%; Precision: 87.0%; Recall: 89.0%; F1-Score: 88.0%. The model performs exceptionally well in identifying cases with no cavities, reflecting a high level of accuracy and reliability in distinguishing non-cavity scans from other classes.
- Class 2 (1-3 Cavities, < 25mm). Accuracy: 78.0%; Precision: 82.0%; Recall: 85.0%; F1-Score: 83.5%. For small-sized cavities (1-3, <25mm), the model demonstrates robust performance, accurately identifying smaller lesions with high precision and recall.
- Class 3 (1-3 Cavities, > 25mm). Accuracy: 76.0%; Precision: 80.0%; Recall: 82.0%; F1-Score: 81.0%. The model's performance slightly decreases for larger cavities (1-3, >25mm), as expected due to the complexity introduced by the size variation. However, it still shows reliable detection capabilities in this class.
- Class 4 (> 3 Cavities, < 25mm). Accuracy: 72.0%; Precision: 78.0%; Recall: 80.0%; F1-Score: 79.0%. In cases with multiple small cavities (>3, <25mm), the model maintains a solid performance, though with a

slight reduction in accuracy due to the increased difficulty of detecting multiple small lesions.

- Class 5 (> 3 Cavities, > 25mm). Accuracy: 74.5%; Precision: 76.0%; Recall: 78.0%; F1-Score: 77.0%. For cases with multiple large cavities (>3, >25mm), the model achieves a lower accuracy and precision, reflecting the challenges associated with identifying extensive, overlapping lesions. Despite these challenges, the model remains effective in handling complex cases with larger lesion sizes.

Overall, these class-wise results provide a comprehensive understanding of the model's performance across different lesion types and sizes. The model performs best in classes with no cavities or fewer small lesions, while performance slightly declines in classes with larger or multiple lesions.

Justification for the Use of Deep Models in Cavity Count and Lesion Localization Tasks. The use of deep learning models for tasks like cavity count and lesion location in pulmonary CT images, although seemingly straightforward, is justified by the complexity and variability inherent in real-world medical images. While counting cavities and identifying lesion locations might appear as simple tasks, the diverse manifestations of pulmonary lesions, variability in image quality, and subtle differences in pathological regions can introduce substantial challenges that are difficult to address with traditional methods or shallow models. Deep models excel in capturing nuanced spatial patterns and subtle textural differences, making them well-suited for consistently detecting and localizing complex lesions across a range

of cases. Moreover, deep learning architectures allow for the integration of multi-scale features, enabling the model to detect lesions at different resolutions and scales, which is particularly beneficial for comprehensive and reliable segmentation and classification in complex medical images. The complexity of lung cavities, which can vary widely in size, shape, and intensity, also demands a model with robust generalization capabilities. By utilizing a deep learning approach, the system can better handle these variations, reducing the likelihood of misidentification and improving overall diagnostic accuracy. In addition, employing a deep model allows for scalability and flexibility, as such models can be easily adapted to include more advanced tasks beyond counting and localization if necessary. The potential to integrate other diagnostic features or related tasks makes deep models a valuable choice for building comprehensive systems that address the multifaceted needs of medical image analysis, thereby supporting clinical decision-making more effectively.

5.5.3 Multi-Task Learning

The purpose of the hyperparameter σ in Formula 5.2 is to adjust the weighting of the overlap between predicted results and true labels in the loss function, thus influencing the degree of adjustment to model predictions during the training process. In the proposed DPTBNet model, the performance of MTL based on different σ values is listed in Table 5.7. We observed that when σ is set to 2, DPTBNet achieves the best performance, and it performs the best in terms of 95HD, IoU, DSC, ACC, FPR, and F1 scores. Although the classification REC and PRE performance reach their best values of 0.813 and

0.820 when $\sigma = 3$ and $\sigma = 0$, the segmentation performance is worse than that of $\sigma = 2$. Among them, $\sigma = 0$ represents the original dice loss.

The purpose of Formula 5.3 and 5.4 is to balance the performance of classification and semantic segmentation in MTL, where the hyperparameter μ is used to control the maximum and minimum values. In the proposed DPTBNet model, the performance of classification and semantic segmentation based on different μ values is listed in Table 5.8. From the table, we observed that when μ is set to 0.3, DPTBNet achieves the best balance between classification loss and semantic segmentation loss, and it performs the best in terms of 95HD, IoU, DSC, ACC, PRE, FPR, and F1 scores.

In the experiments, we set $\sigma = 0.8$ and $\mu = 0.3$. The proposed method was compared with other MTL approaches, with results summarized in Table 5.9. To reduce potential biases from training on a single dataset, four existing MTL methods were trained on both the DeepPulmoTB and DeepCervix (Yang et al., 2024) datasets.

As shown in Table 5.9, the proposed DPTBNet model outperforms all other methods across all metrics except for precision (PRE), where it achieves the second-highest score among the methods tested.

Explanation of Performance Differences Between DPTBNet and Other Techniques, Focusing on Instances Where Other Methods Show Superior Results. The differences in results between DPTBNet and other techniques, as observed in Table 5.9, can be attributed to the unique design

and objectives of each model, as well as the specialized optimizations used by DPTBNet to enhance its performance specifically for TB lesion detection and segmentation:

1. Gordaliza et al. (2019): This model utilizes radiologist reports as ground truth instead of manual segmentation masks, which may lead to less precise spatial information for training. Additionally, it is based on the V-Net model and uses a novel loss function incorporating model uncertainty for multi-task learning. While effective for broader TB manifestations, the approach is less focused on fine-grained spatial details, which could explain its lower performance in segmentation metrics (e.g., IoU and DSC) compared to DPTBNet.

2. Song et al. (2020): Song et al. focuses on skin lesion analysis, with adaptations for dealing with imbalanced datasets and improving segmentation through a combination of focal loss and Jaccard distance. The model's three-phase joint training strategy is tailored for skin lesion segmentation rather than pulmonary applications, potentially limiting its efficacy on lung CT images. This difference in domain focus may explain why it underperforms in TB-related tasks relative to DPTBNet, which is specifically designed to capture TB lesion characteristics.

3. Y. Zhou et al. (2021): Zhou et al. propose a multi-task learning framework for tumor classification and segmentation in breast ultrasound images, utilizing iterative training to refine feature maps. The model is optimized for ultrasound images, which have different imaging characteristics compared to CT scans

of the lungs. This difference in imaging modality and target task could contribute to its lower segmentation and classification scores on TB CT data, as DPTBNet is specifically tuned for the spatial and textural properties of lung CT images.

4. Alebiosu et al. (2023): Alebiosu et al. introduced DAvoU-Net with Multi-Scale Residual Blocks and a bi-directional LSTM for segmenting TB-affected areas in CT images. While effective, this model is complex and designed to generalize across various CT datasets, not specifically optimized for multi-task learning in TB lesion segmentation and classification. In contrast, DPTBNet's integration of SwinUnetR with a multi-scale classification network and iterative optimization gives it an edge in capturing TB-specific features, explaining its superior performance in most metrics except precision (PRE).

5. DPTBNet: DPTBNet leverages an optimized architecture combining SwinUnetR for segmentation and a lightweight multi-scale network for classification. The model's unique iterative optimization algorithm refines feature maps by integrating probability maps from previous iterations, enhancing its capacity to focus on TB lesion boundaries and fine details. These optimizations are particularly effective for the DeepPulmoTB and DeepCervix datasets, resulting in improved performance across most metrics. The slight underperformance in precision compared to some methods may be due to DPTBNet's focus on maximizing recall and segmentation accuracy, which are critical for identifying TB lesions comprehensively.

Overall, DPTBNet's tailored architecture, with its specific focus on TB lesion features and iterative optimization, enables it to outperform other models in most metrics. However, certain methods may achieve higher precision due to their focus on specific structures or simpler tasks, while DPTBNet is designed to handle the complex multi-task requirements of TB lesion detection and segmentation.

5.6 Summary

To explore the effectiveness of MTL technology in CAD of pulmonary TB, this chapter designs a novel MTL model (DPTBNet). This model can simultaneously handle two tasks: semantic segmentation of pulmonary TB lesion tissues and attribute classification. Additionally, to further enhance the model's capability to capture pulmonary TB lesion features, this chapter also proposes an improved iterative optimization algorithm. This algorithm intricately refines feature maps by integrating the probability maps from previous iterations. Studies show that by providing a variety of interactively learnable imaging features, DeepPulmoTB offers a richer and more comprehensive lesion representation for the DPTBNet model, improving the accuracy and robustness of CAD. Moreover, DPTBNet provides better experimental results in pulmonary TB segmentation and pulmonary attribute classification than the STL model and achieves state-of-the-art performance among mainstream MTL methods used for medical image analysis.

CHAPTER 6

LUNG CAVITY WEAKLY SUPERVISED SEMANTIC SEGMENTATION FRAMEWORK

6.1 Introduction

Radiological diagnosis of LC is crucial for identifying pulmonary TB. In clinical applications, semantic segmentation techniques based on the fully supervised learning paradigm rely on large amounts of high-quality pixel-level LC segmentation ground truth data for training. However, obtaining precise pixel-level segmentation ground truth annotations for pulmonary TB CT data is an extremely time-consuming and costly task. This is because this process necessitates experienced medical experts to annotate these three-dimensional CT data on a per-pixel, per-layer basis. Furthermore, pixel-level annotation data may have significant subjective differences in the annotation process among different experts.

WSSS technology is a common solution to the problems mentioned above. This technology aims to use inexpensive coarse-grained or incomplete LC annotation information to train models for efficient acquisition of LC pixel annotation data. For example, using WSSS technology, radiologists only need to classify the number and approximate locations of LC in a CT scan (e.g., left lung, right lung, or both) to achieve efficient segmentation of LC. This can not only reduce the workload of medical experts in data annotation but also lower the cost and complexity of data preparation. Especially in resource-limited medical settings, WSSS technology can help healthcare providers overcome challenges in data acquisition and processing, thus facilitating more effective

pulmonary TB diagnosis. Therefore, to reduce the reliance on pixel-level annotations of pulmonary TB LC and improve LC feature acquisition efficiency, this section proposed the following two main objectives centered around WSSS techniques:

- i) The second objective is to propose a novel medical 3D imaging analysis model called SwinUNeLCsT to address the limitations of traditional transformer architectures in capturing fine-grained structural features of LC.
- ii) The third objective is to design a WSSS framework. This framework utilizes multi-class labels from CT images to generate pixel-level weakly supervised regions, which are then refined to produce pseudo-semantic segmentation labels for pulmonary TB lesions. Based on this LC WSSS framework, even with coarse-grained LC annotations, we can assist physicians in identifying and segmenting pulmonary TB lesions, providing information that aids clinical decision-making.

6.2 Description of the Proposed Hybrid CNN-Transformer Network

As discussed in section 3.2.4.2, this study introduces a novel CNN-Transformer hybrid model, SwinUNeLCsT, addressing Objective 2. Specifically, this model combines the advantages of CNN and Transformer architecture. It uses CNNs to capture local detail features and Transformers to model global dependencies, thereby enhancing the overall feature representation of TB CT scan images.

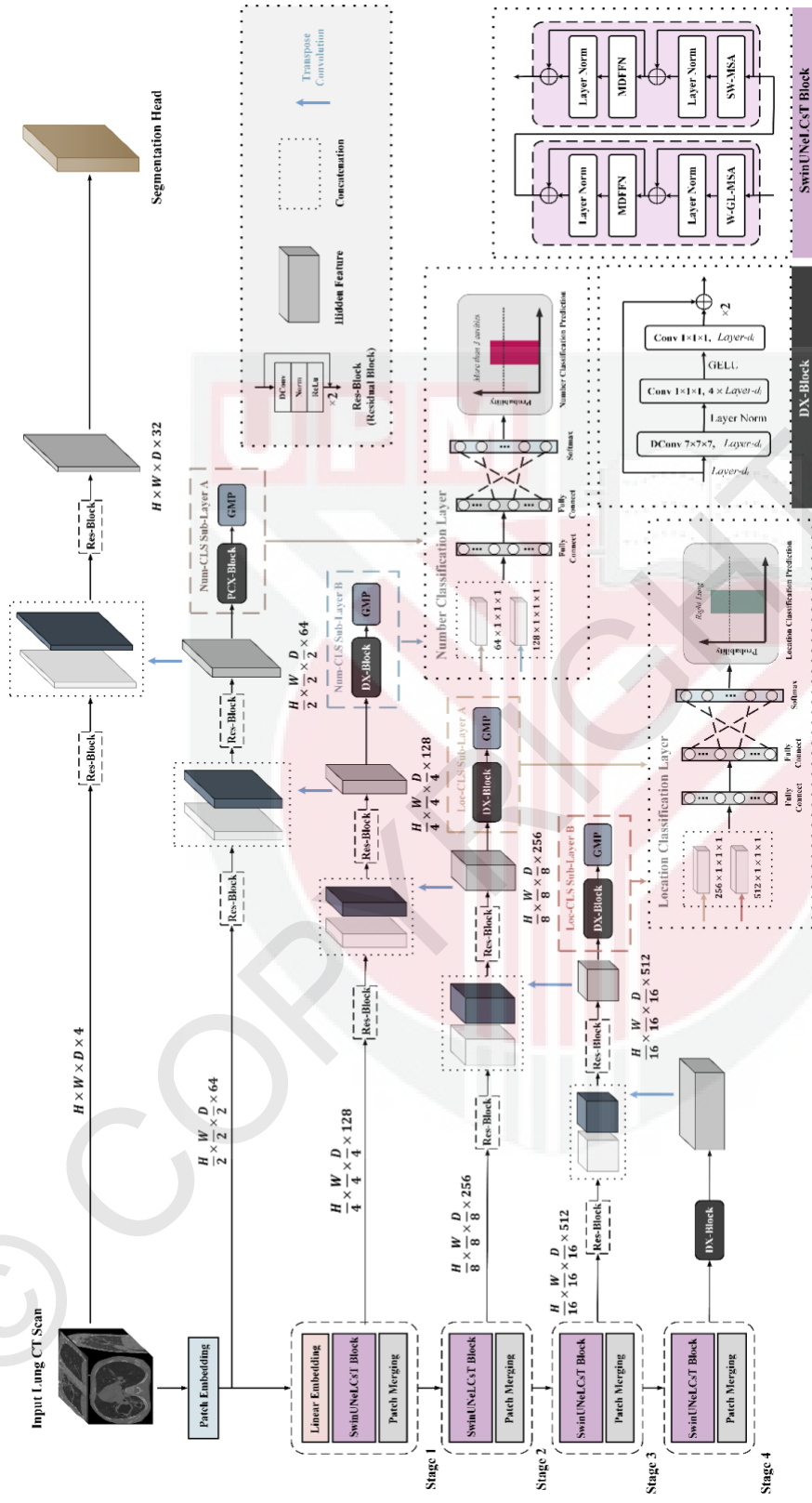


Figure 6.1: Overview of the SwinUNetCST Architecture

6.2.1 SwinUNeLCsT Encoder

The overall SwinUNeLCsT architecture is shown in Figure 6.1. Denote the input pulmonary TB CT scan image as a sub-volume $X \in \mathbb{R}^{H \times W \times D \times S}$ 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 represents the dimensionality of the embedding space. Following (Tang et al., 2022), 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:

$$\begin{cases} \hat{z}^l = \text{W-GL-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\ z^l = \text{MDFFN}(\text{LN}(\hat{z}^l)) + \hat{z}^l \\ \hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l \\ z^{l+1} = \text{MDFFN}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{cases} \quad (6.1)$$

In formula (6.1), 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 $\hat{z}^{l+1}$ are the outputs of W-GL-MSA and SW-MSA; LN and MDFFN denote layer normalization and multi-DConv feed forward network, respectively. Overall, the encoder uses a patch size of $2 \times 2 \times 2$. Each patch is mapped to an 8-dimensional feature space (as $2 \times 2 \times 2 \times 1 = 8$), where

1 represents that the image has only one input channel. The initial embedding space dimension is $C=64$ dimensions. The 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 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 $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$.

6.2.2 Multi-DConv Feed-Forward Network

In traditional swin transformers (Liu et al., 2021), the MLP (Multilayer Perceptron) usually consists of a series of FC 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, this study proposes the

MDFFN (multi-DConv feed-forward network), as shown in Figure 6.2 (a). MDFFN employs depth-wise separable convolutional layers, $DConv_{3 \times 3 \times 3}$, 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 X_l is the input to the l -th layer of MDFFN, the operations in one branch of the network would be:

$$Y_{l,DConv} = GELU\left(DConv_{3 \times 3 \times 3}\left(GELU\left(Conv_{1 \times 1 \times 1}(X_l)\right)\right)\right) \quad (6.2)$$

$$Y_{l,conv} = Conv_{1 \times 1 \times 1}(Y_{l,DConv}) \quad (6.3)$$

In formula (6.2), GELU represents Gaussian Error Linear Unit (Kang et al., 2022). And then the output 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:

$$Y_l = Y_{l,conv}^{(1)} \square Y_{l,conv}^{(2)} \oplus X_l \quad (6.4)$$

6.2.3 Window-Based Global-Local Multi-Head Self-Attention

In complex LC 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 pulmonary TB CT scans.

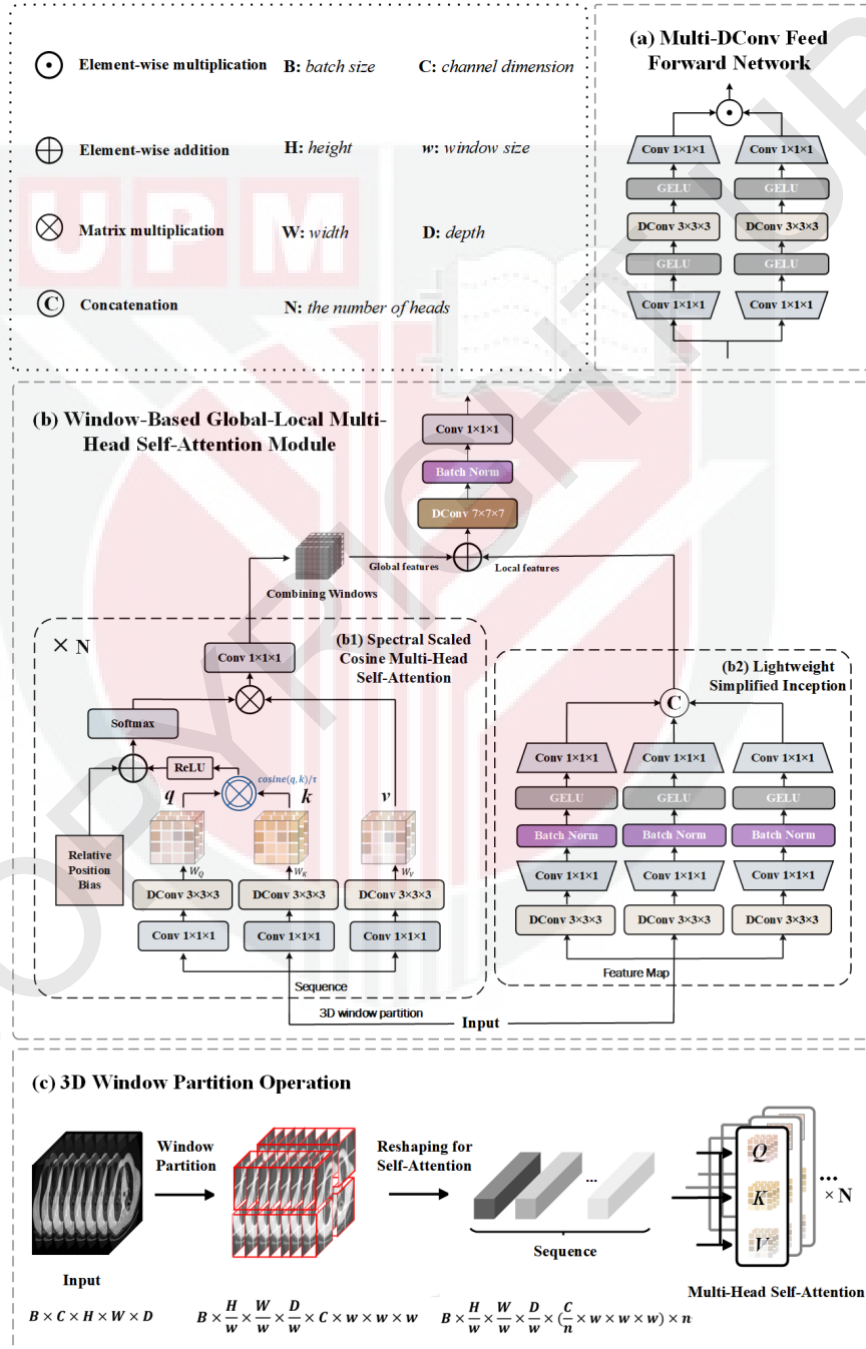


Figure 6.2: Components of the SwinUNeLCst Block

Therefore, to fully leverage the advantages of both architectures, this study develops a W-GL-MSA (window-based global-local multi-head self-attention), as shown in Figure 6.2 (b). W-GL-MSA contains two parallel branches: SSC-MSA (spectral scaled cosine multi-head self-attention) and LSI (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:

$$O_{\text{W-GL-MSA}} = \text{Conv}_{1 \times 1 \times 1} \left(\text{BN} \left(\text{DConv}_{7 \times 7 \times 7} (I + X) \right) \right) \quad (6.5)$$

In formula (6.5), I is the output of the lightweight simplified inception module, and X is the output processed by the window-based spectral scaled cosine self-attention module.

6.2.3.1 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 Figure 6.2 (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 $B \times \frac{H}{w} \times \frac{W}{w} \times \frac{D}{w} \times \left(\frac{C}{h} \times w \times w \times w \right)$ 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.

6.2.3.2 Window-based Spectral Scaled Cosine Self-Attention

The self-attention mechanism in transformer models is key for capturing global image features, but it's also a major source of computational overhead. In the field of medical imaging, the standard self-attention mechanism (Huang et al., 2024; Liu et al., 2021) 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 $O(W^2 H^2 D^2)$ 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 (Dong et al., 2023; Liu et al., 2022; Wang et al., 2022; Zamir et al., 2022), this study proposes a novel SSC-MSA (spectral scaled cosine multi-head self-attention) with linear complexity, as illustrated in Figure 6.3 (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 $Q = W_d^Q W_p^Q X$, the key vector $K = W_d^K W_p^K X$, and the value vector $V = W_d^V W_p^V X$. 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 Q and 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:

$$X = W_p \cdot BN \cdot W_{Dc7} \cdot \text{Attention}(Q, K, V) \quad (6.6)$$

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{\text{ReLU}(\cosine(Q, K))}{\tau} \right) V \quad (6.7)$$

In formula (6.6) and (6.7), $Q \in \mathbb{R}^{HWD \times C}$, $K \in \mathbb{R}^{C \times HWD}$, and $V \in \mathbb{R}^{HWD \times C}$ are matrices obtained after reshaping tensors from their original size $\mathbb{R}^{H \times W \times D}$.

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.

6.2.3.3 Lightweight Simplified Inception

The LSI (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 (Yao et al., 2021), 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 Figure 6.3 (b2). Let F be the feature map input to the module. The module's output I can be represented as:

$$O_i = Conv_{1 \times 1 \times 1} \left(ReLU \left(BN \left(Conv_{1 \times 1 \times 1} \left(DConv_{3 \times 3 \times 3} (F_i) \right) \right) \right) \right) \quad (6.8)$$

$$I = \text{Concat} (O_1, O_2, O_3) \quad (6.9)$$

6.2.4 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 the 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) 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 pulmonary TB CT medical images.

6.2.4.1 DX-block

To learn richer feature representations in pulmonary TB CT imaging, we develop a novel large-kernel depth-wise separable convolution module (as shown in Figure 6.2), inspired by the ConvNeXt architecture (Zhuang Liu et al., 2022), 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 X with a dimension of layer- d_i , where i represents a specific layer in the network. Then, the output of a basic DX-block can be expressed as:

$$\bar{X} = \text{Conv}_{1 \times 1 \times 1} \left(\text{GELU} \left(\text{Conv}_{1 \times 1 \times 1} \left(\text{LN} \left(\text{DConv}_{7 \times 7 \times 7} (X) \right) \right) \right) \right) \quad (6.10)$$

$$O_{\text{DX-block}} = X + \bar{X} \quad (6.11)$$

6.3 Description of the Lung Cavity Weakly Supervised Semantic Segmentation Framework

As mentioned in section 3.2.4.1, to reduce the dependence on accurate LC pixel-level annotations, this study constructs a WSSS framework for LC, addressing Objective 3. This framework first uses the SwinUNeLCsT model to jointly train tasks for classifying the number and location of LC, to obtain CAM for LC from these two classification layers. Then, to better represent LC features, we propose a technique called CFFG (CAM feature fusion gate), which merges the CAM of these two layers, producing CAM that simultaneously possesses regions of interest for both the location and number

of LC. Given the CAM $M \in \mathbb{R}^{h \times w \times d \times C}$, the overall CAM $M(i, j, s)$ is defined as follows:

$$M(i, j, s, :) = \text{CFFG}(M^{\text{num}}(i, j, s, :), M^{\text{loc}}(i, j, s, :)) \quad (6.12)$$

In formula (6.12), $M(i, j, s, :)$ denotes the amalgamated CAM that fuses both number classification layer's CAM ($M^{\text{num}}(i, j, s, :)$) and location classification layer's CAM ($M^{\text{loc}}(i, j, s, :)$). Moreover, CAM for each classification layer is formed by merging the CAM of two sub-layers through CFFG:

$$M^{\text{num}}(i, j, s, :) = \text{CFFG}(M^{\text{num}}_{\text{layer-A}}(i, j, s, :), M^{\text{num}}_{\text{layer-B}}(i, j, s, :)) \quad (6.13)$$

$$M^{\text{loc}}(i, j, s, :) = \text{CFFG}(M^{\text{loc}}_{\text{layer-A}}(i, j, s, :), M^{\text{loc}}_{\text{layer-B}}(i, j, s, :)) \quad (6.14)$$

In formula (6.13) and (6.14), $M^{\text{num}}_{\text{layer-A}}$ and $M^{\text{num}}_{\text{layer-B}}$ denote number classification sub-layers A and B. $M^{\text{loc}}_{\text{layer-A}}$ and $M^{\text{loc}}_{\text{layer-B}}$ denote location classification sub-layers A and B.

Subsequently, through a series of steps detailed in section 5.4.2, we refine these regions of interest and generate semantic segmentation pseudo-annotations for LC. Finally, we use these pseudo-annotations for WSSS of LC, as shown in Figure 6.3.

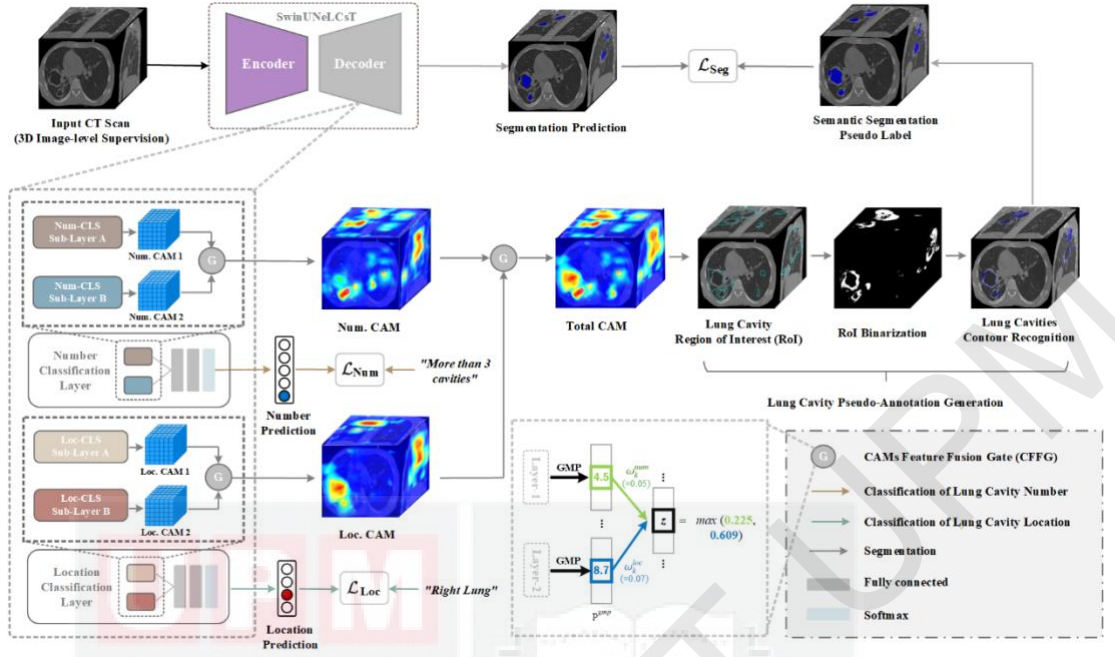


Figure 6.3: Overview of the Proposed Lung Cavity Weakly Supervised Semantic Segmentation Framework

6.3.1 CFFG

The CFFG (CAM feature fusion gate) 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 M^c represents the contribution of the given class c to the heatmap. To highlight regions of interest in the heatmap M^c and suppress irrelevant areas, this study applies a ReLU operation to M^c . The overall computation is expressed as follows:

$$M^c = ReLU\left(\sum_k \omega_k^c \cdot A_k\right) \quad (6.15)$$

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

$$M(i, j, s, :) = \max\left(M_{\text{layer-1}}^c(i, j, s, :), M_{\text{layer-2}}^{\hat{c}}(i, j, s, :)\right) \quad (6.16)$$

In formula (6.16), the categories of classification c and $\hat{c}$ in $M_{\text{layer-1}}^c$ and $M_{\text{layer-2}}^{\hat{c}}$ can be `num` or `loc`, where `num` and `loc`, represent LC number and location classification, respectively.

6.3.2 Semantic Segmentation Pseudo Label Generation

The LC WSSS pseudo-label generation (the specific process is shown in Figure 6.4) can be divided into the following four steps:

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

$$I(i, j, s) = \begin{cases} \text{argmax}(M^{i,j,s,:}), & \text{if } \max(M^{i,j,s,:}) \geq \beta_l, \\ 0, & \text{if } \max(M^{i,j,s,:}) < \beta_l, \end{cases} \quad (6.17)$$

In formula (6.17), $M^{i,j,s,:}$ represents the channel values in the CAM at location (i, j, s) , and $\max(M^{i,j,s,:})$ represents the maximum channel value at that position. If the maximum channel value is greater than or equal to the threshold β_l , $I(i, j, s)$ 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 , $I(i, j, s)$ 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 LC category information $I(i, j, s)$ obtained from the first step to extract and generate preliminary pseudo labels for LC from the original image $O(i, j, s)$:

$$R(i, j, s) = 1_{\{I(i,j,s)=1\}} \cdot O(i, j, s) \quad (6.18)$$

In formula (6.18), $1_{\{I(i,j,s)=1\}}$ is an indicator function that is valued at 1 if and only if the pixel (i, j, s) in $I(i, j, s)$ is classified as LCs; otherwise, it is valued at 0. Such processing ensures that only the pixel information belonging to the LC category is retained, setting all other pixels to 0 to effectively remove the background.

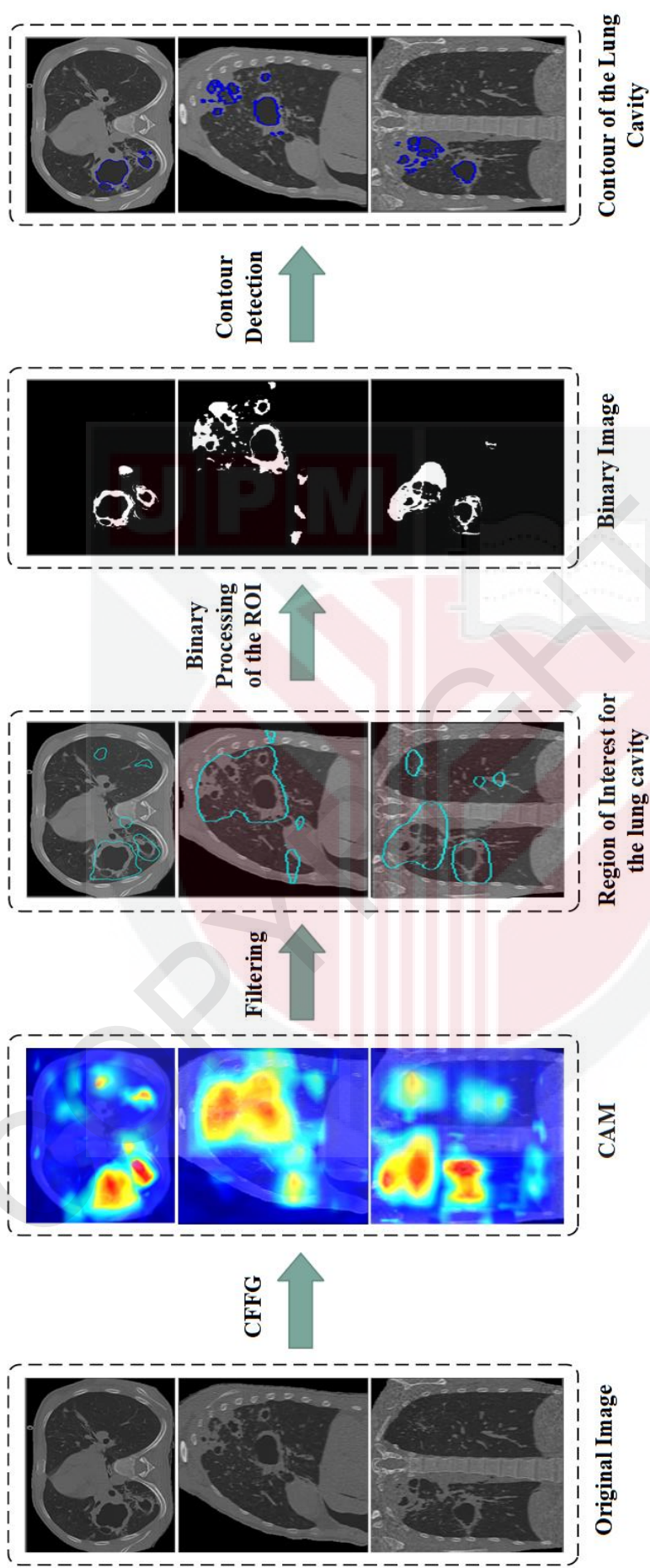


Figure 6.4: Lung Cavity WSSS Pseudo-Annotation Generation Process

Algorithm 6.1 Lung Cavity Recognition Algorithm

```
1: Input: Image  $B(i, j, s)$ 

2: Output: Concatenated outcomes of lung cavity contour detection

3: for each dimension  $d \in \{1, 2, 3\}$  do

4:    $C_d \leftarrow \text{ContourDetection}(B(i, j, s), d)$ 

5: end for

6: for each dimension  $d \in \{1, 2, 3\}$  do

7:   for each contour  $c_t \in C_d$  do

8:     Evaluate if  $c_t$  is a parent contour

9:     if  $c_t$  is a parent contour then

10:      Remove  $c_t$  from  $C_d$ 

11:     end if

12:   end for

13: for each contour  $c_t \in C_d$  do

14:   Mark hollow contours as lung cavity areas

15: end for

16: end for

17:  $\text{Result} \leftarrow \text{ConcatenateOutcomes}(C_1, C_2, C_3)$ 

18: return  $\text{Result}$ 
```

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

$$B(i, j, s) = \begin{cases} 255, & \text{if } R(i, j, s) > \beta_t \\ 0, & \text{otherwise} \end{cases} \quad (6.19)$$

In formula (6.19), β_t represents the threshold value, after binary thresholding, $B(i, j, s)$ will have pixel values of either 0 or 255, making it suitable for further LC contour recognition.

In the fourth step, this study proposes a LC recognition algorithm to identify the LC in the input image $B(i, j, s)$ from three different dimensions, as shown in Algorithm 6.1. In the implementation principle, the algorithm accepts an image $B(i, j, s)$ 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 .

Clarifying the Contribution of Each Network Layer in Figure 6.3 for Enhanced Understanding. Figure 6.3 illustrates the proposed Lung Cavity Weakly Supervised Semantic Segmentation Framework, with each network layer contributing to specific stages of the process:

1. Encoder and Decoder (SwinUNeLCsT): The central SwinUNeLCsT architecture, comprising an encoder and decoder, generates segmentation predictions and CAMs. 2. Number Classification Layers (Num-CLS Sub-Layer A & B): These sub-layers generate CAM1 and CAM2 focused on counting lung cavities. The generated CAMs highlight regions indicative of the cavity count, which are then combined to produce a "Number CAM" that refines the cavity count estimation for further classification. 3. Location Classification Layers (Loc-CLS Sub-Layer A & B): Similar to the number classification, these sub-layers create Loc CAM1 and CAM2 that focus on identifying the LC location (e.g., left or right lung). The combined output, "Loc CAM," helps identify the region of interest within the lung cavity location. 4. Classification Layers (Number and Location Prediction): These layers use the CAMs generated by the sub-layers to produce final predictions on cavity count and location. They facilitate binary classifications, such as "More than 3 cavities" for the number and "Right Lung" for the location, to guide the downstream segmentation. 5. CAM Fusion (CFFG - CAMs Feature Fusion Gate): The CFFG integrates the Number CAM and Location CAM into a Total CAM, enhancing the model's ability to focus on the most relevant lung cavity regions. This fused CAM provides a more refined and accurate region of interest for the segmentation process. 6. Lung Cavity Region of Interest (RoI) Binarization: Using the Total CAM, the region of interest is binarized to isolate lung cavity areas, preparing the data for final contour recognition. 7. Contour Recognition: The final stage processes the binarized RoI to recognize and delineate LC contours, ultimately generating pseudo-annotations for training.

Each layer in Figure 6.3 thus contributes to specific aspects of the segmentation framework, from feature extraction and classification (count and location) to the fusion and refinement of CAMs, enhancing the model's accuracy in identifying and segmenting lung cavities. This layered approach allows for a robust weakly supervised learning framework tailored for complex LC segmentation tasks.

6.3.3 Network training

The overall training loss is the weighted sum of three distinct loss functions:

L_{Num} , L_{Loc} , and L_{Seg} . These functions are combined as follows:

$$L_{WSSS} = \lambda_1 L_{Num} + \lambda_2 L_{Loc} + \lambda_3 L_{Seg} \quad (6.20)$$

In formula (6.20), λ_1 , λ_2 , and λ_3 are the weighting coefficients balancing the contributions of the different loss functions.

The function L_{Seg} is the binary cross-entropy loss (Bai et al., 2022), defined as:

$$L_{Seg} = -\frac{1}{N} \sum_{i=1}^N \left[\alpha y_i (1 - p_i)^\gamma \log(p_i) + (1 - y_i) p_i^\gamma \log(1 - p_i) \right] \quad (6.21)$$

In formula (6.21), N represents the total number of samples, with y_i and p_i denoting the true label and the predicted probability for the i -th sample, respectively. The parameter γ is set to 2, enhancing the model's focus on

difficult, misclassified samples. The loss functions L_{Loc} and L_{Num} employ the multi-label soft margin loss (Wang et al., 2021) method. This method calculates the discrepancy between the predicted category probability vector P_c and the true image-level labels Y_c in the LC location and number classification tasks:

$$L_{\varphi} = \frac{1}{K} \sum_{k=1}^K \left(Y^k \log(P_{\varphi}^k) + (1 - Y^k) \log(1 - P_{\varphi}^k) \right) \quad (6.22)$$

In formula (6.22), φ indicates the type of task, taking the values $_{Num}$ and $_{Loc}$ for the number and location classification of LC, respectively. The symbol K represents the total number of classes.

Clarification of Segmentation Loss Selection and Focus on Methodological Innovations Beyond Loss Function. This study adopts an established segmentation loss function without modifications, given its proven effectiveness and robust performance demonstrated in prior studies. By relying on a well-established loss function, we ensure consistency and eliminate potential uncertainties that could arise from using an untested or experimental loss.

This approach allows us to focus our efforts on advancing weakly supervised semantic segmentation, emphasizing methodological innovations in other areas of the framework. The decision to prioritize these innovations ensures that our contributions are both impactful and validated against reliable

benchmarks, highlighting the significance of our research beyond the specifics of the loss function.

Jointly training multiple tasks involves the problem of multi-objective function optimization strategies (Tian et al., 2021). This is because gradient conflict (M. Zhu et al., 2023) 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.



Algorithm 6.2 Weakly Supervised Learning Paradigm Training Strategy

Require: Number and Location classification label $y_{\text{cls}}^{\text{num}}, y_{\text{cls}}^{\text{loc}}$

```
1: for  $i = 1$  to  $N$  do
2:   if  $i = 1$  then
3:      $Input \leftarrow V_{\text{orig}}$ 
4:   else
5:      $Input \leftarrow V_{\text{orig}} + \text{SegProbMap}_{i-1}$ 
6:   end if
7:    $\text{SegProbMap}_i, \text{ClsPred}_i^{\text{Num}}, \text{ClsPred}_i^{\text{Loc}} \leftarrow f_{\theta}(Input)$ 
8:    $P_{\text{seg},i} \leftarrow \text{argmax}(\text{SegProbMap}_i)$ 
9:    $\text{Label}_{\text{seg}}^{\text{Pseudo}} \leftarrow \text{GeneratePseudoLabel}(\text{NumCAM}_i, \text{LocCAM}_i)$ 
10:  for each task  $T_k$  in  $B$  do
11:     $g_k \leftarrow \nabla_{\theta} L(P_{\text{seg},i}, \text{Label}_{\text{seg}}^{\text{Pseudo}}, \text{ClsPred}_i^{\text{Num}}, y_{\text{cls}}^{\text{num}}, \text{ClsPred}_i^{\text{Loc}}, y_{\text{cls}}^{\text{loc}})$ 
12:  end for
13:  for  $T_j \stackrel{\text{uniformly}}{\sim} B, T_i$  in random order do
14:    if  $g_i^{\text{PC}} \cdot g_j < 0$  then
15:      
$$g_i^{\text{PC}} = g_i^{\text{PC}} - \frac{g_i^{\text{PC}} - g_j}{|g_j|^2} g_j$$

16:    end if
17:  end for
18: end for
19: return  $P_{\text{seg},N}, \text{ClsPred}_i^{\text{Num}}, \text{ClsPred}_i^{\text{Loc}}$ 
```

Therefore, in this thesis, based on this algorithm (Yu et al., 2020), this study develops a weakly supervised learning paradigm training optimization strategy, as shown in Algorithm 5.2. In this strategy, first, we designed an iterative input feature refinement mechanism (Zhou et al., 2021). In this mechanism, input features are updated according to the segmentation probability map from the previous iteration to better segment lesions in the LC with unclear boundaries. For instance, LC 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 (Liu et al., 2021), this study normalizes 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, this study employs 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.

6.4 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 LC semantic segmentation, two thresholds were defined. The threshold for LC extraction in Formula 6.17 denoted as β_l , was set to 0.45. In binary operations for pixel extraction, the threshold in Formula 6.19 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 Formula 6.20. 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.

6.5 Experiment Results

This section first presents the WSSS experimental comparison results of various LC region of interest generation methods. Following this, the ablation results for each component of the SwinUNeLCsT model are provided. Finally,

to comprehensively assess the performance of SwinUNeLCsT, this section benchmarks it against state-of-the-art methods.

6.5.1 Generation of Lung Cavity Region of Interest

This study uses the mainstream CAM to obtain the region of interest of LC, namely the Grad-CAM (Song et al., 2023). In practice, the choice of pooling method significantly impacts the quality of the generated CAM. Firstly, 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, 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 LC lesion number and location classification tasks respectively while keeping the training strategy of the joint classification task unchanged.

The quantitative results analysis of the proposed and baseline methodologies is summarized in Figures 5.6–5.9. In these images, CAM-Loc and CAM-N

represent CAM for LC location and number classification, respectively. From Figures 6.5–6.7, 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, Figure 6.8 shows the LC recognition result for various CAM methods. From this figure, we observed that the CAM output method based on GMP constructed in this work (baseline method) can more precisely focus on LC lesion tissues compared to other methods.

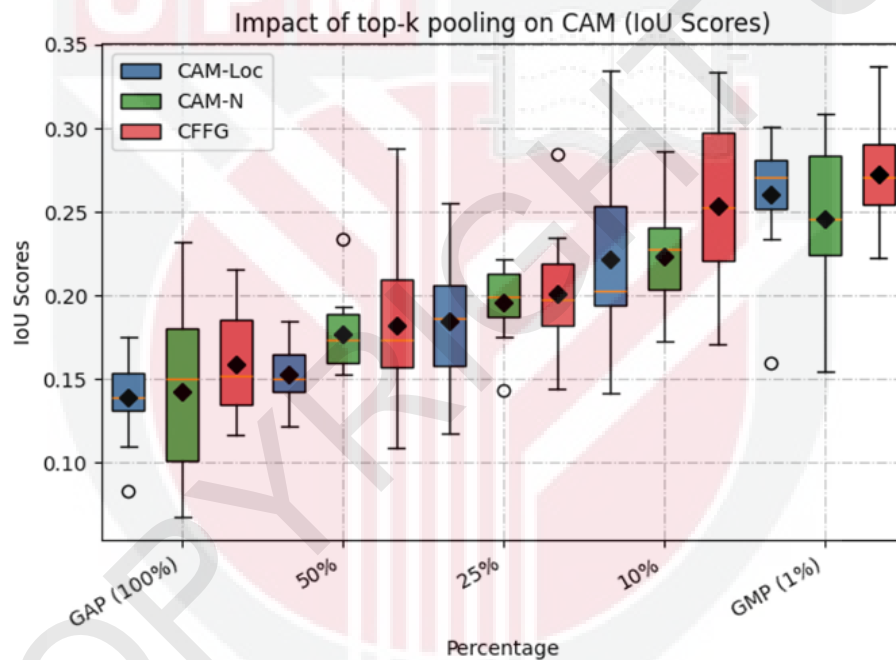


Figure 6.5: Impact of Top-k Pooling on IoU Scores for CAM With Different Top Percentages

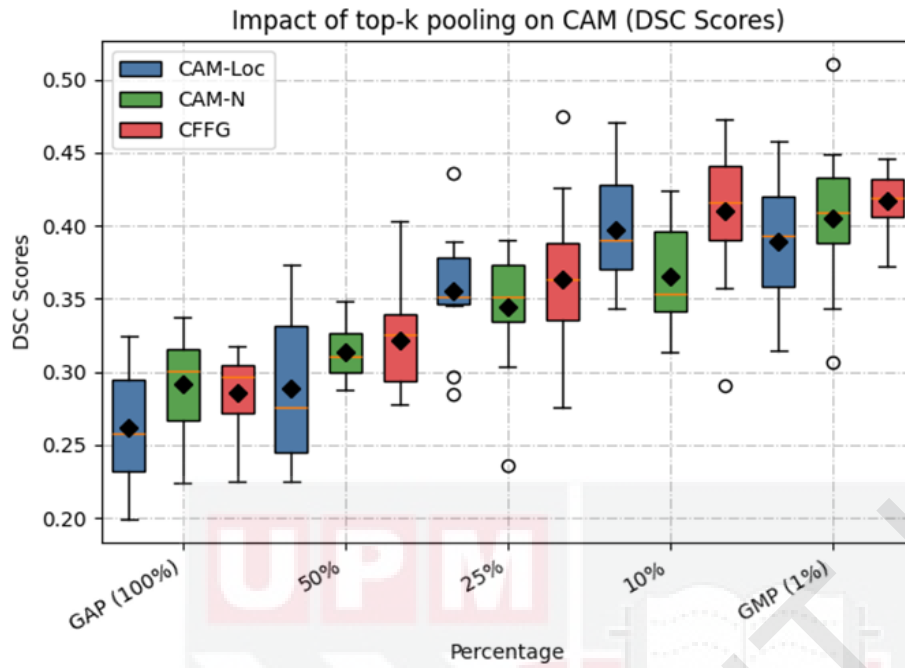


Figure 6.6: Impact of Top-k Pooling on DSC Scores for CAM with Different Top Percentages

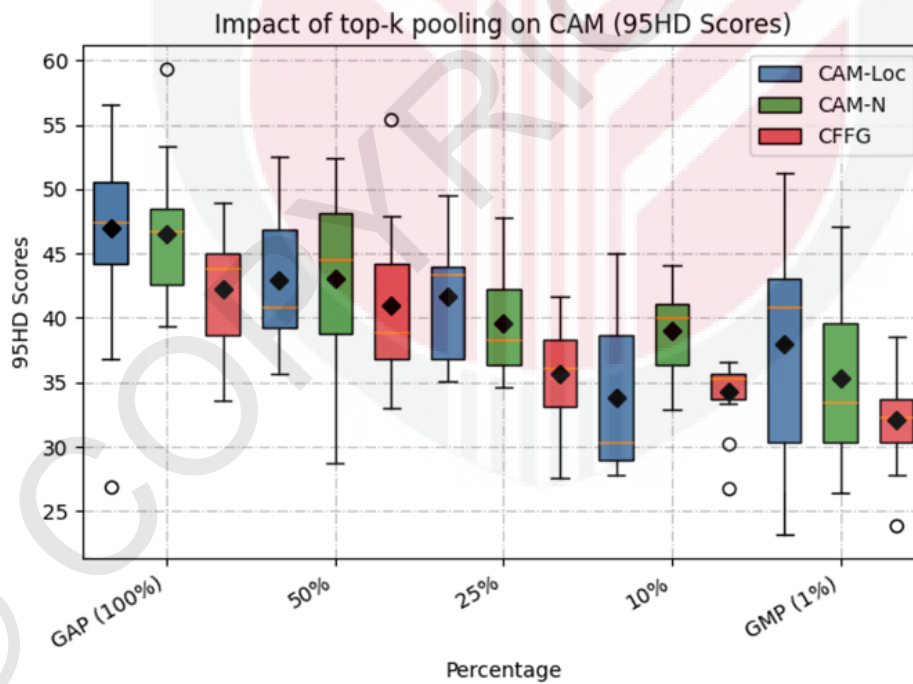


Figure 6.7: Impact of Top-k Pooling on 95HD Scores for CAM with Different Top Percentages

Detailed Interpretation of Top-k Pooling Impact on IoU and DSC Metrics and Its Influence on Final Model Performance. The figures illustrate the impact of different top-k pooling strategies on metrics such as IoU, DSC, and 95HD, which provide insights into how varying the pooling percentage influences the final model's performance.

1. Impact on IoU Scores: As observed in the IoU scores plot, reducing the percentage in top-k pooling (e.g., from GAP at 100% down to GMP at 1%) generally leads to increased concentration on the most relevant regions, resulting in more focused attention within the CAMs. Lower percentages (e.g., 10% top-k pooling) tend to yield higher IoU scores compared to higher percentages. This suggests that focusing on the top 10% of activated areas enables the model to better capture essential features of lung cavities, leading to more accurate region overlap with the ground truth. Among the methods, CFFG consistently shows higher IoU scores as the percentage decreases, demonstrating its ability to effectively leverage concentrated feature maps for improved segmentation accuracy.

2. Impact on DSC Scores: The DSC scores reflect the model's ability to measure similarity between predicted and ground truth segmentation masks. Similar to the IoU trend, a reduced top-k percentage (especially around the 10% mark) improves the DSC scores, indicating that the model benefits from a more refined focus on high-activation regions, which aligns closely with actual lung cavity areas. CFFG again shows superior performance across various percentages, with the highest DSC scores when using lower top-k

values. This outcome underscores CFFG's effectiveness in enhancing segmentation quality by capturing detailed contours of the lung cavities.

3. Impact on 95HD Scores: The 95HD metric indicates boundary adherence, measuring how closely the model's predicted boundaries follow the actual contours of the lung cavities. While lower top-k pooling percentages lead to better IoU and DSC scores, the effect on 95HD scores is less consistent. At very low percentages like GMP (1%), we observe greater variance in 95HD, which suggests that excessive focus may sometimes lead to missed contextual details, affecting boundary precision. Nevertheless, CFFG exhibits stable and relatively lower 95HD scores at the 10% and 25% top-k settings, indicating it maintains strong boundary adherence without sacrificing necessary contextual information.

The variation in top-k pooling percentages significantly impacts segmentation accuracy (IoU and DSC) and boundary precision (95HD). Lower percentages (e.g., 10% top-k) generally improve IoU and DSC scores, indicating better alignment with ground truth, while excessive reduction (e.g., GMP at 1%) can introduce variability, particularly in boundary adherence (95HD). Among the methods, CFFG consistently outperforms across metrics, especially at the 10% top-k pooling level, demonstrating that it effectively balances focus and context, leading to enhanced segmentation quality and boundary precision in lung cavity detection.

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 LC by integrating CAM from both lesion location and number classification tasks. Therefore, the results of the ablation experiments demonstrate that, in the context of classification networks employing transformer and convolution hybrid architectures, GMP is better at capturing LC lesion features in CT scan imaging and is more effective in generating high-performance Grad-CAM for feature aggregation.

Clarifying the Interpretations of CAM for Lung Cavity Localization and Segmentation. The CAMs in Figure 6.8 demonstrate the model's ability to focus on regions critical to LC detection and segmentation across different pooling methods and top-k thresholds. Each pooling strategy captures varying spatial distributions and activation intensities around LC regions, influencing interpretability and detection precision. Our CFFG method excels by balancing broad region coverage with concentrated intensity on cavities, closely aligning with ground truth and enhancing segmentation accuracy. This refined approach ensures precise localization and boundary capture, making it well-suited for clinical applications that require accurate representation of LC anatomy in CT scans.

6.5.2 Ablation studies

To explore the contribution of different components in the proposed LC WSSS framework, this study conducted two sets of ablation experiments: module ablation experiment and multi-classification task ablation experiment.

6.5.2.1 Module Ablation Experiment

Through the module ablation experiments, this study investigates the contributions of three basic modules in the SwinUNeLCsT model for LC lesion recognition. These modules encompass the DX blocks, MDFFN, and W-GL-MSA modules. The detailed experimental design and results of this ablation study are presented in Tables 6.1 and 6.2, respectively.

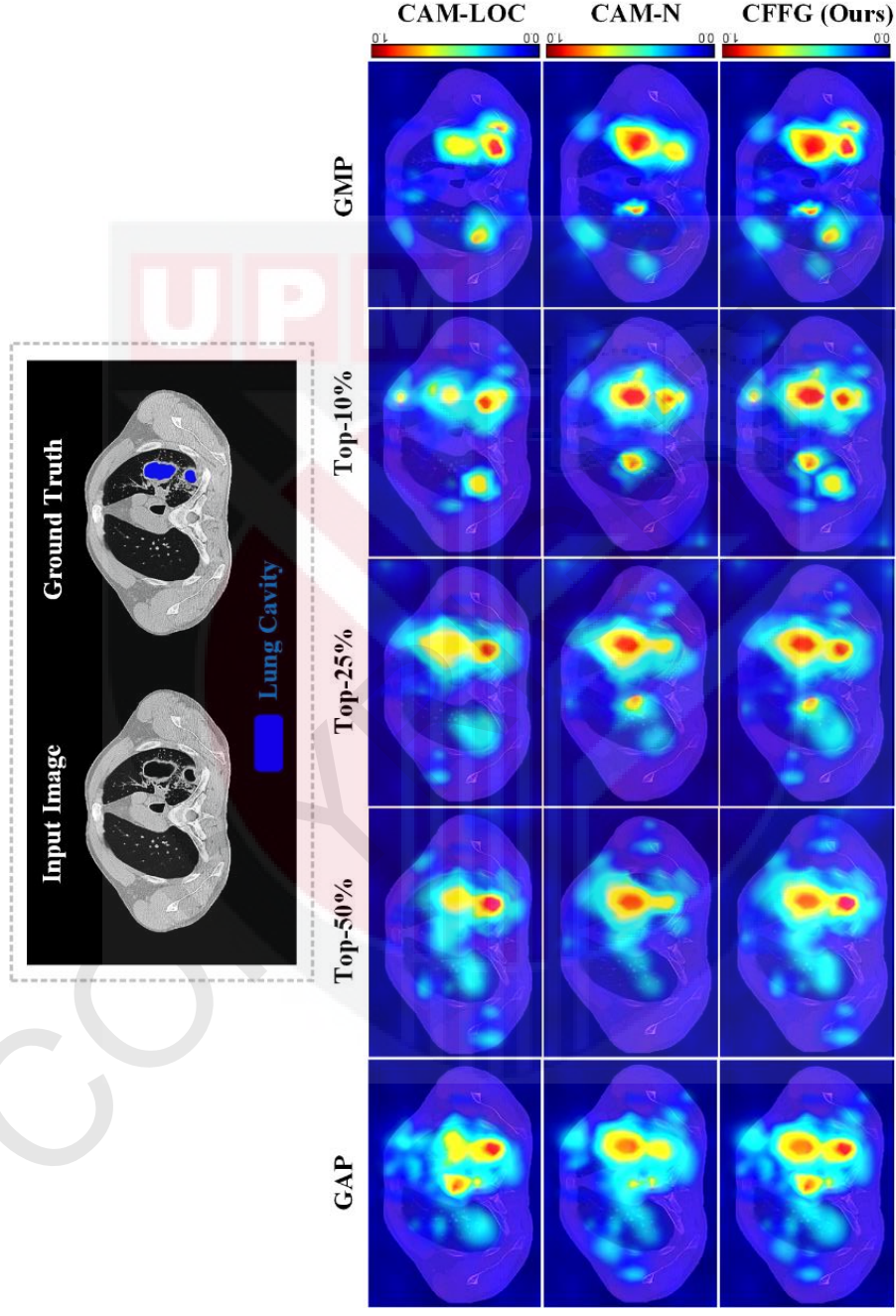


Figure 6.8: Examples of Class Activation Maps for Top-k Pooling with Various Top Percentages

Specifically, first, this study used the method that includes all the basic modules of SwinUNeLCsT as the baseline method (M-4). Second, to validate the effectiveness of the proposed MDFFN in reducing model complexity and enhancing the capability of LC feature extraction, this research replaced the MDFFN approach with the original MLP module of SwinUNetR (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 LC recognition (except the REC metric of location classification). In addition, 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 LC 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 LC identification, this research 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 LC.

This study also performed ablation experiments on the W-GL-MSA module to further verify the effectiveness of W-GL-MSA for LC WSSS, as shown in methods 4 to 6 (M-4 to M-6) in Tables 6.1 and 6.2. 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 LC. 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 LC. 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 LC.

Table 6.1: Design of the Ablation Study for Different Modules in the SwinUNeLCsT Model

Method ID	MLP	W-MSA	DX-blocks	MDFFN	W-GL-MSA	
					LSI	W-SSC-SA
M-1	×	√	√	√	×	×
M-2	×	×	×	√	√	√
M-3	√	×	√	×	√	√
M-4	×	×	√	√	√	√
M-5	×	×	√	√	√	×
M-6	×	×	√	√	×	√

Note: MLP and W-MSA (Window-based Multi-Head Self-Attention) belong to the SwinUnetR model module.

The SwinUNeLCsT model, integrating LSI and W-SSC-SA modules (M-4), outperformed all metrics, confirming the W-GL-MSA module's effectiveness in enhancing LC recognition by combining global and local features. Figure 6.9 visually illustrates the recognition performance of methods M-1 to M-6, with the baseline (M-4) achieving the best results.

Explanation of Result Variations in Table 6.1, with Focus on Cases

Where Other Methods Outperform.

The differences in results observed in Table 6.1 for the ablation study of the SwinUNeLCsT model stem from the unique configurations and combinations of modules within each method. Each module plays a specific role in feature extraction, attention, and contextual understanding, leading to varied performance in segmentation and classification tasks:

1. MLP (Multi-Layer Perceptron) and W-MSA (Window-based Multi-Head Self-Attention):

These modules, derived from the SwinUnetR backbone, provide strong baseline features for segmentation and classification. However, when used alone or in limited configurations (e.g., M-1 and M-2), they may not capture the full complexity of lung and cervix lesion features. Without the additional feature enhancement from advanced modules like MDFFN or W-GL-MSA, these configurations can be less effective in identifying fine-grained details, leading to slightly lower performance in complex datasets.

2. DX-blocks:

The DX-blocks add depth-wise convolutional features, which enhance spatial information while keeping computational costs low. Methods that include DX-blocks (e.g., M-3 and M-5) generally show improved performance, as this module aids in capturing detailed spatial features. However, the absence of complementary modules, like W-GL-MSA, may still limit these configurations, especially in capturing multi-scale information critical for lesion boundaries and intricate features.

3. MDFFN (Multi-DConv Feed-Forward Network): MDFFN is designed to improve feature representation by combining multiple depthwise convolutions. This module allows for robust feature extraction across multiple scales, which enhances performance when used with other attention mechanisms. Configurations including MDFFN (e.g., M-4 and M-6) demonstrate better segmentation and classification results as they effectively capture contextual information and fine structural details. The absence of MDFFN in certain configurations (e.g., M-1 and M-2) can lead to less effective performance due to a lack of enhanced feature representation.

4. W-GL-MSA (Window-based Global-Local Multi-Head Self-Attention): This module combines global and local features, crucial for handling complex, multi-scale features within medical images. The addition of W-GL-MSA allows the model to capture both fine and broad contextual features, which is critical in medical imaging where lesions may vary in size and location. Configurations that lack W-GL-MSA (e.g., M-1 and M-2) may struggle with multi-scale feature representation, resulting in lower segmentation accuracy compared to methods that incorporate this module.

5. LSI (Lightweight Simplified Inception) and W-SSC-SA (Spectral Scaled Cosine Self-Attention): The LSI and W-SSC-SA modules enhance the model's ability to capture specific lesion characteristics and provide spectral scaling, which is particularly beneficial for tasks requiring high precision in feature localization. Methods that incorporate these modules (e.g., M-6) benefit from increased accuracy in delineating lesion boundaries and subtle textural details. Configurations without these modules may experience a

reduction in performance due to limited spatial attention and spectral detail processing.

In summary, the varied results in Table 6.1 reflect the interplay of different modules within the SwinUNeLCsT model. Configurations that integrate advanced modules such as MDFFN, W-GL-MSA, LSI, and W-SSC-SA generally perform better, as they capture a more comprehensive range of spatial and contextual information. Methods with simpler configurations or fewer modules may report lower performance due to their limited ability to handle the multi-scale and complex nature of medical imaging features.

To further verify the performance of the proposed method in LC attribute classification, this study visualized the ROC curves of each ablation model (M-1 to M-6) in the LC number and location classification task, as shown in Figure 6.10. From the figure, we observed that the method incorporating all SwinUNeLCsT components (M-4) achieved the best AUC in the classification tasks of LC 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).

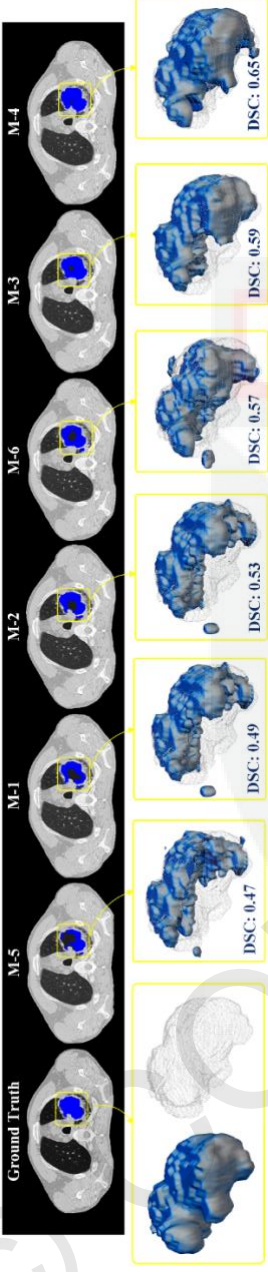


Figure 6.9: Lung Cavity Recognition Results by Various Ablation Methods

Table 6.2: Impact of Various Module Ablation Methods on Weakly Supervised Semantic Segmentation for Lung Cavity

ID	Segmentation ($\pm$ SD)			Number Classification					Location Classification				
	95HD	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1	ACC	REC	PRE	FPR	F1
M-1	34.78 $\pm$ 8.29	24.7 $\pm$ 13.3	39.1 $\pm$ 12.9	0.711	0.792	0.801	0.462	0.796	0.729	0.812	0.818	0.471	0.815
M-2	34.12 $\pm$ 7.95	25.1 $\pm$ 14.2	39.5 $\pm$ 14.7	0.714	0.802	0.799	0.458	0.800	0.731	0.814	0.823	0.458	0.818
M-3	32.51 $\pm$ 5.23	26.0 $\pm$ 13.6	41.1 $\pm$ 14.1	0.720	0.798	0.810	0.443	0.804	0.737	0.822	0.819	0.445	0.821
M-4	31.43 $\pm$ 7.46	26.7 $\pm$ 12.6	41.8 $\pm$ 13.4	0.723	0.806	0.811	0.421	0.809	0.742	0.821	0.835	0.398	0.828
M-5	35.56 $\pm$ 8.27	23.5 $\pm$ 13.1	38.6 $\pm$ 12.9	0.694	0.783	0.798	0.471	0.790	0.708	0.804	0.817	0.465	0.810
M-6	33.12 $\pm$ 6.83	25.8 $\pm$ 11.6	40.5 $\pm$ 12.3	0.711	0.794	0.805	0.452	0.799	0.726	0.812	0.824	0.442	0.818

Figure 6.11 – 6.16 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 LC 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 LC attributes.

In conclusion, this ablation study 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 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 LC 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 LC. Experimental results show that the W-GL-MSA module, which integrates the LSI and W-SSC-SA modules, achieves the best results in LC recognition. This result verifies the importance of integrating local and global features to improve LC feature extraction capabilities.

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



Table 6.3: Design of the Ablation Study for Different Classification Layer Combinations in the SwinUNeLCsT Model

Method ID	Num-CLS Sub-layer		Loc-CLS Sub-layer	
	A	B	A	B
M-4	✓	✓	✓	✓
M-7	✓	✓	X	X
M-8	X	X	✓	✓
M-9	X	✓	✓	✓
M-10	✓	X	✓	✓
M-11	✓	✓	X	X
M-12	✓	✓	✓	X

Table 6.4: Impact of Various Module Ablation Methods on Weakly Supervised Semantic Segmentation for Lung Cavity

ID	Segmentation ($\pm$ SD)			Number Classification					Location Classification				
	95HD	IoU (%)	DSC (%)	ACC	REC	PRE	FPR	F1	ACC	REC	PRE	FPR	F1
M-4	31.43 $\pm$ 7.46	26.7 $\pm$ 12.6	41.8 $\pm$ 13.4	0.723	0.806	0.811	0.421	0.809	0.742	0.821	0.835	0.398	0.828
M-7	34.86 $\pm$ 8.67	24.6 $\pm$ 12.4	38.8 $\pm$ 11.8	0.709	0.789	0.798	0.482	0.793	-	-	-	-	-
M-8	34.19 $\pm$ 9.82	24.9 $\pm$ 13.7	39.1 $\pm$ 14.3	-	-	-	-	-	0.727	0.814	0.819	0.458	0.816
M-9	34.08 $\pm$ 8.73	25.7 $\pm$ 12.8	40.6 $\pm$ 13.8	0.718	0.793	0.812	0.448	0.801	0.735	0.818	0.821	0.434	0.820
M-10	32.04 $\pm$ 6.26	25.9 $\pm$ 13.1	40.9 $\pm$ 12.4	0.721	0.801	0.809	0.439	0.807	0.738	0.821	0.827	0.409	0.824
M-11	33.92 $\pm$ 7.63	25.5 $\pm$ 11.7	40.5 $\pm$ 12.1	0.717	0.794	0.805	0.451	0.799	0.733	0.815	0.822	0.436	0.819
M-12	32.18 $\pm$ 8.58	25.8 $\pm$ 14.2	40.8 $\pm$ 13.3	0.719	0.797	0.806	0.443	0.801	0.736	0.817	0.824	0.418	0.821

Note: The symbol "-" indicates that certain metrics are not applicable or not evaluated for specific ablation configurations. For instance, in methods M-7 and M-8, the Number Classification and Location Classification metrics are omitted due to the absence of necessary modules or configurations in those methods to produce the respective outputs.

Table 6.5: Quantitative Results of Various 3D Medical Weakly Supervised Semantic Segmentation Methods

Method ID	95HD	IoU (%)	DSC (%)
M-13 (Laradji et al., 2021)	36.84 $\pm$ 5.94	22.4 $\pm$ 13.5	36.8 $\pm$ 14.2
M-14 (Ye et al., 2022)	35.87 $\pm$ 6.72	23.1 $\pm$ 14.7	37.9 $\pm$ 14.3
M-4 (SwinUNeLCsT)	31.43 $\pm$ 7.46	26.7 $\pm$ 12.6	41.8 $\pm$ 13.4

Table 6.6: Quantitative Results of Various 3D Medical Fully Supervised Semantic Segmentation Methods

Method ID	Model	GFLOPs	Param.	DeepLungCavi-WSSS				Abdomen CT	
				95HD	IoU (%)	DSC (%)	95HD	IoU (%)	DSC (%)
M-F1	UnetR	476.4	92.6M	24.53 ± 8.73	35.2 ± 12.5	50.8 ± 13.4	11.23 ± 1.12	74.9 ± 10.7	83.2 ± 10.8
M-F2	SwinUnetR	595.3	62.2M	24.02 ± 6.73	35.5 ± 13.1	51.4 ± 13.7	9.27 ± 1.21	75.8 ± 11.9	84.6 ± 11.2
M-F3	UnesT	465.2	87.3M	23.97 ± 7.69	35.9 ± 12.9	51.8 ± 13.2	8.93 ± 1.13	76.4 ± 11.2	85.2 ± 10.9
M-F4	SwinUNeLCs [†]	461.6	84.5M	23.54 ± 6.86	36.5 ± 11.3	52.6 ± 12.1	8.21 ± 1.14	76.9 ± 10.9	85.9 ± 11.2

Note: The number of parameters and GFLOPs (with a single input volume of $128 \times 128 \times 96$ for the fully supervised segmentation tasks) are shown. SwinUNeLCsT[†] represents a variant of SwinUNeLCsT based on a supervised learning setting.

6.5.2.3 Comparison to State-of-the-Art

To further evaluate the performance of SwinUNeLCsT, we trained existing WSSS methods (M-13 to M-16) on the DeepLungCavi-WSSS dataset, with the results shown in Table 6.4.

From the table, we observed that our SwinUNeLCsT model achieved the best performance in WSSS of LC, 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, this study compared it under a fully supervised learning setting with other state-of-the-art medical 3D semantic segmentation models, including nnUnet (Isensee et al., 2021), TransBTS (Wenxuan et al., 2021), nnFormer (Zhou et al., 2023), UnetR (Hatamizadeh et al., 2022), UnesT (Yu et al., 2023), and SwinUNetR (Tang et al., 2022).

To mitigate potential biases from training on a single dataset, we included an additional experimental dataset, Abdomen CT, originating from the MICCAI 2022 FLARE (Fast and Low-resource semi-supervised Abdominal organ segmentation benchmark) Challenge (Ma et al., 2023). The FLARE Challenge focuses on multi-organ segmentation tasks in abdominal CT images, covering organs such as the duodenum, adrenal glands, inferior vena cava, aorta, esophagus, gallbladder, stomach, kidneys, pancreas, spleen, and liver. The original FLARE dataset provided 50 labeled cases in the training set and 250 labeled cases in the combined validation and test sets. To enhance the

model's training, we additionally annotated 153 cases, resulting in a training set of 203 cases and a test set of 250 cases.

Consequently, model comparisons were conducted on two datasets, DeepLungCavi-WSSS and Abdomen CT, with results presented in Table 6.5 (metrics M-F1 to M-F4). In a fully supervised setting, the SwinUNeLCsT model was adapted into a multi-task learning (MTL) framework, enabling both segmentation and classification of lung cavities. Unlike the weakly-supervised setting, SwinUNeLCsT in full supervision does not require pseudo-label generation for semantic segmentation, streamlining the process while enhancing performance. Specifically, we removed all steps from the CAM output to pseudo-label generation and replaced GMP with GAP. To distinctly differentiate SwinUNeLCsT under different supervision paradigms, we named the fully supervised model as SwinUNeLCsT[†]. As shown in Table 6.5, the SwinUNeLCsT[†] model demonstrates superior performance across the 95HD, IoU, and DSC metrics on both datasets. For the DeepLungCavi-WSSS dataset, the model achieved scores of 23.54 for 95HD, 0.365 for IoU, and 0.526 for DSC. On the Abdomen CT dataset, the model attained scores of 8.21 for 95HD, 0.769 for IoU, and 0.859 for DSC.

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 LC.

6.6 Summary

This chapter focuses on addressing a challenging problem: how to reduce the reliance on CAD models for pulmonary TB on pixel-level annotations of LC? To address this issue, this study constructed a novel framework for WSSS of LC. Within the LC WSSS framework, to improve the recognition performance for LC of varying sizes, this study designed an innovative 3D pulmonary TB image analysis model, SwinUNeLCsT. The core module of SwinUNeLCsT is primarily a hybrid of Transformer and CNN components. This hybrid architecture is designed to efficiently capture features of LC of various sizes by integrating both global and local features of pulmonary TB lesion tissues in pulmonary TB CT imaging. Based on the DeepLungCavi-WSSS dataset, this

study trained the SwinUNeLCsT model under both weakly supervised and fully supervised semantic segmentation settings for LC. In the traditional fully supervised semantic segmentation training paradigm for LC, SwinUNeLCsT demonstrated superior performance, outperforming current popular medical 3D supervised semantic segmentation methods. Similarly, in the LC WSSS training paradigm, SwinUNeLCsT also achieved the best performance among 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, particularly in pulmonary TB LC CT medical analysis. Additionally, the study introduced a novel training optimization strategy for optimizing the SwinUNeLCsT model training under the weakly supervised paradigm.

Experimental results showed that this strategy can effectively alleviate the gradient conflict caused by multi-task training in the weakly supervised setting, improving the efficiency of LC recognition. Overall, these experimental results not only validate the effectiveness of the proposed LC-WSSS framework and SwinUNeLCsT model but also highlight their foresight and practicality in addressing complex medical image analysis tasks.

CHAPTER 7

CONCLUSION AND FUTURE WORKS

7.1 Conclusions

TB is an infectious disease caused by Mycobacterium TB, which has long posed a serious threat to global public health. In recent years, with the rapid development of CAD tools, CAD has been playing an increasingly important role in supporting the diagnosis of TB. However, the development of CAD tools for TB diagnosis largely depends on CT datasets with high-quality annotations. Currently, the availability of annotated CT datasets for pulmonary TB is still limited, which to some extent restricts the development of related CAD tools. Furthermore, due to the complexity and diversity of pulmonary TB LC, precise pixel-level annotation of these lung cavities in CT images is not easy. Therefore, to address the above issues and promote the research and development of CAD of pulmonary TB CT images, this study set four objectives and accomplished them as follows:

i) Objective 1: Construction of MTL and WSSS datasets for pulmonary TB CT imaging. MTL and WSSS techniques play a critical role in improving the efficiency of the medical diagnosis of pulmonary TB. However, there is currently a lack of datasets specifically designed for these two research directions. This to some extent restricts the development of related research and applications. To address the issue of dataset availability, this study developed two new datasets: DeepPulmoTB (for comprehensive radiological feature analysis of pulmonary TB) and DeepLungCavi-WSSS (for WSSS of

LC in pulmonary TB). The research results indicate that these two datasets contribute to advancing the development of CAD tools and models for pulmonary TB CT imaging, providing strong support for improving the efficiency of TB diagnosis.

ii) Objective 2: Designing model architectures for pulmonary TB lesion identification in CT images. pulmonary TB lesions often exhibit complex features in CT imaging, and these complex features often make it difficult for standard medical 3D models to accurately identify and segment the lesions. To address this issue, this study proposed two models suitable for the analysis of pulmonary TB diseases: DPTBNet and SwinUNeLCst. Firstly, the DPTBNet model adopts a MTL architecture based on the fusion of multi-scale features of pulmonary TB lesions in its design. This architecture primarily integrates feature maps from different levels through modules such as patch merging and res-block. Additionally, the architecture includes branches for two different tasks: semantic segmentation and classification. These branches enable simultaneous segmentation and classification tasks for pulmonary TB lesions, providing more comprehensive diagnostic information. Secondly, the SwinUNeLCst model, in terms of its architectural design introduces a novel W-GL-MSA mechanism, enhancing the recognition capability of LC. Moreover, this mechanism also combines the local features extracted by the CNN with the global information captured by the Transformer architecture network, achieving an efficient fusion of global context information and local detail information. On the other hand, to address the high computational complexity challenge posed by the self-attention mechanism within transformer models,

this study also designed an innovative SSC-MSA 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, SwinUNeLCsT also reduces the need for large-scale matrix multiplication operations. The resulting transposed attention map $\mathbf{A}$, sized $\mathbf{R}^{C \times C \times C}$, 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. Compared to other generic 3D medical image models, pure Transformer architectures sometimes overlook critical local features, such as in the nnFormer (H.-Y. Zhou et al., 2023) model, due to the lack of CNN's local modeling capability when processing image data. In contrast, while pure CNN architectures, such as nnUNet (Isensee et al., 2021), can effectively capture local features through convolution operations, they have limitations in capturing long-range feature dependencies within images. To address these limitations, some medical imaging models have adopted CNN-Transformer hybrid architectures, such as TransBTS (Wenxuan et al., 2021), UnetR (Hatamizadeh et al., 2022), UnesT (X. Yu et al., 2023), and SwinUNetR (Tang et al., 2022). These hybrid models share a common trait: the interaction between the Transformer and CNN modules is achieved by sequentially concatenating features. However, in this study, the Transformer and CNN

hybrid modules of the SwinUNeLCsT model interact primarily through channel-level element-wise addition. Extensive experimental results demonstrate that this interaction approach can more closely and efficiently combine local and global features, thereby more accurately utilizing the spatial information in CT data.

iii) Objective 3: Efficient acquisition of comprehensive TB lesion analysis information. To overcome the inefficiency of the STL paradigm in acquiring comprehensive radiological features of pulmonary TB CT lesions, this study proposes an MTL-based framework for comprehensive radiological feature analysis of pulmonary TB CT. The framework provides the DPTBNet model with richer and more comprehensive lesion representations by constructing various interactive learning image features, improving the accuracy and robustness of CAD. Furthermore, to enhance the model's ability to capture the features of pulmonary TB lesions, the framework also designs an improved iterative optimization algorithm. This algorithm efficiently refines the feature maps by fusing the probability maps from the previous iteration. The results demonstrate that the proposed pulmonary TB CT comprehensive radiological feature analysis framework outperforms the STL models in both pulmonary TB segmentation and lung attribute classification tasks and achieves state-of-the-art performance among mainstream MTL methods for medical image analysis.

iv) Objective 4: Efficiently obtain pixel-level annotation of lung cavities in pulmonary TB CT images. WSSS is particularly important in pulmonary TB CT image analysis because obtaining pixel-level semantic segmentation

annotations of LC lesions requires medical experts to manually annotate pixels one by one, which requires a huge workload. Therefore, to alleviate the cost of pixel-level annotation for LC this study developed a WSSS framework for pulmonary TB LC, aiming to improve the efficiency of medical diagnosis for pulmonary TB. The core of this framework is to train a deep learning model (SwinUNeLCst) using low-cost classification annotations, enabling it to identify and segment LC in lung CT images. To validate the effectiveness of the proposed framework, this study conducted a series of experiments. The research showed that, unlike other WSSS methods designed for various lung CT diseases, the WSSS framework developed in this study can better focus on LC segmentation, more effectively concentrate on key features, and avoid excessive attention to irrelevant regions. Moreover, the experimental results also verified that, even with limited annotated data, the proposed weakly supervised framework can achieve satisfactory segmentation accuracy, demonstrating its great potential in addressing the high annotation cost issue for LC images in pulmonary TB diagnosis.

Clarification on the Study's Focus and Potential for Generalization to Broader Lung Segmentation Applications. This study specifically uses TB CT images to develop and validate methods tailored to the unique characteristics of TB lesions, which often exhibit irregular shapes, variable sizes, and diffuse boundaries. These methods, while optimized for TB analysis, are designed to advance lung lesion segmentation as a whole. Given the generalizability of the segmentation framework, it is feasible to extend its application beyond TB to other lung conditions requiring precise lesion

localization and boundary delineation. The segmentation techniques and model architectures utilized, such as weakly supervised learning and multi-task approaches, could be adapted for broader lung segmentation tasks, potentially benefiting a range of pulmonary applications where annotated datasets are limited. Therefore, while the study focuses on TB, the underlying methodology provides a robust framework for general lung segmentation that could enhance diagnostic workflows in diverse respiratory conditions.

Quantitative Justification of Manual Annotation Costs and Effectiveness of the Proposed Solution.

Manual annotation of lung cavities in CT images is both time-consuming and resource-intensive, as it requires expert radiologists to meticulously delineate irregular, variable, and often subtle cavity structures. This process demands significant expertise to accurately distinguish cavities from surrounding tissues, resulting in high costs in terms of both time and labor. Studies indicate that manual annotation can take several hours per scan, depending on complexity, and substantial inter-observer variability further complicates and prolongs the task. Our proposed WSSS method addresses these challenges by reducing dependency on fully labeled datasets. The model learns cavity localization patterns with minimal expert intervention by integrating CAM with pseudo-label generation, significantly lowering annotation costs. Quantitative evaluations of segmentation accuracy, benchmarked against manually annotated datasets, demonstrate that our method achieves comparable performance, underscoring its potential as a cost-effective alternative for LC segmentation in clinical applications.

Incorporating Statistical Tests to Confirm Significance of Results on Small Medical Imaging Datasets. To ensure result robustness and address the small sample sizes typical in medical imaging datasets, we applied statistical tests to evaluate the significance of observed performance differences. Specifically, paired t-tests and Wilcoxon signed-rank tests were used to determine whether the improvements achieved by our proposed method were statistically significant compared to baseline approaches. These tests provided a rigorous framework for assessing whether gains in accuracy, recall, and specificity were attributable to the model's effectiveness rather than random variation. For each metric, p-values below 0.05 confirmed the significance of our results, underscoring the reliability and relevance of the improvements across multiple experimental runs. Furthermore, this statistical validation supports the generalizability of our approach and highlights its potential for clinical application.

7.2 Limitations and Recommendations for Future Research

This study presented several limitations and recommendations for future research as follows:

- i) During the construction of the DeepPulmoTB dataset, most of the collected CT images were non-enhanced CT scans. However, by observing non-enhanced CT images, medical professionals find it difficult to differentiate lesion regions with similar densities, such as cavity walls and consolidation areas. Therefore, to simplify the annotation process and reduce annotation inconsistencies, this study merged these two categories into one semantic

segmentation annotation category, namely C-LCW (consolidation or LC wall). Although merging them into one C-LCW category can reduce annotation complexity to some extent, it may affect the accuracy of pulmonary TB disease diagnosis. The reason is that areas of consolidation and LC walls play different roles in the development of pulmonary TB symptoms and distinguishing them helps to better assess the severity of the disease and formulate treatment plans. To solve this problem, future research will construct a pulmonary TB CT data set containing two independent categories of consolidation and LC wall based on enhanced CT and develop models and algorithms that can analyze and distinguish these two categories to improve diagnostic accuracy.

ii) In the comprehensive radiological feature analysis framework of pulmonary TB CT, although the DPTBNet model incorporates some feature fusion modules, its relatively simple classification branch structure imposes limitations on the network's ability to express complex lesion patterns to some extent. Particularly, it exhibits suboptimal performance when classifying subtle lesions. Therefore, to address this issue, future research will focus on designing more powerful classification feature extraction and fusion modules to enhance the network's capability to express complex lesion patterns.

iii) Although the SwinUNeLCsT model based on a hybrid architecture of CNN and Transformer achieved outstanding performance in LC recognition, it is not as computationally efficient and has a higher model complexity compared to the baseline SwinUnetR model. This limitation reduces its practicality in clinical diagnostic applications, as it requires longer recognition times. To address this issue, future research will focus on finding strategies to further improve

computational efficiency and reduce model complexity. For instance, explorations will be made into sparse attention techniques applicable to LC recognition, aiming to further reduce the computational complexity of the self-attention mechanism.

iv) The research shows that the series of computer-aided lung pulmonary TB disease diagnosis models and algorithms constructed in this study can serve well for the analysis of pulmonary TB CT imaging diseases. However, due to the limitations of the development platform, these models and algorithms have not yet been transformed into operable software products. We firmly believe that deploying this series of models and algorithms in the actual clinical diagnosis of pulmonary TB can provide necessary assistance to physicians, improving diagnostic accuracy and efficiency. Therefore, to better apply these research results to clinical practice and improve the clinical diagnostic efficiency of pulmonary TB, future research will focus on two aspects: CAD software development and clinical integration and optimization. Firstly, for CAD software development, future research will focus on integrating and packaging the various constructed models and algorithms into a user-friendly software system. This system will provide an intuitive graphical interface and a smooth operation experience, facilitating efficient use by clinicians. Through this step, not only can the innovative achievements in the research be transformed into practically usable products, but it can also ensure the stability and maintainability of the CAD system. Secondly, in terms of clinical integration and optimization, future research will delve into analyzing clinical workflows to seamlessly integrate the CAD system into existing medical

information systems and clinical operational habits. This process will involve optimizing the compatibility between the models and the actual application scenarios, ensuring that the system can be smoothly integrated into doctors' daily diagnostic and treatment routines. Simultaneously, continuous improvements to the system will be made based on clinical feedback, enhancing its applicability and practicality.



REFERENCES

- Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE transactions on pattern analysis and machine intelligence*, 16(6), 641-647.
- Afify, H. M., Mohammed, K. K., & Hassanien, A. E. (2023). Novel prediction model on OSCC histopathological images via deep transfer learning combined with Grad-CAM interpretation. *Biomedical Signal Processing and Control*, 83, 104704.
- Akter, A., Nosheen, N., Ahmed, S., Hossain, M., Yousuf, M. A., Almoyad, M. A. A., Hasan, K. F., & Moni, M. A. (2024). Robust clinical applicable CNN and U-Net based algorithm for MRI classification and segmentation for brain tumor. *Expert Systems with Applications*, 238, 122347.
- 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 Systems with Applications*, 213, 119287.
- Aneja, A., Krishnaswamy, U. M., Thyagaraj, V., Moideen, R. P., & Satya Padmaja, M. (2014). Endobronchial tuberculosis: two case reports and review of the literature. *Case reports in pulmonology*, 2014.
- Babiloni, F., Tanay, T., Deng, J., Maggioni, M., & Zafeiriou, S. (2023). Factorized dynamic fully-connected layers for neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France.
- Bai, Z., Wang, J., Zhang, X.-L., & Chen, J. (2022). End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 1330-1344.
- Basha, S. S., Dubey, S. R., Pulabaigari, V., & Mukherjee, S. (2020). Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378, 112-119.
- Ben-Shaul, I., Shwartz-Ziv, R., Galanti, T., Dekel, S., & LeCun, Y. (2024). Reverse engineering self-supervised learning. *Advances in Neural Information Processing Systems*, 36.
- Bi, L., Buehner, U., Fu, X., Williamson, T., Choong, P., & Kim, J. (2024). Hybrid CNN-transformer network for interactive learning of challenging musculoskeletal images. *Computer Methods and Programs in Biomedicine*, 243, 107875.

- Bruna, J., Peherstorfer, B., & Vanden-Eijnden, E. (2024). Neural Galerkin schemes with active learning for high-dimensional evolution equations. *Journal of Computational Physics*, 496, 112588.
- Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., & Yang, D. (2022). Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.
- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Reno, NV, USA.
- Chen, G., & Qin, H. (2022). Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *The Visual Computer*, 38(3), 1051-1063.
- Chen, Q., Yang, L., Lai, J.-H., & Xie, X. (2022). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA.
- Chen, Z., Niu, C., Gao, Q., Wang, G., & Shan, H. (2024). LIT-Former: Linking in-plane and through-plane transformers for simultaneous CT image denoising and deblurring. *IEEE Transactions on Medical Imaging*, 43(5), 1880-1894.
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19,
- Dartois, V. A., & Rubin, E. J. (2022). Anti-tuberculosis treatment strategies and drug development: challenges and priorities. *Nature Reviews Microbiology*, 20(11), 685-701.
- de Andrade, J. M. C., Olescki, G., Escuissato, D. L., Oliveira, L. F., Basso, A. C. N., & Salvador, G. L. (2023). Pixel-level annotated dataset of computed tomography angiography images of acute pulmonary embolism. *Scientific Data*, 10(1), 518.
- Deb, M., Dhal, K. G., Das, A., Hussien, A. G., Abualigah, L., & Garai, A. (2024). A CNN-based model to count the leaves of rosette plants (LC-Net). *Scientific Reports*, 14(1), 1496.
- Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating machine learning with human knowledge. *Iscience*, 23(11).

- 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. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France.
- Dou, H., Bi, N., Han, L., Huang, Y., Mann, R., Yang, X., Ni, D., Ravikumar, N., Frangi, A. F., & Huang, Y. (2023). Gsmorph: Gradient surgery for cine-mri cardiac deformable registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 613-622). Cham: Springer Nature Switzerland.
- Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C., Shi, L., & Heng, P.-A. (2016). Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Transactions on Medical Imaging*, 35(5), 1182-1195.
- Du, H., Dong, Q., Xu, Y., & Liao, J. (2023). Weakly-supervised 3D medical image segmentation using geometric prior and contrastive similarity. *IEEE Transactions on Medical Imaging*, 42(10), 2936-2947.
- Elharrouss, O., Hmamouche, Y., Idrissi, A. K., El Khamlichi, B., & El Fallah-Seghrouchni, A. (2023). Refined edge detection with cascaded and high-resolution convolutional network. *Pattern Recognition*, 138, 109361.
- Fini, E., Astolfi, P., Alahari, K., Alameda-Pineda, X., Mairal, J., Nabi, M., & Ricci, E. (2023). Semi-supervised learning made simple with self-supervised clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada.
- Gao, D., Li, P., Wang, M., Liang, Y., Liu, S., Zhou, J., Wang, L., & Zhang, Y. (2024). CSF-GTNet: A novel multi-dimensional feature fusion network based on Convnext-GeLU-BiLSTM for EEG-signals-enabled fatigue driving detection. *IEEE Journal of Biomedical and Health Informatics*, 28(5), 2558-2568.
- Gao, Z., Duan, Y., Yang, Y., Rui, L., & Zhao, C. (2022). FedHF: A High Fairness Federated Learning Algorithm Based on Deconfliction in Heterogeneous Networks. In *International Conference on Service-Oriented Computing* (pp. 553-566). Cham: Springer Nature Switzerland.
- Gong, Q., Kang, W., & Fahroo, F. (2023). Approximation of compositional functions with ReLU neural networks. *Systems & Control Letters*, 175, 105508.
- 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. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy.

- Graham, S., Vu, Q. D., Jahanifar, M., Raza, S. E. A., Minhas, F., Snead, D., & Rajpoot, N. (2023). One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83, 102685.
- Guo, Y., Zhang, Y., Chen, Y., Peng, W., Liu, X., Zhang, L., Huang, X., & Ma, Z. (2023). Membrane potential batch normalization for spiking neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France.
- Haji, S. H., & Abdulazeez, A. M. (2021). Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4), 2715-2743.
- Han, D., Pan, X., Han, Y., Song, S., & Huang, G. (2023). Flatten transformer: Vision transformer using focused linear attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France.
- Han, Q., Qian, X., Xu, H., Wu, K., Meng, L., Qiu, Z., Weng, T., Zhou, B., & Gao, X. (2024). DM-CNN: Dynamic Multi-scale Convolutional Neural Network with uncertainty quantification for medical image classification. *Computers in Biology and Medicine*, 168, 107758.
- Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, Palo Alto, California, USA.
- Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., & Xu, D. U. (2022). Transformers for 3D medical image segmentation. 2022 IEEE/CVF winter conference on applications of computer vision (WACV). *Waikoloa, HI, USA: IEEE*.
- He, J., Erfani, S., Ma, X., Bailey, J., Chi, Y., & Hua, X.-S. (2021). α -IoU: A family of power intersection over union losses for bounding box regression. *Advances in Neural Information Processing Systems*, 34, 20230-20242.
- Hicham, K., Laghmati, S., & Tmiri, A. (2022). Artificial intelligence for colorectal polyps classification using 3D CNN. *International Conference on Integrated Design and Production*, Paris, France.
- Hosain, A. A., Santhalingam, P. S., Pathak, P., Rangwala, H., & Kosecka, J. (2021). Hand pose guided 3D pooling for word-level sign language recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, USA.
- Hou, B., Liu, Y., Ling, N., Liu, L., & Ren, Y. (2021). A fast lightweight 3D separable convolutional neural network with multi-input multi-output for moving object detection. *IEEE Access*, 9, 148433-148448.

- Hu, Y., Xian, R., Wu, Q., Fan, Q., Yin, L., & Zhao, H. (2024). Revisiting scalarization in multi-task learning: A theoretical perspective. *Advances in Neural Information Processing Systems*, 36.
- Hua, B.-S., Tran, M.-K., & Yeung, S.-K. (2018). Pointwise convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA.
- Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2023). Normalization techniques in training dnns: Methodology, analysis and application. *IEEE transactions on pattern analysis and machine intelligence*, 45(8), 10173-10196.
- Huang, Q., Ding, H., & Razmjoooy, N. (2024). Oral cancer detection using convolutional neural network optimized by combined seagull optimization algorithm. *Biomedical Signal Processing and Control*, 87, 105546.
- Huang, S., Wang, K., Liu, H., Chen, J., & Li, Y. (2023). Contrastive semi-supervised learning for underwater image restoration via reliable bank. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada.
- Huang, W., Deng, Y., Hui, S., Wu, Y., Zhou, S., & Wang, J. (2024). Sparse self-attention transformer for image inpainting. *Pattern Recognition*, 145, 109897.
- Huang, Y., Lyu, J., Cheng, P., Tam, R., & Tang, X. (2024). SSiT: Saliency-guided self-supervised image transformer for diabetic retinopathy grading. *IEEE Journal of Biomedical and Health Informatics*, 28(5), 2806-2817.
- Huang, Z., Liang, M., Qin, J., Zhong, S., & Lin, L. (2023). Understanding Self-attention Mechanism via Dynamical System Perspective. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France.
- Ionescu, B., Müller, H., Péteri, R., Abacha, A. B., Sarrouiti, M., Demner-Fushman, D., Hasan, S. A., Kozlovski, S., Liauchuk, V., & Cid, Y. D. (2021). Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12,

- Ionescu, B., Müller, H., Péteri, R., Cid, Y. D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A. B., & Hasan, S. A. (2019). ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10,
- Ionescu, B., Müller, H., Péteri, R., Rückert, J., Abacha, A. B., de Herrera, A. G. S., Friedrich, C. M., Bloch, L., Brüngel, R., & Idrissi-Yaghir, A. (2022). Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications. *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2022)*, Bologna, Italy.
- 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. *Nature methods*, 18(2), 203-211.
- Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2023). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.
- Kang, J., Liu, R., Li, Y., Liu, Q., Wang, P., Zhang, Q., & Zhou, D. (2022). An improved 3D human pose estimation model based on temporal convolution with gaussian error linear units. 2022 8th International Conference on Virtual Reality (ICVR), Nanjing, China.
- Khan, F. H., Pasha, M. A., & Masud, S. (2023). Towards designing a hardware accelerator for 3D convolutional neural networks. *Computers and Electrical Engineering*, 105, 108489.
- Koloskova, A., Hendriks, H., & Stich, S. U. (2023). Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. *International Conference on Machine Learning, Honolulu, Hawaii, USA*.
- Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., & Müller, H. (2020). Overview of ImageCLEF tuberculosis 2020 automatic CT-based report generation. Proceedings of the CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece.
- Kraft, D. (2020). Computing the Hausdorff distance of two sets from their distance functions. *International Journal of Computational Geometry & Applications*, 30(01), 19-49.
- Kuang, Z., Yan, Z., & Yu, L. (2024). Weakly supervised learning for multi-class medical image segmentation via feature decomposition. *Computers in Biology and Medicine*, 171, 108228.

- Kweon, H., Yoon, S.-H., & Yoon, K.-J. (2023). Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada.
- 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. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, USA.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., & Zhang, S. (2024). One-shot weakly-supervised segmentation in 3D medical images. *IEEE Transactions on Medical Imaging*, 43(1), 175-189.
- Li, G., Zhang, M., Zhang, Q., & Lin, Z. (2022). Efficient binary 3D convolutional neural network and hardware accelerator. *Journal of Real-Time Image Processing*, 19(1), 61-71.
- Li, K., Qian, Z., Han, Y., Eric, I., Chang, C., Wei, B., Lai, M., Liao, J., Fan, Y., & Xu, Y. (2023). Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, 86, 102791.
- Li, R., Mai, Z., Zhang, Z., Jang, J., & Sanner, S. (2023). Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. *Journal of Visual Communication and Image Representation*, 92, 103800.
- Li, S., Li, T., Sun, C., Yan, R., & Chen, X. (2023). Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis. *Journal of Manufacturing Systems*, 69, 20-30.
- Li, X., Zhou, Y., Du, P., Lang, G., Xu, M., & Wu, W. (2021). A deep learning system that generates quantitative CT reports for diagnosing pulmonary tuberculosis. *Applied Intelligence*, 51, 4082-4093.
- Lin, L., Wu, J., Liu, Y., Wong, K. K., & Tang, X. (2023). Unifying and personalizing weakly-supervised federated medical image segmentation via adaptive representation and aggregation. *Proceedings of the International Workshop on Machine Learning in Medical Imaging (MLMI 2023)*, Held at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vancouver, Canada.

- Lin, X., Yu, L., Cheng, K.-T., & Yan, Z. (2023). BATFormer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(7), 3501–3512.
- Liu, B., Guo, J., Chen, C. P., Wu, X., & Zhang, T. (2024). Fine-grained interpretability for EEG emotion recognition: Concat-aided Grad-CAM and systematic brain functional network. *IEEE Transactions on Affective Computing*, 15(2), 671–684.
- Liu, B., Liu, X., Jin, X., Stone, P., & Liu, Q. (2021). Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34, 18878-18890.
- Liu, M., Chen, L., Du, X., Jin, L., & Shang, M. (2021). Activated gradients for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4), 2156-2168.
- Liu, S., Chen, Z., Liu, Y., Wang, Y., Yang, D., Zhao, Z., Zhou, Z., Yi, X., Li, W., & Zhang, W. (2023). Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France.
- Liu, Y., Lu, Y., Liu, H., An, Y., Xu, Z., Yao, Z., Zhang, B., Xiong, Z., & Gui, C. (2023). Hierarchical prompt learning for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada.
- Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., & He, Z. (2024). A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7478-7498.
- Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., & Dong, L. (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA.

- 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 Recognition*, 144, 109861.
- Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157, 106791.
- Miao, J., Wei, Y., Wang, X., & Yang, Y. (2023). Temporal pixel-level semantic understanding through the VSPW dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 11297-11308.
- Mikami, N., Ueki, Y., Shibahara, M., Aizawa, K., & Ara, K. (2024). CNN-based acoustic identification of gas–liquid jet: Evaluation of noise resistance and visual explanation using Grad-CAM. *International Journal of Multiphase Flow*, 171, 104688.
- Ming, Q., & Xiao, X. (2024). Towards accurate medical image segmentation with gradient-optimized dice loss. *IEEE Signal Processing Letters*, 31, 191-195.
- Nijati, M., Guo, L., Abulizi, A., Fan, S., Wubuli, A., Tursun, A., Nijati, P., Xia, L., Hong, K., & Zou, X. (2023). Deep learning and radiomics of longitudinal CT scans for early prediction of tuberculosis treatment outcomes. *European Journal of Radiology*, 169, 111180.
- Opfer, R., Krüger, J., Spies, L., Ostwaldt, A.-C., Kitzler, H. H., Schippling, S., & Buchert, R. (2023). Automatic segmentation of the thalamus using a massively trained 3D convolutional neural network: higher sensitivity for the detection of reduced thalamus volume by improved inter-scanner stability. *European radiology*, 33(3), 1852-1861.
- Pan, X., Wang, S., Liu, Y., Wen, L., & Lu, M. (2024). iPCa-Former: A multi-task transformer framework for perceiving incidental prostate cancer. *IEEE Signal Processing Letters*, 31, 785-789.
- Piao, M., & Jin, C. H. (2023). CNN and ensemble learning based wafer map failure pattern recognition based on local property based features. *Journal of Intelligent Manufacturing*, 34(8), 3599-3621.
- Raina, V., Molchanova, N., Graziani, M., Malinin, A., Muller, H., Cuadra, M. B., & Gales, M. (2023). Tackling bias in the dice similarity coefficient: Introducing nDSC for white matter lesion segmentation. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Cartagena, Colombia.
- Rao, J., Ke, A., Liu, G., & Ming, Y. (2023). MS-GAN: multi-scale GAN with parallel class activation maps for image reconstruction. *The Visual Computer*, 39(5), 2111-2126.

- Reus-Muns, G., Alemdar, K., Sanchez, S. G., Roy, D., & Chowdhury, K. R. (2023). AirFC: Designing Fully Connected Layers for Neural Networks with Wireless Signals. *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, Washington, D.C., USA.
- Rong, S., Tu, B., Wang, Z., & Li, J. (2023). Boundary-enhanced co-training for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada.
- Ru, L., Zhan, Y., Yu, B., & Du, B. (2022). Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA.
- Ru, L., Zheng, H., Zhan, Y., & Du, B. (2023). Token contrast for weakly-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada.
- Schaeffer, R., Khona, M., Ma, T., Eyzaguirre, C., Koyejo, S., & Fiete, I. (2024). Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems*, 36.
- 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. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy.
- Sharma, S. K., Mohan, A., & Sharma, A. (2016). Miliary tuberculosis: A new look at an old foe. *Journal of Clinical Tuberculosis and other mycobacterial diseases*, 3, 13-27.
- Shen, Y., Moore, R. H., & Deo, A. (2023). Visualizing Abaqus output database in ParaView: A universal converter in Python and C++. *SoftwareX*, 22, 101331.
- Song, D., Yao, J., Jiang, Y., Shi, S., Cui, C., Wang, L., Wang, L., Wu, H., Tian, H., & Ye, X. (2023). A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. *Computer Methods and Programs in Biomedicine*, 235, 107527.
- Song, L., Lin, J., Wang, Z. J., & Wang, H. (2020). An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2912-2921.

- Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., & Liu, L. (2021). Pixel difference networks for efficient edge detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada.
- Sun, W., Feng, X., Liu, J., & Ma, H. (2023). Weakly supervised segmentation of COVID-19 infection with local lesion coherence on CT images. *Biomedical Signal Processing and Control*, 79, 104099.
- 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA.
- 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 Computing Surveys (CSUR)*, 54(8), 1-34.
- Tremeau, A., & Borel, N. (1997). A region growing and merging algorithm to color segmentation. *Pattern Recognition*, 30(7), 1191-1203.
- Ullah, U., Saleem, S., Farooq, M., Yameen, B., & Cheema, M. I. (2024). Lipoarabinomannan-based tuberculosis diagnosis using a fiber cavity ring down biosensor. *Biomedical Optics Express*, 15(3), 1428-1436.
- Wan, S.-Y., & Higgins, W. E. (2003). Symmetric region growing. *IEEE Transactions on Image Processing*, 12(9), 1007-1015.
- Wang, G., Luo, X., Gu, R., Yang, S., Qu, Y., Zhai, S., Zhao, Q., Li, K., & Zhang, S. (2023). PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation. *Computer Methods and Programs in Biomedicine*, 231, 107398.
- Wang, H., Shao, Y., Zhou, S., Zhang, C., & Xiu, N. (2021). Support Vector Machine Classifier via $\frac{1}{2} L_0 + \frac{1}{2} L_1$ Soft-Margin Loss. *IEEE transactions on pattern analysis and machine intelligence*, 44(10), 7253-7265.
- Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA.
- Wang, J., Bhalerao, A., Yin, T., See, S., & He, Y. (2024). CAMANet: Class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4), 2199-2210.

- Wang, J., & Xia, B. (2024). Weakly supervised image segmentation beyond tight bounding box annotations. *Computers in Biology and Medicine*, 169, 107913.
- 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 Journal of Photogrammetry and Remote Sensing*, 190, 196-214.
- Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141, 61-67.
- Wang, S., & Zhang, Y. (2023). Grad-CAM: understanding AI models. *Comput. Mater. Contin*, 76, 1321-1324.
- Wang, Y., Li, X., Zhao, Y., & Zhang, J.-F. (2024). Threshold Selection and Resource Allocation for Quantized Identification. *Journal of Systems Science and Complexity*, 37(1), 204-229.
- Wang, Y., Shi, Q., & Chang, T.-H. (2023). Batch normalization damages federated learning on non-IID data: Analysis and remedy. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 1-5.
- Wang, Y., Yang, G., Li, S., Li, Y., He, L., & Liu, D. (2023). Arrhythmia classification algorithm based on multi-head self-attention mechanism. *Biomedical Signal Processing and Control*, 79, 104206.
- Wang, Z., & Voiculescu, I. (2023). Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels. *MICCAI Workshop on Data Engineering in Medical Imaging*, Vancouver, Canada.
- Wei, T., & Gan, K. (2023). Towards realistic long-tailed semi-supervised learning: Consistency is all you need. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada.
- Wei, X., Deng, Z., Zheng, X., He, B., & Hu, Y. (2024). Weakly supervised glottis segmentation on endoscopic images with point supervision. *Biomedical Signal Processing and Control*, 92, 106113.
- Wen, S., Liu, J., & Xu, W. (2021). A novel lesion segmentation algorithm based on U-net network for tuberculosis CT image. *2021 International Conference on Control, Automation and Information Sciences (ICCAIS)*, Xi'an, China.

- Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., & Jiangyun, L. (2021). Transbts: Multimodal brain tumor segmentation using transformer. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Strasbourg, France. Springer.
- Weszka, J. S. (1978). A survey of threshold selection techniques. *Computer graphics and Image processing*, 7(2), 259-265.
- Wollenberg, K. R., Jeffrey, B. M., Harris, M. A., Gabrielian, A., Hurt, D. E., & Rosenthal, A. (2022). Patterns of genomic interrelatedness of publicly available samples in the TB portals database. *Tuberculosis*, 133, 102171.
- Wu, W., Zhao, Y., Shou, M. Z., Zhou, H., & Shen, C. (2023). Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France.
- Wu, X., Tan, G., Luo, H., Chen, Z., Pu, B., Li, S., & Li, K. (2024). A knowledge-interpretable multi-task learning framework for automated thyroid nodule diagnosis in ultrasound videos. *Medical Image Analysis*, 91, 103039.
- Xiao, H., Li, L., Liu, Q., Zhu, X., & Zhang, Q. (2023). Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control*, 84, 104791.
- Xiao, Z., Zhang, X., Liu, Y., Geng, L., Wu, J., Wang, W., & Zhang, F. (2023). RNN-combined graph convolutional network with multi-feature fusion for tuberculosis cavity segmentation. *Signal, Image and Video Processing*, 17(5), 2297-2303.
- Xie, Y., Zhang, Z., Chen, S., & Qiu, C. (2023). Detect, Grow, Seg: A weakly supervision method for medical image segmentation based on bounding box. *Biomedical Signal Processing and Control*, 86, 105158.
- Xing, X., Li, X., Wei, C., Zhang, Z., Liu, O., Xie, S., Chen, H., Quan, S., Wang, C., & Yang, X. (2024). DP-GAN+ b: A lightweight generative adversarial network based on depthwise separable convolutions for generating CT volumes. *Computers in Biology and Medicine*, 108393.
- Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. (2020). On layer normalization in the transformer architecture. *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria.
- Xu, D., Qiu, H., Gao, L., Yang, Z., & Wang, D. (2022). A novel dual-stream self-attention neural network for remaining useful life estimation of mechanical systems. *Reliability Engineering & System Safety*, 222, 108444.

- Xu, L., Bennamoun, M., Boussaid, F., Ouyang, W., Sohel, F., & Xu, D. (2024). Auxiliary Tasks Enhanced Dual-Affinity Learning for Weakly Supervised Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, Advance online publication.
- Yamashita, H., Okamoto, T., Takashima, R., Ohtani, Y., Takiguchi, T., Toda, T., & Kawai, H. (2024). Fast neural speech waveform generative models with fully-connected layer-based upsampling. *IEEE Access*, 12, 31409-31421.
- Yan, B., Chang, X., Anastasopoulos, A., Fujita, Y., & Watanabe, S. (2024). Cross-Modal Multi-Tasking for Speech-to-Text Translation via Hard Parameter Sharing. *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea.
- Yang, E., Pan, J., Wang, X., Yu, H., Shen, L., Chen, X., Xiao, L., Jiang, J., & Guo, G. (2023). Adatask: A task-aware adaptive learning rate approach to multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence, Washington, D.C., USA*.
- Yang, C., Tan, Z., Wang, Y., Bi, R., Shi, T., Yang, J., Huang, C., Jiang, P., & Fu, X. (2024). SwinUNeCCT: bidirectional hash-based agent transformer for cervical cancer MRI image multi-task learning. *Scientific Reports*, 14(1), 24621.
- Ye, Q., Gao, Y., Ding, W., Niu, Z., Wang, C., Jiang, Y., Wang, M., Fang, E. F., Menpes-Smith, W., & Xia, J. (2022). Robust weakly supervised learning for COVID-19 recognition using multi-center CT images. *Applied Soft Computing*, 116, 108291.
- Ye, R., Liu, F., & Zhang, L. (2019). 3d depthwise convolution: Reducing model parameters in 3d vision tasks. *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*.
- Yin, S., Wang, L., Shafiq, M., Teng, L., Laghari, A. A., & Khan, M. F. (2023). G2Grad-CAMRL: an object detection and interpretation model based on gradient-weighted class activation mapping and reinforcement learning in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 3583-3598.
- Ying, J., Huang, W., Fu, L., Yang, H., & Cheng, J. (2023). Weakly supervised segmentation of uterus by scribble labeling on endometrial cancer MR images. *Computers in Biology and Medicine*, 167, 107582.

- Yoon, I., Hong, J. H., Witanto, J. N., Yim, J.-J., Kwak, N., Goo, J. M., & Yoon, S. H. (2023). Mycobacterial cavity on chest computed tomography: clinical implications and deep learning-based automatic detection with quantification. *Quantitative Imaging in Medicine and Surgery*, 13(2), 747.
- Yu, D., Lin, J., Cao, T., Chen, Y., Li, M., & Zhang, X. (2023). SECS: An effective CNN joint construction strategy for breast cancer histopathological image classification. *Journal of King Saud University-Computer and Information Sciences*, 35(2), 810-820.
- Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33, 5824-5836.
- Yu, X., Yang, Q., Zhou, Y., Cai, L. Y., Gao, R., Lee, H. H., Li, T., Bao, S., Xu, Z., & Lasko, T. A. (2023). UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, 90, 102939.
- Yuan, D., Chang, X., Liu, Q., Yang, Y., Wang, D., Shu, M., He, Z., & Shi, G. (2023). Active learning for deep visual tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10), 13284-13296.
- Yuan, F., Zhang, Z., & Fang, Z. (2023). An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognition*, 136, 109228.
- 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 (CVPR)*, New Orleans, LA, USA.
- Zhang, C., Lam, K.-M., Liu, T., Chan, Y.-L., & Wang, Q. (2024). Structured adversarial self-supervised learning for robust object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, Article 5613720.
- Zhang, J., Li, X., Tian, J., Luo, H., & Yin, S. (2023). An integrated multi-head dual sparse self-attention network for remaining useful life prediction. *Reliability Engineering & System Safety*, 233, 109096.
- Zhang, L., Yang, Q., Liu, X., & Guan, H. (2022). Rethinking hard-parameter sharing in multi-domain learning. 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan.
- Zhang, L., Yang, Q., Liu, X., & Guan, H. (2023). An Alternative Hard-Parameter Sharing Paradigm for Multi-Domain Learning. *IEEE Access*, 11, 10440-10452.

- Zhang, T., Wei, D., Zhu, M., Gu, S., & Zheng, Y. (2024). Self-supervised learning for medical image data with anatomy-oriented imaging planes. *Medical Image Analysis*, 103151.
- Zhang, X., Dong, X., Saripan, M. I. b., Du, D., Wu, Y., Wang, Z., Cao, Z., Wen, D., Liu, Y., & Marhaban, M. H. (2023). Deep learning PET/CT - based radiomics integrates clinical data: A feasibility study to distinguish between tuberculosis nodules and lung cancer. *Thoracic Cancer*, 14(19), 1802-1811.
- Zhao, D., Liu, Y., Yin, H., & Wang, Z. (2023). An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image. *Expert Systems with Applications*, 211, 118672.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2921-2929.
- Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., & Yu, Y. (2023). nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32, 4036-4045.
- Zhou, M., Xu, Z., Zhou, K., & Tong, R. K.-y. (2023). Weakly Supervised Medical Image Segmentation via Superpixel-Guided Scribble Walking and Class-Wise Contrastive Regularization. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Vancouver, Canada.
- Zhou, Q., Huang, Z., Ding, M., & Zhang, X. (2023). Medical image classification using light-weight CNN with spiking cortical model based attention module. *IEEE Journal of Biomedical and Health Informatics*, 27(4), 1991-2002.
- Zhou, S., Zhang, W., Jiang, J., Zhong, W., Gu, J., & Zhu, W. (2022). On the convergence of stochastic multi-objective gradient manipulation and beyond. *Advances in Neural Information Processing Systems*, 35, 38103-38115.
- Zhou, Y.-q., Son, G.-h., Shi, Y.-q., Yu, Y.-j., Li, M.-y., Zhang, Q., Zou, D.-h., Zhang, Z.-y., Yang, C., & Wang, S.-y. (2021). Quantitative segmentation analysis of the radiological changes by using ITK-SNAP: Risk assessment of the severity and recurrence of medication-related osteonecrosis of the jaw. *International Journal of Medical Sciences*, 18(10), 2209.
- Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.-T., & Shen, D. (2021). Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Medical Image Analysis*, 70, 101918.

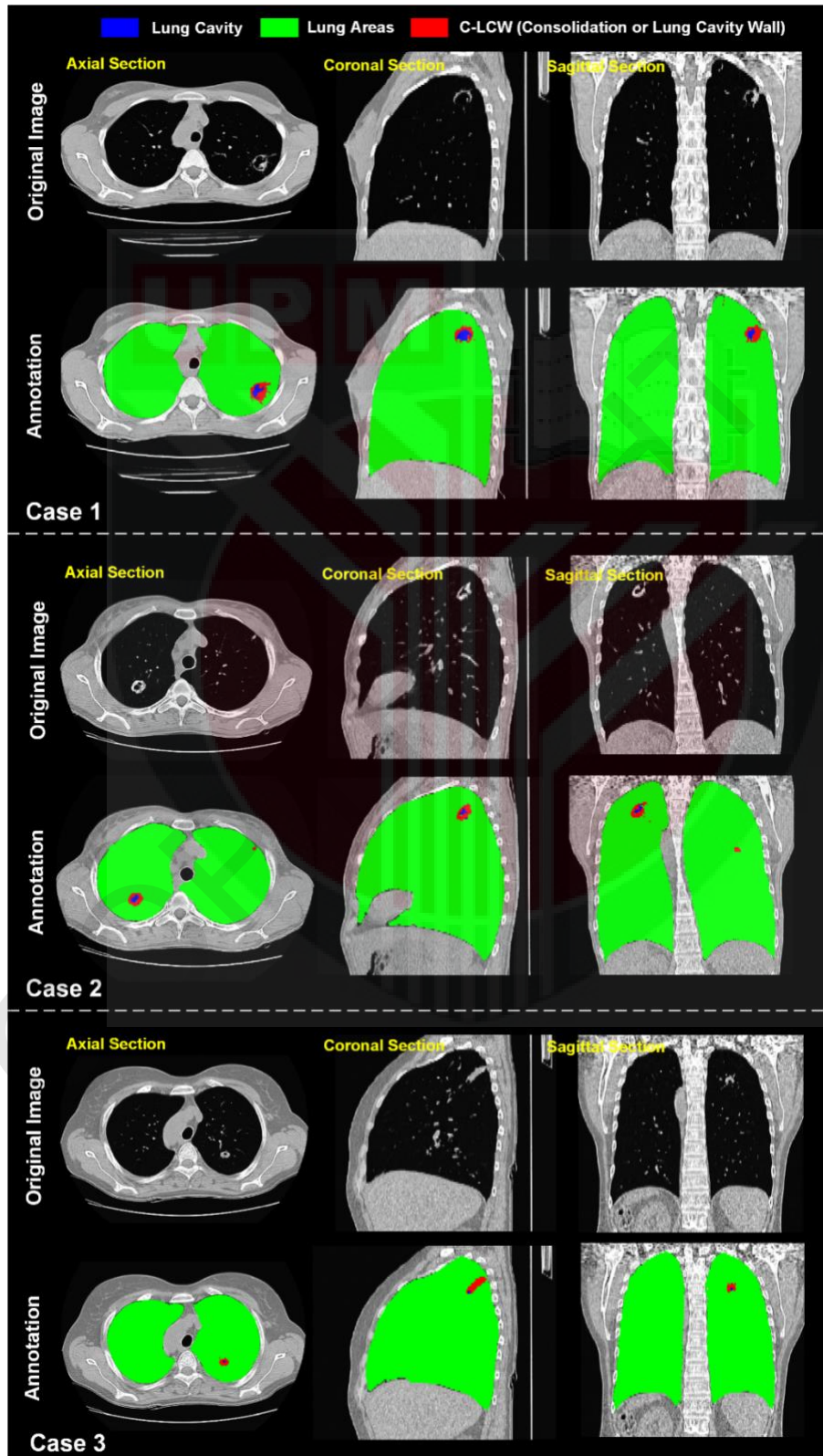
Zhu, J., Wang, C., Shen, X., Zhang, Y., & Qiu, T. (2023). A Discriminant Face Recognition Algorithm Based on Improved Softmax Loss Algorithm. *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)*, Chengdu, China.

Zhu, M., Chen, Z., & Yuan, Y. (2023). FedDM: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting. *IEEE Transactions on Medical Imaging*, 42(6), 1632-1643.



APPENDICES

Some examples of DeepPulmoTB data set:



BIODATA OF STUDENT

Tan Zhuoyi received the B.S. degree from Guangzhou University, Guangzhou, China, in 2018, and the M.S. degree from the Information Institute, Guangdong University of Finance and Economics, in 2021. He is currently pursuing the Ph.D. degree with Universiti Putra Malaysia. His research interests include medical imaging, pattern recognition, and deep learning.



LIST OF PUBLICATIONS

- 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*.
- Tan, Z., Madzin, H., Norafida, B., Rahmat, R. W. O., Khalid, F., & Sulaiman, P. S. (2024). SwinUNeLCST: Global-local spatial representation learning with hybrid CNN-transformer for efficient tuberculosis lung cavity weakly supervised semantic segmentation. *Journal of King Saud University-Computer and Information Sciences*, 102012.
- Tan, Z., Madzin, H., & Ding, Z. (2023). Semi-supervised Semantic Segmentation Methods for UW-OCTA Diabetic Retinopathy Grade Assessment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 25th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2022, September 18, 2022 - September 22, 2022, Singapore, Singapore.
- Tan, Z., Madzin, H., & Ding, Z. (2023). Image Quality Assessment Based on Multi-model Ensemble Class-Imbalance Repair Algorithm for Diabetic Retinopathy UW-OCTA Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 25th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2022, September 18, 2022 - September 22, 2022, Singapore, Singapore.
- Tan, Z., Madzin, H., Khalid, F., & Beng, N. S. (2023). Segmentation of Pulmonary Cavity in Lung CT Scan for Tuberculosis Disease. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 33(2), 98-106.