Chest-OMDL: Organ-specific Multidisease Detection and Localization in Chest Computed Tomography using Weakly Supervised Deep Learning from Free-text Radiology Report

Xuguang Bai[*1] BXG21@MAILS.TSINGHUA.EDU.CN
Mingxuan Liu[*1] ARKTIS@QQ.COM
Yifei Chen[1] JUSTLFC03@GMAIL.COM
Hongjia Yang[1] YANGHJ23@MAILS.TSINGHUA.EDU.CN
Qiyuan Tian[†1] QIYUANTIAN@TSINGHUA.EDU.CN
[1] *School of Biomedical Engineering, Tsinghua University, Beijing, China*

Editors: Under Review for MIDL 2025

Abstract

Deep learning (DL) models designed to detect abnormalities in chest computed tomography (CT) reduce radiologists' workload. However, training multidisease diagnostic models requires large expert-annotated datasets, significantly increasing model development cost. To address this challenge, we propose a weakly supervised learning (WSL) framework entitled Chest-OMDL for Organ-specific Multidisease Detection and Localization in chest CT. Chest-OMDL trains DL models using disease labels extracted by RadBERT from free-text radiology reports and multi-organ segmentation masks generated by the Segment Anything by Text (SAT) model, therefore reducing the need for manual annotation. Specifically, Chest-OMDL employs a Y-shaped Mamba model (Y-Mamba), comprising a feature extractor, an organ segmentation decoder, and a disease anomaly map generator. By incorporating multidisease anatomical knowledge, Y-Mamba is trained with a multi-task loss for organ-level weak supervision. Chest-OMDL was trained and validated on the large-scale CT-RATE dataset (25,692 non-contrast 3D chest CT scans from 21,304 patients) and tested on the external RAD-ChestCT dataset (3,630 scans), outperforming CT-CLIP (contrastive language-image pre-training) and CT-Net (full supervision). The proposed Chest-OMDL can be applied to multiple anatomical sites, potentially streamlining diagnostics.

Keywords: Radiology report, Chest computed tomography, Weakly supervised learning, Multidisease detection

1. Introduction

Chest computed tomography (CT) is widely used in clinical practice for diagnosing causes of signs or symptoms of chest diseases, such as cough, shortness of breath, chest pain, or fever. CT played a crucial role in the fight against COVID-19 (Ma et al., 2021; Carter et al., 2020). However, each 3D chest CT volume comprises millions of voxels and exhibits significant variations in individual characteristics and imaging conditions, making it cumbersome for radiologists to examine CT volumes slice by slice (Khanna et al., 2020)). To support clinical diagnosis and decision-making while reducing radiologists' workload, many deep learning (DL) models have been developed to assist in CT interpretation and received FDA approvals.

[*] Contributed equally
[†] Corresponding author

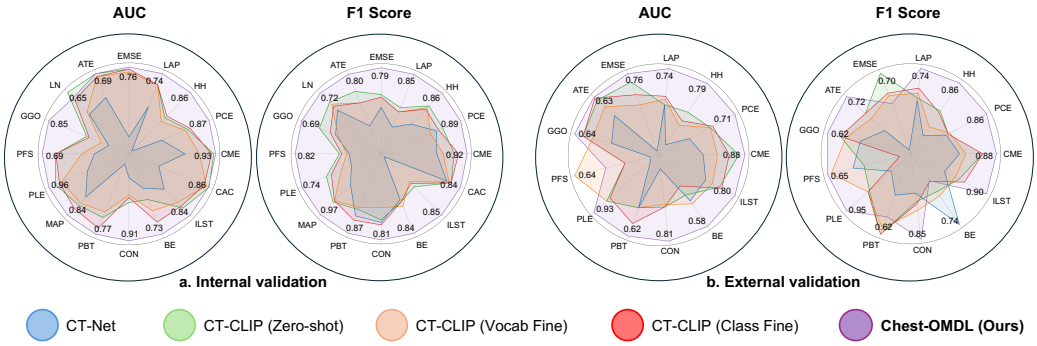

Figure 1: AUROC and F1 scores of Chest-OMDL and comparative methods on internal and external validation datasets. Each dimension of the radar chart represents a specific disease (Abbreviations are defined in Appendix B, detailed results of different methods are provided in Appendix E).

As of 2023, 190 FDA-approved AI-based radiology software devices exist, with 42 of them specifically developed for thoracic radiology (Milam and Koo, 2023).

However, most existing DL models are trained on limited datasets and target only a single disease, raising concerns about the generalizability and robustness of these models in clinical practice (Yang et al., 2024; Chen et al., 2024). Achieving generalizable multidisease detection and lesion localization in chest CT requires large labeled datasets, which are difficult to obtain because annotating medical images is time-consuming and cost-intensive (Rajpurkar et al., 2022; Liu et al., 2024, 2023b; Cao et al., 2022, 2024; Xu et al., 2022; Huang et al., 2023; Fang et al., 2024).

Because ample training data already exist in electronic health records, an alternative approach is to extract information from more accessible free-text radiology reports to train DL models. This primarily involves training contrastive language-image pretraining (CLIP) frameworks (Radford et al., 2021) on large-scale paired image-text datasets for zero-shot classification or applying natural language processing (NLP) techniques (Bergomi et al., 2024) to extract classification labels from text reports for supervised learning. A representative study of the former is CT-CLIP (Hamamci et al., 2024), which was trained with CT-RATE, a large-scale dataset comprising 3D chest CT scans and paired text reports. Hamamci et al. demonstrated that, in multi-abnormality detection, CT-CLIP outperformed state-of-the-art (SOTA) fully supervised models across all key metrics. A notable example of the latter approach is the multidisease classifiers for body CT scans developed by Tushar et al., designed for three different organ systems using automatically extracted labels from radiology text reports (Tushar et al., 2021). Their main contribution involved employing rule-based algorithms to extract 19,225 disease labels from 13,667 body CT scans. Furthermore, Sato et al. (Sato et al., 2024) recently developed a DL-based pipeline to detect abnormalities in the liver, gallbladder, pancreas, spleen, and kidneys, also leveraging information from free-text radiology reports rather than manual annotations. However, the common limitation of the aforementioned methods is that they only perform disease classification while cannot localize abnormal locations, results in poor interpretability for clinical

use. Liu et al. proposed a cross-modality learning framework Cross-DL (Liu et al., 2023a) for detecting four abnormality types across 17 regions in head CT with voxel-level localization. But applying this method to chest CT is challenging due to its large 3D coverage, millions of voxels, and significant variability in imaging conditions.

In this study, to effectively utilize information from free-text radiology reports and achieve simultaneous organ-specific multidisease detection and localization, we propose a novel weakly-supervised learning framework Chest-OMDL for chest CT. Specifically, Chest-OMDL leverages classification labels extracted by RadBERT (Yan et al., 2022) from text reports and segmentation masks generated by the Segment Anything by Text (SAT) model (Zhao et al., 2024) from CT images as weak supervision to train a Y-shaped mamba model (Y-Mamba). The Y-Mamba consists of a feature extractor, an organ segmentation decoder, and a disease anomaly map generator, producing organ segmentation results and lesion heatmaps for chest CT. By incorporating anatomical prior knowledge of each disease during training, the trained Y-Mamba generates interpretable pixel-level lesion localization.

We trained Chest-OMDL on CT-RATE (Hamamci et al., 2024), the largest publicly available chest CT dataset, and compared it with existing methods on both an internal validation set and an external validation set (RAD-ChestCT) (Draelos et al., 2021). Chest-OMDL achieved SOTA performance on multidisease classification tasks across 9 organs. The radar plot in Fig. 1 visually compares the classification AUROC and F1-score of various models. Furthermore, we quantitatively evaluated the localization performance of Chest-OMDL on an external COVID-19 CT dataset (Ma et al., 2020). Despite relying only on organ-level weak supervision, the model achieved a pixel-level segmentation Dice Similarity Coefficient (DSC) of 0.450.

2. Materials and Methods

2.1. Training and Internal Validation Datasets

We utilized the recently curated and open-sourced CT-RATE dataset by Hamamci et al. (Hamamci et al., 2024) as the training and internal validation dataset, which includes 25,692 non-contrast 3D chest CT volumes along with paired radiology text reports from 21,304 individual patients from Istanbul Medipol University Mega Hospital. The CT-RATE dataset is divided into two groups: 20,000 patients (24,128 volumes) for training and 1,304 (1,564 volumes) for validation. CT volumes are stored in multiple matrix sizes (65.4% at 512×512 pixels, 4.2% at 768×768 pixels, 30.4% at 1024×1024 pixels). The pixel spacing in the axial (XY) plane ranges from 0.227 to 1.416 mm, with a mean of 0.605 mm. The slice thickness varies from 0.035 to 6 mm, with a mean of 1.231 mm. All CT volumes were first resized to 128×128×64 voxels, followed by windowing and histogram equalization for image enhancement.

2.2. External Validation Datasets

To evaluate the performance of different methods on out-of-distribution (OoD) data, we utilized the RAD-ChestCT (Draelos et al., 2021) misaligned external validation dataset, which includes 3,630 non-contrast chest CT volumes uniformly reconstructed using a single technique from the Duke University Health System. have a matrix size of 512×512 pixels,

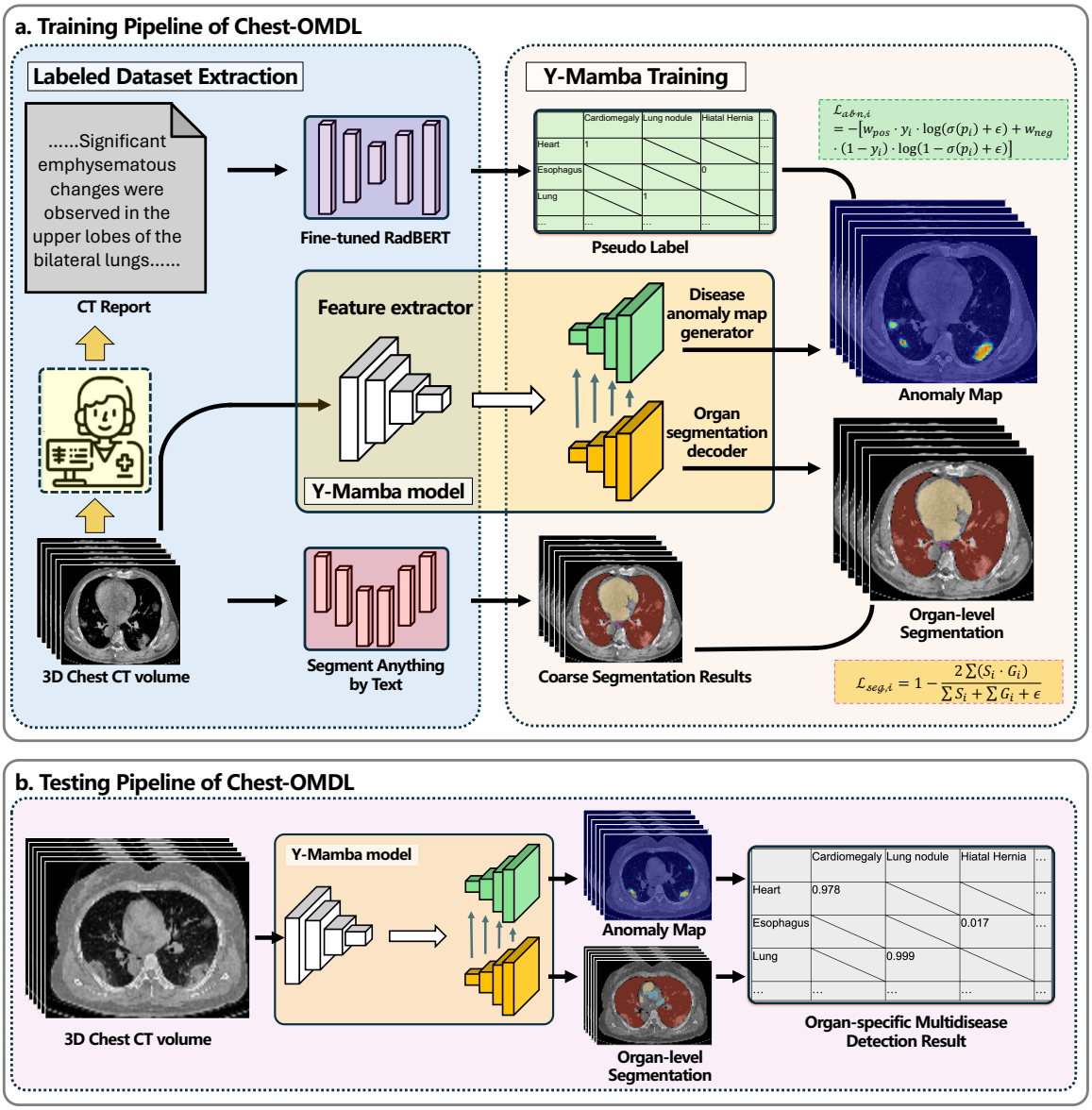

Figure 2: Overview of the proposed Chest-OMDL pipeline.

with axial (XY) pixel spacing ranging from 0.189 to 0.977 mm (mean: 0.692 mm) and slice thickness varying between 0.125 and 5 mm (mean: 0.706 mm). Since neither CT-RATE nor RAD-ChestCT includes lesion segmentation labels, we further evaluated the localization ability of Chest-OMDL using 10 labeled COVID-19 CT scans from the external COVID-19 CT dataset (Ma et al., 2020), where the infections were annotated by two radiologists and verified by an experienced radiologist. To ensure consistent evaluation, we apply identical preprocessing methods to the external datasets as used with CT-RATE.

2.3. Chest-OMDL Pipeline

The overall pipeline of the proposed Chest-OMDL is illustrated in Figure 2. The training process consists of two main steps: (1) Labeled Dataset Extraction: Automatically extracting meaningful supervisory information from both text reports and 3D CT images. (2) Y-Mamba Training: Training the Y-Mamba model using a multi-task loss function.

Labeled Dataset Extraction: Chest-OMDL assumes that a sufficiently large dataset can make the model robust to noisy labels (Rolnick et al., 2017; Karimi et al., 2020), enabling the direct use of outputs from existing automated methods as training labels. To extract disease labels from CT reports, pre-trained language models can be fine-tuned. In constructing the CT-RATE dataset, Hamamci et al. utilized the RadBERT-RoBERTa-4m model (Yan et al., 2022) to identify disease labels from free-text reports, which we adopt in this study. These labels are linked to six specific organs, creating a tabular pseudo-label for each subject (detailed associations between diseases and organs are in Appendix B). For segmentation labels, we use the recent Segment Anything by Text (SAT) model (Zhao et al., 2024), a knowledge-enhanced approach leveraging natural language prompts to segment 3D medical volumes. Zhang et al. extended CT-RATE by introducing RadGenome-Chest CT (Zhang et al., 2024), which includes SAT-based segmentation results. We directly use this dataset, retaining segmentation masks for six disease-related organs (Lung, Trachea and Bronchie, Pleura, Mediastinum, Heart, Esophagus) out of the nine available regions.

Y-Mamba Training: To achieve multidisease classification and pixel-level lesion segmentation under the weak supervision of organ-level disease localization, we constructed a Y-Mamba model based on the architecture of SegMamba (Xing et al., 2024) (details of the structure are provided in Appendix A). This model simultaneously predicts a segmentation mask for each anatomical organ (from organ segmentation decoder) and an anomaly map for each disease (from disease anomaly map generator). We constructed a multi-task loss function for training Y-Mamba, starting with the Dice loss for the segmentation task:

$$\mathcal{L}_{seg,i} = 1 - \frac{2\sum(S_i \cdot G_i)}{\sum S_i + \sum G_i + \epsilon} \tag{1}$$

where S_i is the predicted segmentation, G_i is the coarse segmentation of the i-th organ generated using SAT, and $\sum(S_i \cdot G_i)$ represents the overlapping pixel count. The totals $\sum S_i$ and $\sum G_i$ correspond to the predicted and ground truth pixel counts, respectively, while ϵ is a small constant to prevent division by zero. We further define the abnormality detection loss as:

$$\mathcal{L}_{abn,i} = -\left[w_{pos} \cdot y_i \cdot \log(\sigma(p_i) + \epsilon) + w_{neg} \cdot (1 - y_i) \cdot \log(1 - \sigma(p_i) + \epsilon)\right] \tag{2}$$

where p_i represents the model's predicted probability for the i-th disease, obtained by averaging the top-k values (with $k = 24$) after element-wise multiplication of the anomaly map and the segmentation mask of the specific organ. y_i is the corresponding ground truth (1 for positive and 0 for negative samples). The terms w_{pos} and w_{neg} denote the positive and negative sample weights, respectively, which are derived based on the disease frequency to mitigate biases caused by imbalanced datasets. The sigmoid function $\sigma(p_i)$ ensures that predictions are mapped onto a probability space, and ϵ serves as a numerical stability constant.

The two loss functions are combined using dynamically decreasing weights λ:

$$\mathcal{L}_{total} = \lambda \sum_i \mathcal{L}_{seg,i} + \sum_i \mathcal{L}_{abn,i} \tag{3}$$

During testing (Fig. 2b), the Y-Mamba model produces segmentation masks for organs and anomaly maps for diseases. By applying the element-wise multiplication of segmentation maps and anomaly maps, as described in Appendix B, the mean of the top-k values is computed as the anomaly score for each disease, consistent with the training phase. Final case-level classifications are based on thresholds, while binarized anomaly maps enable pixel-level segmentation.

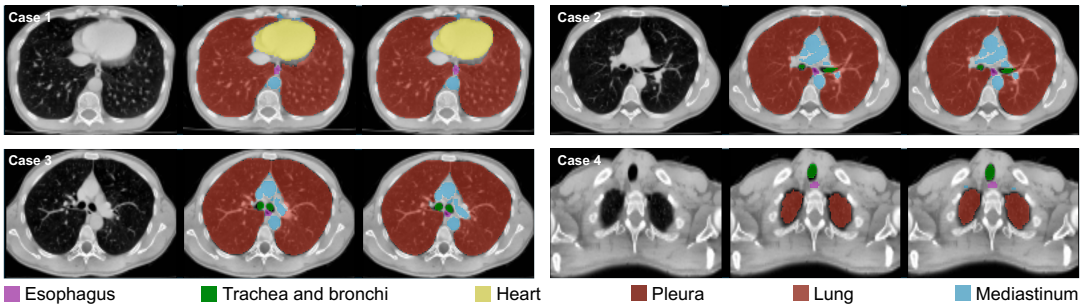

Figure 3: Representative slice-level examples of multi-organ segmentation.

3. Experiments and Results

3.1. Training Setup

A total of 22,620 CT volumes from the CT-RATE training dataset were used for training, while the remaining 1,508 volumes were reserved for early stopping and threshold selection. The model with the highest average AUROC on these 1,508 volumes during training was selected, and the threshold was determined using the Jaccard Index. The entire framework was optimized with the AdamW (Loshchilov and Hutter, 2017) optimizer using four NVIDIA A800 GPUs (requiring 2 days 20 hours per training). The number of training epochs was set to 150, with a batch size of 8. The loss function weights were dynamically adjusted during training according to: $\lambda = \max\left(\text{initial\_weight} \cdot e^{-\frac{\text{decay\_rate}}{\text{total\_epochs}} \cdot \text{epoch}}, 0.5\right)$, ensuring gradual decay over the course of the training process. The disease distribution across all datasets is detailed in Appendix C.

3.2. Multi-organ Segmentation Results

Table 1: Organ segmentation DSC metric on the internal validation dataset. Each metric is presented as mean ± standard deviation (SD).

Organ	Lung	Trachea and Bronchi	Pleura	Mediastinum	Heart	Esophagus
Dice	0.9697±0.0872	0.9002±0.0906	0.9697±0.0872	0.8499±0.0867	0.9110±0.1095	0.8515±0.0959

Accurate segmentation is crucial for Chest-OMDL to identify effective regions in anomaly maps for multiple diseases. The DSC for six organs on the internal validation set are presented in Table 1. Representative slice-level examples are shown in Fig. 3. From left to right (for each case): the input, the coarse segmentation label generated by the SAT model, and the segmentation mask output by Chest-OMDL. For all organs, the average DSC exceeds 84%, with the lungs and pleura achieving the highest DSC values (both at 0.9697±0.0872), likely due to their highly overlapping anatomical structures. In contrast, the mediastinum has the lowest DSC (0.8499±0.0867), as its boundaries are indistinct and its intensity falls within the range of normal soft tissue distributions.

3.3. Organ-specific Multidisease Detection Performance of Different Methods

Table 2: The performance of multidisease classification is evaluated on both internal and external validation sets using three key metrics: AUROC, accuracy, and F1 score.

Dataset	Numbers of scans	Metric	CT-Net	CT-CLIP (Zero-shot)	CT-CLIP (VocabFine)	CT-CLIP (ClassFine)	**Chest-OMDL (ours)**
Internal valid (CT-RATE)	1564	AUROC	0.628	0.723	0.749	0.751	**0.807**
		F1 score	0.664	0.701	0.729	0.720	**0.828**
		Accuracy	0.613	0.662	0.696	0.684	**0.754**
Enternal valid (RAD-ChestCT)	3630	AUROC	0.549	0.624	0.651	0.643	**0.720**
		F1 score	0.594	0.641	0.665	0.655	**0.723**
		Accuracy	0.543	0.591	0.617	0.610	**0.659**

Comparative Methods: We compare Chest-OMDL with four SOTA methods that also utilize information extracted from radiology reports to train DL models.: (1) CT-Net (Draelos et al., 2021), a fully supervised traditional classification model trained directly with disease labels; (2) CT-CLIP (zero-shot) (Hamamci et al., 2024), a visual-language foundation model based on CLIP that automatically learns semantic knowledge through contrastive learning; (3) CT-CLIP (VocabFine) and (4) CT-CLIP (ClassFine), two variants of CT-CLIP fine-tuned using different methods specifically for disease classification tasks. These comparative models were trained and validated on the same dataset, and the experimental results from Hamamcı et al.'s study (Hamamci et al., 2024) are directly used for comparison in this work.

Validation Results: The average AUROC, F1 score, and accuracy for detecting 16 diseases (internal validation) and 13 diseases (external validation) are summarized in Table 2. Chest-OMDL shows significant improvements, with higher mean AUROC (+7.74% internal, +10.60% external), mean F1 score (+13.58% internal, +8.72% external), and mean accuracy (+8.33% internal, +6.81% external) compared to CT-CLIP (VocabFine). While its performance on the external dataset is lower than internal dataset, similar to other methods, Chest-OMDL maintains an AUROC above 0.7, comparable to the in-distribution validation performance of comparison methods, demonstrating strong generalization. ROC curves of Chest-OMDL for each disease are in Appendix D, and detailed results for all methods are in Appendix E.

4. Lesion Localization Results

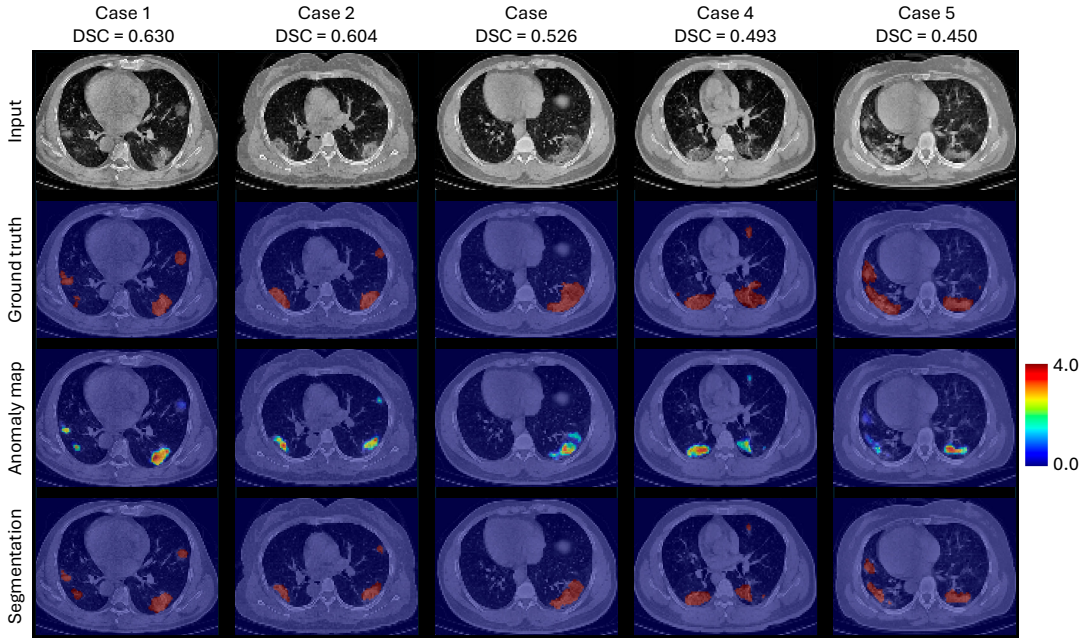

Figure 4: Abnormality localization results on the Covid-19 CT dataset.

We evaluated our model's localization on 10 annotated Covid-19 CT cases (Ma et al., 2020). Covid-19 manifests in CT scans as lung-specific abnormalities like ground-glass opacity (GGO), pulmonary fibrotic sequela (PFS), and consolidation (CON) (Mumoli et al., 2021). To segment these lesions, we overlaid the Chest-OMDL anomaly maps for lung-related diseases (as shown in Appendix B) and used a binarization threshold of 0.1.

Figure 4 shows anomaly maps and segmentation results for the top five DSC cases, with lower-scoring cases in Appendix F. Our method achieved an average DSC of 0.450, compared to 0.673 ± 0.223 (Ma et al., 2021) reported by Ma et al. using a supervised approach. This shows that Chest-OMDL, despite relying solely on organ-level weak supervision during training, achieved segmentation accuracy comparable to supervised methods.

5. Conclusion

We propose Chest-OMDL, a weakly supervised framework for disease detection and localization in chest CT scans. It uses an automated method to extract pseudo-labels from radiology reports and anatomical priors from images to train the Y-Mamba model. Experiments on the CT-RATE and RAD-ChestCT datasets show Chest-OMDL surpasses CT-Net and CT-CLIP, while tests on a small Covid-19 CT dataset demonstrate its ability for pixel-level segmentation. By eliminating manual annotations, it reduces costs and streamlines learning, making it effective for diagnosing chest-related, organ-specific diseases. Future work aims to develop Chest-OMDL into a foundation model for tasks such as radiology report generation and fine-grained lesion segmentation.

Acknowledgments

Funding was provided by the Tsinghua University Startup Fund.

References

Laura Bergomi, Tommaso M Buonocore, Paolo Antonazzo, Lorenzo Alberghi, Riccardo Bellazzi, Lorenzo Preda, Chandra Bortolotto, and Enea Parimbelli. Reshaping free-text radiology notes into structured reports with generative transformers. *arXiv preprint arXiv:2403.18938*, 2024.

Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022.

Yunkang Cao, Xiaohao Xu, Chen Sun, Liang Gao, and Weiming Shen. Bias: Incorporating biased knowledge to boost unsupervised image anomaly localization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(4):2342–2353, 2024. doi: 10.1109/TSMC.2023.3344383.

Brett W Carter, Melissa L Rosado-de Christenson, John P Lichtenberger III, and Santiago Martínez-Jiménez. *ExpertDDx: Chest*. Elsevier Health Sciences, 2020.

Yifei Chen, Binfeng Zou, Zhaoxin Guo, Yiyu Huang, Yifan Huang, Feiwei Qin, Qinhai Li, and Changmiao Wang. Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7759–7767, 2024.

Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021.

Zhaojie Fang, Shenghao Zhu, Yifei Chen, Binfeng Zou, Fan Jia, Linwei Qiu, Chang Liu, Yiyu Huang, Xiang Feng, Feiwei Qin, et al. Gfe-mamba: Mamba-based ad multimodal progression assessment via generative feature extraction from mci. *arXiv preprint arXiv:2407.15719*, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *PREPRINT (Version 1) available at Research Square*, October 2024. doi: 10.21203/rs.3.rs-5271327/v1. URL https://doi.org/10.21203/rs.3.rs-5271327/v1.

Yiming Huang, Guole Liu, Yaoru Luo, and Ge Yang. Adfa: Attention-augmented differentiable top-k feature adaptation for unsupervised medical anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 206–210, 2023. doi: 10.1109/ICIP49359.2023.10222528.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

Anita Khanna, Narendra Digambar Londhe, and Shubhrata Gupta. Automated localized approach for airway segmentation in 3d chest ct volume. *Biomedical and Pharmacology Journal*, 13(4):1671–1682, 2020.

Aohan Liu, Yuchen Guo, Jinhao Lyu, Jing Xie, Feng Xu, Xin Lou, Jun-hai Yong, and Qionghai Dai. Automatic intracranial abnormality detection and localization in head ct scans by learning from free-text reports. *Cell Reports Medicine*, 4(9), 2023a.

Mingxuan Liu, Yunrui Jiao, and Hong Chen. Skip-st: Anomaly detection for medical images using student-teacher network with skip connections. In *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2023b.

Mingxuan Liu, Yunrui Jiao, Jingqiao Lu, and Hong Chen. Anomaly detection for medical images using teacher–student model with skip connections and multiscale anomaly consistency. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024. doi: 10.1109/TIM.2024.3406792.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Jun Ma, Cheng Ge, Yixin Wang, Xingle An, Jiantao Gao, Ziqi Yu, Minqing Zhang, Xin Liu, Xueyuan Deng, Shucheng Cao, Hao Wei, Sen Mei, Xiaoyu Yang, Ziwei Nie, Chen Li, Lu Tian, Yuntao Zhu, Qiongjie Zhu, Guoqiang Dong, and Jian He. Covid-19 ct lung and infection segmentation dataset (version 1.0), 2020. URL https://doi.org/10.5281/zenodo.3757476. [Data set].

Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021.

ME Milam and CW Koo. The current status and future of fda-approved artificial intelligence tools in chest radiology in the united states. *Clinical Radiology*, 78(2):115–122, 2023.

Nicola Mumoli, Aldo Bonaventura, Alessandra Colombo, Alessandra Vecchié, Marco Cei, José Vitale, Luca Pavan, Antonino Mazzone, and Francesco Dentali. Lung function and symptoms in post–covid-19 patients: A single-center experience. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 5(5):907–915, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *ArXiv*, abs/1705.10694, 2017. URL https://api.semanticscholar.org/CorpusID:3511978.

Junya Sato, Kento Sugimoto, Yuki Suzuki, Tomohiro Wataya, Kosuke Kita, Daiki Nishigaki, Miyuki Tomiyama, Yu Hiraoka, Masatoshi Hori, Toshihiro Takeda, Shoji Kido, and Noriyuki Tomiyama. Annotation-free multi-organ anomaly detection in abdominal ct using free-text radiology reports: a multi-centre retrospective study. *eBioMedicine*, 110:105463, 2024. ISSN 2352-3964. doi: 10.1016/j.ebiom.2024.105463. URL https://www.sciencedirect.com/science/article/pii/S2352396424004997.

Fakrul Islam Tushar, Vincent M D'Anniballe, Rui Hou, Maciej A Mazurowski, Wanyi Fu, Ehsan Samei, Geoffrey D Rubin, and Joseph Y Lo. Classification of multiple diseases on body ct scans using weakly supervised deep learning. *Radiology: Artificial Intelligence*, 4(1):e210026, 2021.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

Rui Xu, Yufeng Wang, Xinchen Ye, Pengcheng Wu, Yen-Wei Chen, Fangyi Xu, Wenchao Zhu, Chao Chen, Yong Zhou, Hongjie Hu, Xiaofeng Qu, Shoji Kido, and Noriyuki Tomiyama. Pixel-level and affinity-level knowledge distillation for unsupervised segmentation of covid-19 lesions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1376–1380, 2022. doi: 10.1109/ICASSP43922.2022.9746715.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10): 2838, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts, 2024. URL https://arxiv.org/abs/2312.17183.

Appendix A. The Architecture of the Proposed Y-Mamba

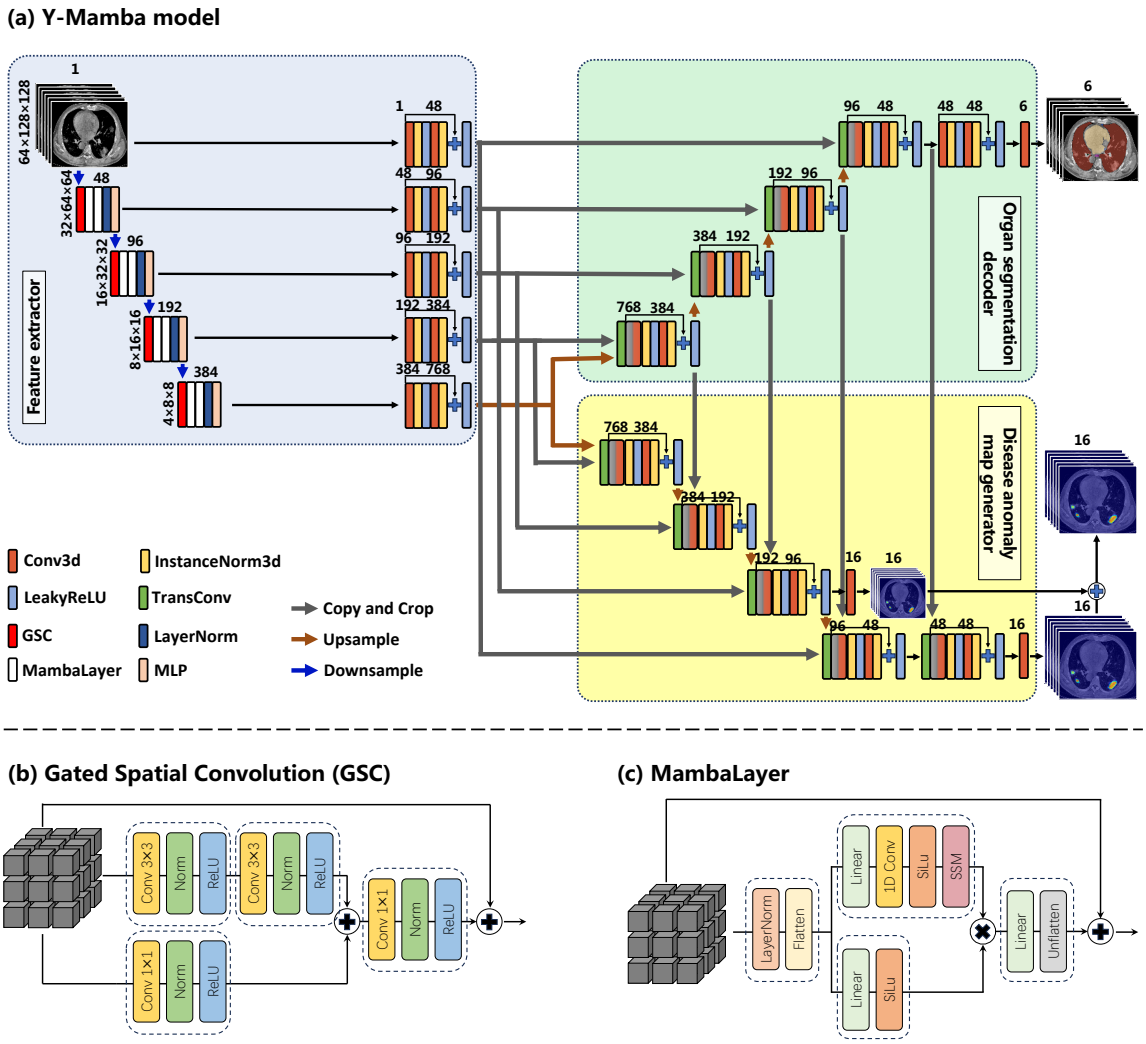

Figure 5: An overview of the proposed Y-Mamba model.

Our proposed Y-Mamba (shown in Fig. 5) is a unified deep learning model designed for organ segmentation and disease anomaly detection in medical imaging. The architecture consists of three main components: a feature extractor, an organ segmentation decoder, and a disease anomaly map generator. The feature extractor encodes multi-scale representations of the input medical images, which are then utilized by the two decoders to perform their respective tasks.

Feature Extractor: The feature extraction module adopts a multi-resolution encoding strategy, leveraging 3D convolutional layers (Conv3D) with LeakyReLU activation and Instance Normalization (InstanceNorm3D) to extract spatial and contextual features. To enhance representational power, we incorporate Gated Spatial Convolution (GSC) and Mamba

layers at deeper levels. The downsampling process is implemented via strided convolutions, progressively reducing spatial resolution while increasing feature channels, resulting in hierarchical feature representations.

1). Gated Spatial Convolution: The GSC module (shown in Fig. 5(b)) enhances spatial feature learning by combining multiple receptive fields. It consists of parallel 3×3 convolutions, followed by normalization and activation. The outputs are gated through an additional 1×1 convolution that learns spatial importance weights. This mechanism effectively suppresses irrelevant features while enhancing critical spatial patterns.

2). Mamba Layer: Each Mamba Layer begins (shown in Fig. 5(c)) with Layer Normalization, which normalizes the input sequence by computing the root mean square value of the input activations. This step is crucial for preventing gradient explosion in deep networks. Following normalization, the Mamba module processes the input sequence, and the resulting output is combined with the input residuals, as expressed in the equation:

$$\mathbf{x}_{i+1} = \text{Unflatten}(\text{Mamba}(\text{Flatten}(\text{LayerNorm}(\mathbf{x}_i)))) + \mathbf{x}_i. \tag{4}$$

Initially, the input features: $\mathbf{x}_i$ undergo a linear transformation and are then split into two components: $\mathbf{y}$ and $\mathbf{z}$. These components are obtained via the operation $\mathbf{y}, \mathbf{z} = \text{split}(\text{linear}(\mathbf{x}_i))$. The $\mathbf{y}$ segment is processed through a 1D convolution, followed by activation and further processing via the Selective Scan Model (SSM):

$$\mathbf{y}_i = \text{SSM}(\text{SiLu}(\text{1D Conv}(\mathbf{y}))). \tag{5}$$

Concurrently, the activated $\mathbf{z}$ segment acts as a gating vector, which is element-wise multiplied with the $\mathbf{y}_i$. Once processed by the Mamba module, $\mathbf{y}_i$ is passed through another linear layer to yield the final result of this module: $\mathbf{x}_{i+1}$.

Organ Segmentation Decoder: The organ segmentation decoder follows an encoder-decoder structure, where high-level semantic features are progressively upsampled and concatenated with corresponding low-level features via skip connections. Transposed convolutions (TransConv) are used for upsampling, while multi-layer perceptrons (MLP) further refine the feature representations. The final segmentation mask is generated through a 6-channel output layer, corresponding to the target organs.

Disease Anomaly Map Generator: Parallel to the segmentation decoder, the disease anomaly map generator utilizes the same encoded features to predict abnormal regions. Instead of producing organ-wise labels, this branch outputs 16-channel anomaly maps, which highlight potential pathological regions. The decoder follows a structure similar to the segmentation decoder, but with additional feature fusion mechanisms to integrate deeper semantic cues. A final element-wise addition step merges multi-scale outputs to generate the anomaly heatmap.

Appendix B. Generation of Pseudo Labels

Table 3: The specific associations between diseases and organs.

	CME	PCE	CAC	HH	LAP	EMSE	ATE	LN	GGO	PFS	PLE	MAP	PBT	CON	BE	ILST
Lung						✓	✓	✓	✓	✓		✓		✓		✓
Trachea and Bronchie													✓		✓	
Pleura											✓					
Mediastinum					✓											
Heart	✓	✓	✓													
Esophagus				✓												

We utilized 16 disease labels extracted from CT-RATE by RadBERT, including Cardiomegaly (CME), Pericardial effusion (PCE), Hiatal Hernia (HH), Lymphadenopathy (LAP), Emphysema (EMSE), Atelectasis (ATE), Lung nodule (LN), Lung opacity (Ground-glass opacity, GGO), Pulmonary fibrotic sequela (PFS), Pleural effusion (PLE), Mosaic attenuation pattern (MAP), Peribronchial thickening (PBT), Consolidation (CON), Bronchiectasis (BE), Interlobular septal thickening (ILST), and Coronary artery wall calcification (CAC). Based on medical domain knowledge, these diseases are associated with different organs to provide coarse localization information. Specifically, as shown in Table 3, lung-related abnormalities are the most frequent, including ATE, LN, GGO, PFS, PBT, CON, BE, and ILST, followed by those related to the heart and trachea. This is because chest CT is primarily used to assess lesions in these regions.

Appendix C. The Disease Distribution of Different Datasets

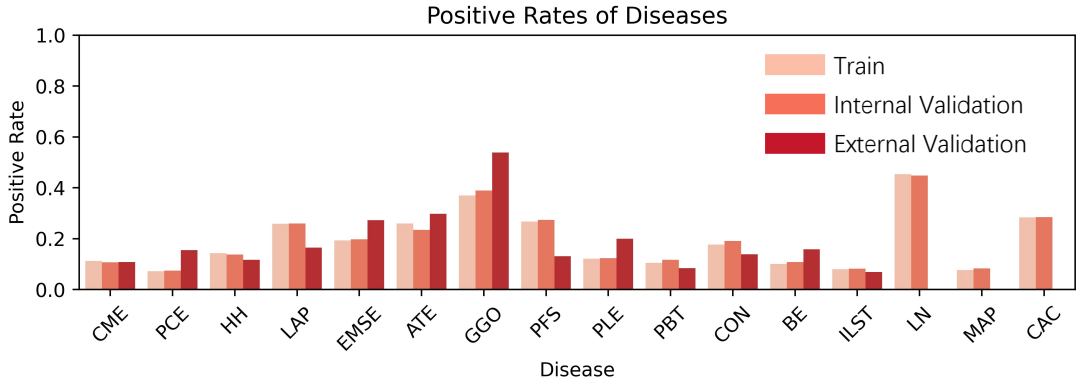

Figure 6: The disease distribution of different datasets.

The positive rates of different diseases in the training, internal validation, and external validation datasets (RAD-ChestCT) are shown in Fig. 6 (LN, MAP, and CAC are unavailable in the external validation set). The data distribution in the training and internal validation sets is consistent, while the positive rates in the external dataset differ significantly from those in the other two sets.

Appendix D. ROC Curve of Chest-OMDL for Multidisease Detection

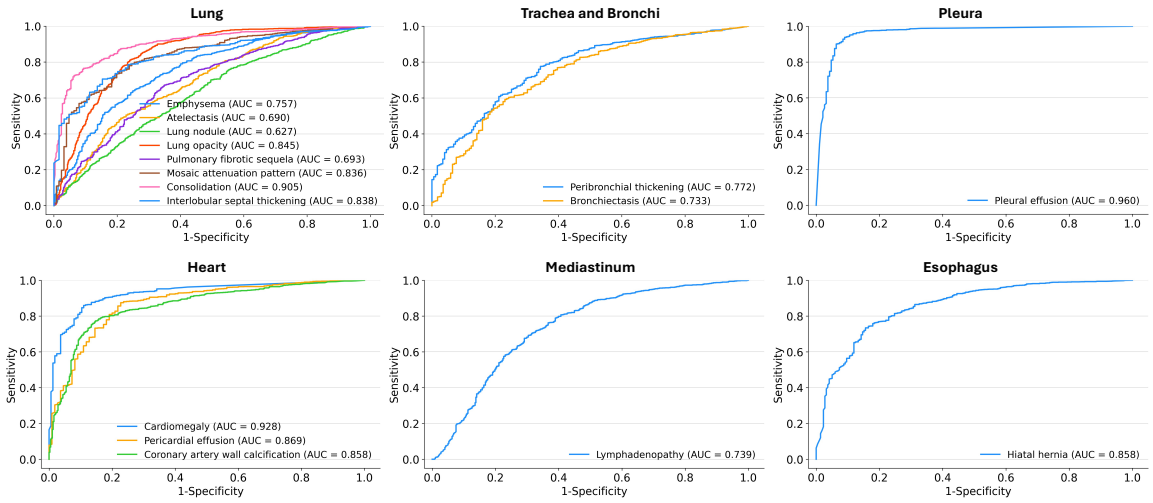

Figure 7: Receiver operating characteristic (ROC) curves for organ-specific multidisease detection (Internal validation dataset).

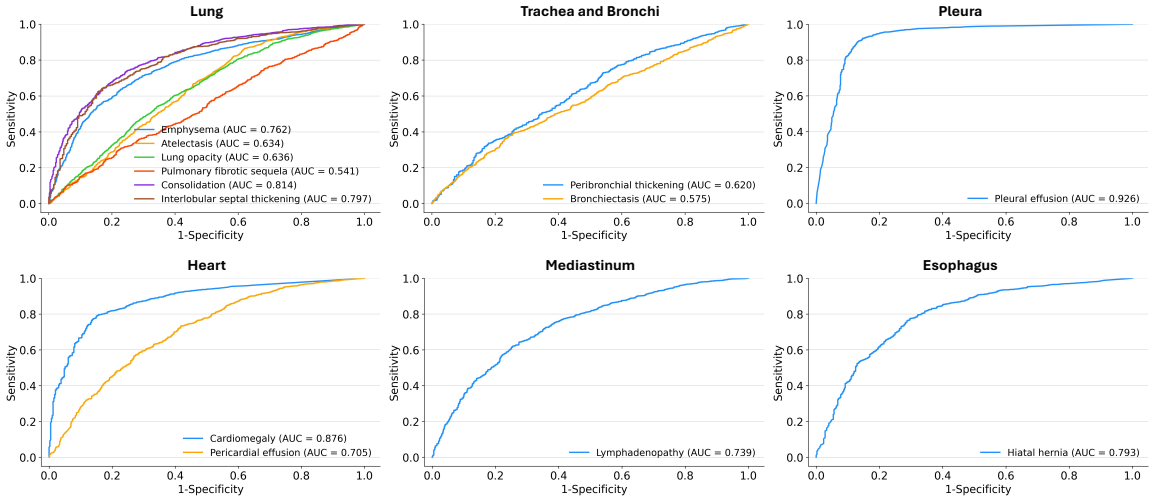

Figure 8: Receiver operating characteristic (ROC) curves for organ-specific multidisease detection (External validation dataset).

Appendix E. Detailed Detection Performance on Each Disease

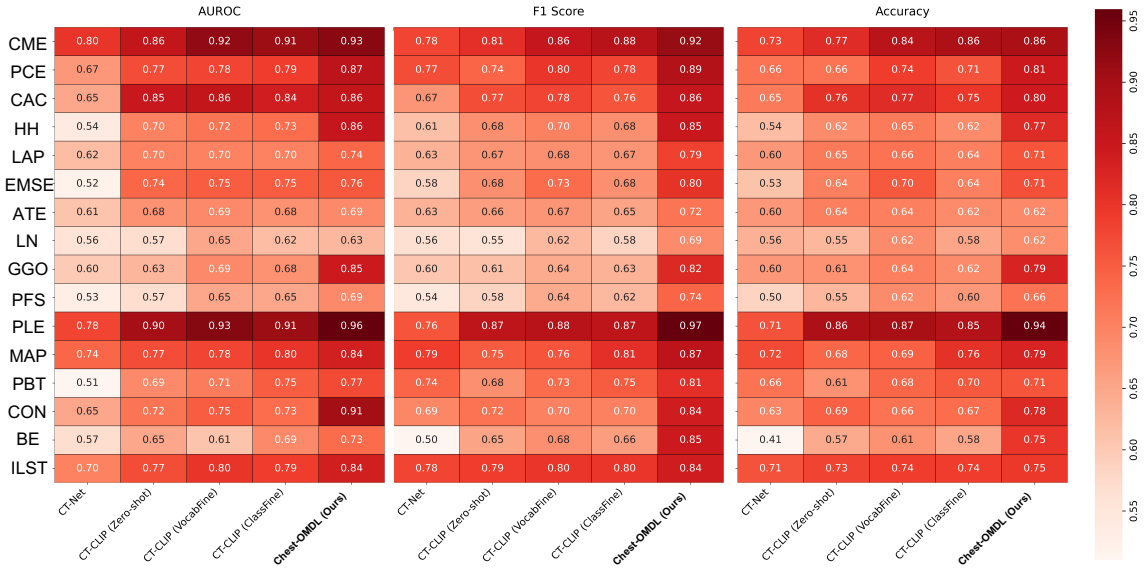

Figure 9: Comparison of anomaly-based performance metrics in the internal validation set.

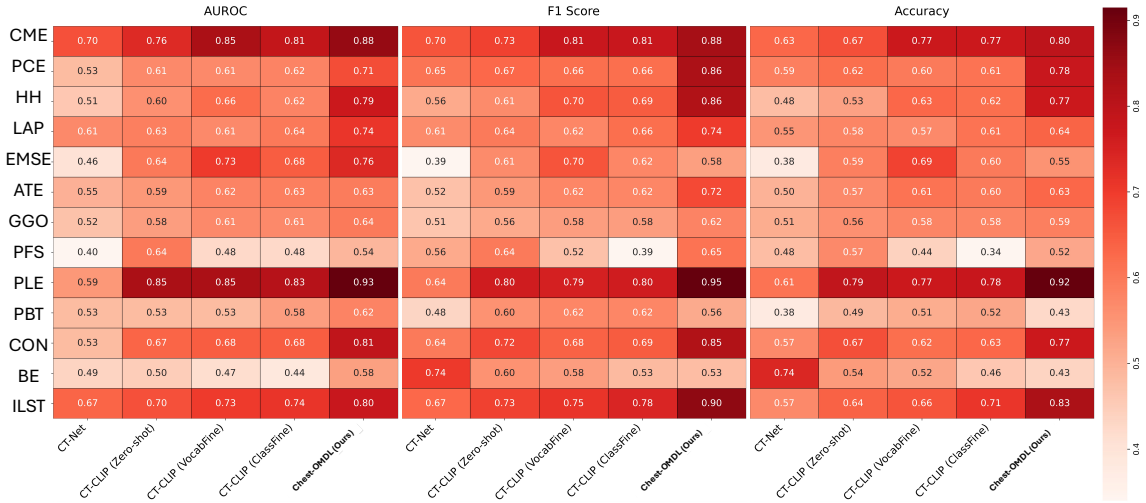

Figure 10: Comparison of anomaly-based performance metrics in the external validation set.

Figures 9 and 10 illustrate the performance of different methods on both internal and external validation datasets across various diseases, evaluated using AUROC, F1 score, and

accuracy. The results show that Chest-OMDL outperforms CT-Net, CT-CLIP, and the two fine-tuned variants of CT-CLIP for most diseases. This highlights the model's exceptional adaptability and superior effectiveness under distribution shifts, setting a new benchmark compared to a fully supervised baseline and contrastive learning-based CLIP models.

Appendix F. Additional Localization Results on the Covid-19 CT Dataset

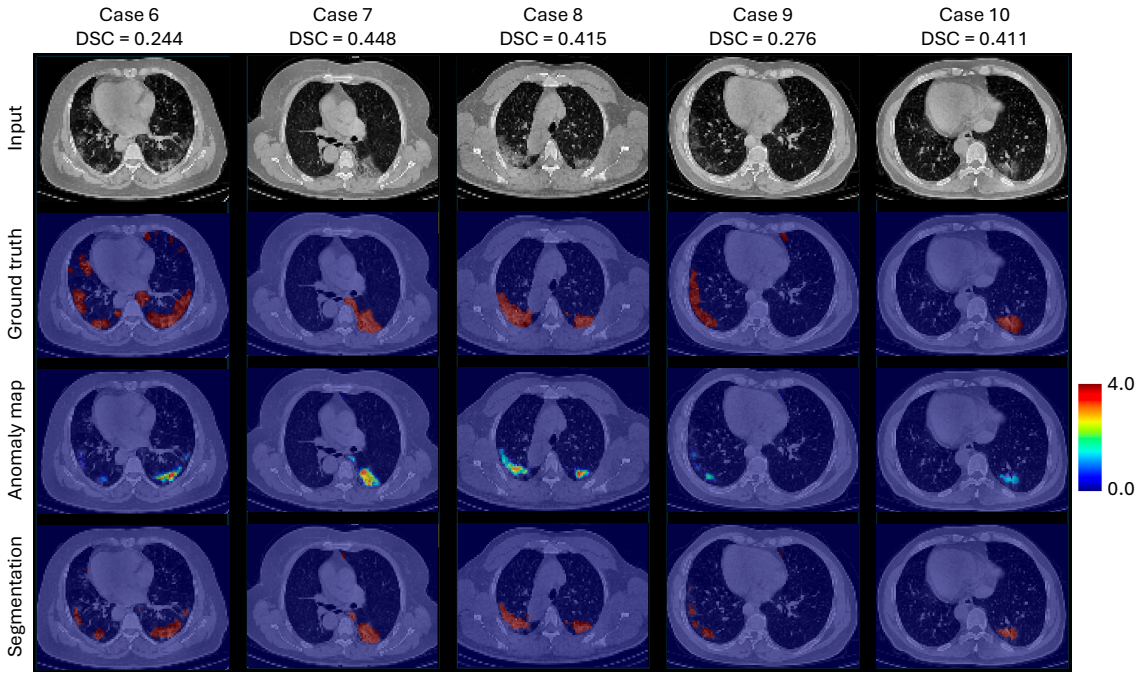

Figure 11: Additional Abnormality localization results on the Covid-19 CT dataset.

Figure 11 shows slice-level examples from five cases with low segmentation DSC metrics. The suboptimal performance is mainly due to the small size of the abnormal regions localized by Chest-OMDL, which fail to fully cover the lesions, reflecting the model's reduced sensitivity to subtle abnormalities without full supervision. However, across all 10 cases (including Figures 4 and 11), Chest-OMDL consistently detected lung abnormalities and localized the regions of interest, despite imprecise segmentation. Fine-tuning on a small labeled dataset could greatly enhance its performance, which will be a focus of future work.