LungCT-Multi: A Comprehensive Lung CT Dataset for Segmentation, Classification, Pre-Training and Dehazing

Abstract

Computed Tomography (CT) is a crucial imaging modality for diagnosing and monitoring pulmonary diseases. Despite the promise of deep learning for automated analysis, progress hindered by the lack of datasets supporting multi-task-learning. Existing datasets are typically limited to a single task, such as segmentation or classification, and rarely provide the large-scale unlabeled data needed for self-supervised pre-training or address common clinical challenges like poor image quality. To bridge this gap, we introduce LungCT-Multi, a comprehensive dataset of **[Number of CT Images]** designed for diverse research applications. It contains 2,450-pixel level lesion masks for segmentation, 27 labeled and 1244 unlabeled slices for classification and pre-training, and a challenging subset of 1000 low-quality slices for image dehazing. We establish strong baselines across all tasks, achieving a 0.714 Dice Score for segmentation. 86.0% few-shot classification accuracy, up to 42.2% improvement via pre-training, and substantial enhancement of image quality through dehazing. LungCT-Multi will be publicly released to advance research in robust, data-efficient pulmonary image analysis.

1 Introduction

Computed Tomography (CT) has emerged as a cornerstone imaging modality in pulmonary medicine, providing non-invasive visualization of lung structures that is essential for diagnosing conditions ranging from infections and inflammation to cancer [1]. The COVID-19 pandemic further underscored CT's clinical value, establishing it as a vital tool for assessing lung involvement and monitoring disease progression [2]. Despite its widespread use, manual interpretation of CT scans is time consuming, subject to inter-observe variability, and increasingly challenged by the growing volume of imaging data. These limitations have spurred the integration of deep learning into medical imaging, where convolutional neural networks (CNNs) and transformer architectures have demonstrated remarkable capabilities in tasks such as lesion segmentation [3], disease classification [4], and quantitative assessment [5]. However, the development of robust, clinically applicable deep learning systems is hindered by a fundamental bottleneck: the scarcity of comprehensively annotated datasets. Current pulmonary CT resources are fragmented. Most existing datasets are designed for isolated tasks they provide either pixel-level annotations for segmentation [6,7] or clinical labels for classification [8,9], but rarely both. This division prevents researchers from exploring multi-task learning paradigms, where a single model could simultaneously localize pathological findings and characterize disease severity [10]. Additionally, the limited availability of large-scale unlabeled data restricts research into self-supervised pre-training methods [11,12], which have shown exceptional performance in other computer vision domains by learning robust feature representations from unlabeled data using contrastive learning or Siamese networks [13,14].

These limitations have practical consequences for clinical translation. Real-world diagnostic workflows typically require both localization and characterization of abnormalities, yet current datasets force researchers to develop separate models for each task or combine heterogeneous datasets with inconsistent annotations. Furthermore, clinical CT imaging often suffers from noise, artifacts, and suboptimal contrast [15], but existing datasets rarely include such challenging cases or provide benchmarks for image enhancement techniques using deep learning.

To address these gaps, we present LungCT-Multi, a comprehensive pulmonary CT dataset specifically designed to support multiple deep learning research directions within a unified framework. The primary contributions of this work are as follows:

- A publicly available lung CT dataset supporting four distinct tasks: Segmentation, Classification, Self-supervised Pre-training, and image Dehazing.
- Comprehensive benchmarking of state-of-the-art methods on all tasks, providing strong baseline performance for future research.

By providing this multi-faceted resource, we enable the research community to explore novel synergies between tasks and establish comprehensive baseline for future deep learning research. LungCT-Multi represents a significant step toward developing robust, data-efficient, and clinically relevant AI systems for pulmonary medicine.

2 Related Work

2.1 Medical Image Dataset

The medical imaging community has benefited from several public datasets. LIDC-IDRI [17] is widely used for lung nodule analysis, while COVID-19 prompted the creation of collections such as COVID-CT [7] and MosMedData [9]. However, these resources are typically single-task, limiting their use for exploring multi-task learning and self-supervised pre-training. COVID-CT focuses on classification [7], whereas MosMedData provides limited segmentation masks [11], highlighting the need for datasets that integrate multiple research tasks.

2.2 Segmentation and classification Methods

U-Net [3] establish a foundational architecture for medical image segmentation, with subsequent advances incorporating transformers [16,18]. For classification, semi-supervised approaches such as Mean Teacher [12] and Adaptive Consistency Regularization [19] have been proposed to address label scarcity. Prior work on Trans-Inf-Net [18] and semi-supervised classification [10] demonstrates the effectiveness of these methods, but also their dependence on well-structured and diverse datasets.

2.3 Self-Supervised Pre-Training

Self-Supervised methods, including contrastive learning [13] and Siamese networks [11,20], have proven effective for learning representations from unlabeled data. In medical imaging, this is particularly valuable given the high cost of annotation. Previous work on SegSiamPT [11] demonstrates the potential of using segmentation masks as a self-supervision signal- a strategy that requires datasets with diverse annotations such as LungCT-Multi.

3. The LungCT-Multi Dataset

This section provides comprehensive description of the LungCT-Multi dataset, detailing its composition, annotation process, and organization into dedicated subsets for different research tasks.

3.1 Data Collection and Sources

The LungCT-Multi dataset aggregates volumetric chest CT scans from multiple clinical sources, with the primary collection originating from Huoshenshan Hospital [7,10]. The dataset comprises scans collected from patients with various pulmonary conditions, ensuring diversity in disease presentation and severity. All data underwent rigorous de-identification following HIPPA guidelines, with all personal identifiers removed and the study approved by institutional review boards of participating medical institutions.

3.2 Dataset Composition and Statistics

LungCT-Multi is organized into four distinct subsets, each tailored to support specific research directions while maintain clinical relevance and annotation quality.

Table 1: LungCT-Multi Dataset Composition and Splits

Subset	Total Slices	Training	Validation	Test
Segmentation				
Classification				
Pre-Training				
Dehazing				
Total				

Table 2: Clinical Severity Distribution in Classification Subset

Severity level	Slices	Clinical Description
Ordinary		Limited pulmonary involvement, mild symptoms
Severe		Significant lesion progression, respiratory symptoms
Critical		Extensive bilateral involvement, respiratory failure
Total		Comprehensive severity spectrum

3.3 Annotation Methodology and Quality Control

Segmentation subset:

Three board-certified radiologists with 5+ years of experience manually delineated pulmonary lesions including ground-glass opacities, consolidation, and mixed patterns characteristic of

COVID-19 pneumonia [6]. The annotation process followed established guidelines with each mask undergoing verification by at least two radiologists. Quality assurance showed high inter-rater reliability with a Dice coefficient of 0.85±0.06.

Classification Subset:

Clinical severity labels were manually assigned by expert radiologists according to Chinese COVID-19 Diagnosis and Treatment Guidelines, with categories reflecting disease progression and treatment urgency. The labeled portion (number of labeled) represents carefully selected cases covering the full severity spectrum, while the unlabeled portion (number of unlabeled) enables semi-supervised and self-supervised learning research.

Pre-training Subset:

This subset consists of completely unlabeled data (unlabeled image numbers) specifically curated for self-supervised learning research. These images contain no manual annotations and are intended for methods that learn representations from unlabeled data, such as contrastive learning and auto-encoding approaches.

Dehazing Subset:

Approximately 1000 slices with inherent quality issues including poor lightning, noise artifacts, and suboptimal contrast were systematically identified and curated. This subset addresses the practical challenge of handling real-world clinical image quality variations.

3.4 Data Preprocessing Pipeline

All CT images underwent standardized preprocessing following established protocols from our previous work. For classification tasks, input CT images were uniformly adjusted to 224*224-pixel resolution [10]. Transformer-based architectures utilized patch-based processing 8*8 patch size configuration [7]. Intensity normalization converted images to tensor format with normalization to the [0-1] range using channel-wise parametrs with mean values of (0.4811,0.4575, 0.4078) and standard deviation of (0.2605, 02533, 0.2683) [10]. Comprehensive data augmentation strategies were employed including random rotation, flipping with probability p=0.5, random cropping with reflection padding, and center cropping for testing and validation phases [7,10]. These preprocessing steps ensure consistency across experiments while maintain the clinical relevance of the CT data for all supported tasks including segmentation, classification, pre-training, and image enhancement.

3.5 Dataset Structure and Accessibility

```
LungCT-Multi/
├── segmentation/
│   ├── images/              # Original CT slices
│   └── masks/               # **Manual** pixel-level annotations
├── classification/
│   ├── labeled/             # 27 expert-**manually** annotated slices
│   └── unlabeled/           # 1,244 slices for SSL
├── pre-training/            # 1,244 unlabeled slices for SSL
├── dehazing/
│   ├── hazy/                # Poor-quality images
│   └── enhanced/            # Reference enhanced images
└── metadata/
    ├── clinical_info.csv    # Patient demographics
    └── splits/              # Train/val/test splits
```

4. Experiments and Results

All experiments were conducted on NVIDIA RTX2080sTi/RTX2080Ti GPUs using PyTorch 1.7.1 and Python 3.8.11 [7, 11]. For segmentation tasks, models were trained for 20,000 iterations with a batch size of 6, using the Stochastic Gradient Descent (SGD) optimizer due to its computational efficiency with mini-batches, adaptability to noisy gradients, and regularization benefits [7]. Classification experiments employed 50 training epochs with batch size 16, using SGD optimizer with initial learning rate of 0.001, weight decay of 0.001, and momentum of 0.9 [10]. The Adam optimizer was utilized for transformer components with learning rate 1e-4 [11]. Standard data augmentations including random rotation and flipping were applied during training [7, 11].

4.2 Segmentation Performance

The segmentation performance was evaluated on COVID ….

4.3 Classification Performance

4.4 Pre-Training Performance

4.5 Image Dehazing for Quality Enhancement

5. Discussion

6. Conclusion

Refernces

[1] J. A. Verschakelen and W. De Wever, Computed Tomography of the Lung: A Pattern Approach. Springer, 2007.

[2] T. Ai et al., "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," Radiology, vol. 296, no. 2, pp. E32-E40, 2020. https://doi.org/10.1148/radiol.2020200642

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[4] L. Wang et al., "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," Scientific Reports, vol. 10, no. 1, p. 19549, 2020. https://doi.org/10.1038/s41598-020-76550-z

[5] F. Shan et al., "Lung infection quantification of COVID-19 in CT images with deep learning," arXiv preprint arXiv:2003.04655, 2020. https://arxiv.org/abs/2003.04655

[6] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, and L. Shao, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images," IEEE Transactions on Medical Imaging (TMI), vol. 39, no. 8, pp. 2626-2637, 2020. https://doi.org/10.1109/TMI.2020.2996645

[7] J. Zhang, F. Li, X. Zhang, Y. Cheng, and X. Hei, "Multi-Task Mean Teacher Medical Image Segmentation Based on Swin Transformer," Applied Sciences, vol. 14, no. 7, p. 2968, 2024. https://doi.org/10.3390/app14072968

[8] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-Dataset: A CT scan dataset about COVID-19," arXiv preprint arXiv:2003.13865, 2020.

https://arxiv.org/abs/2003.13865

[9] S. P. Morozov et al., "MosMedData: Chest CT scans with COVID-19 related findings," arXiv preprint arXiv:2005.06465, 2020. https://arxiv.org/abs/2005.06465

[10] J. Zhang, J. Wang, Y. Zhang, K. Xiang, M. Kang and Z. Pan, "A Transfer Learning Image Classification Method Using Self-Supervised Information," in 2022 8th International Conference on Virtual Reality (ICVR), 2022. https://doi.org/10.1109/ICVR55215.2022.9848136

[11] L. Zhu, M. Zhao, R. Peng, X. Hei, M. Sahrim, J. Zhang, "Pre-training with Siamese Networks using self-supervised Information for Unlabeled Images," International Conference on Parallel and Distributed Systems (ICPADS), 2025. [To be published]

[12] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Advances in Neural Information Processing Systems (NeurIPS), 2017. https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International Conference on Machine Learning (ICML), 2020, pp. 1597-1607. https://doi.org/10.5555/3524938.3525087

[14] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," Advances in Neural Information Processing Systems, vol. 33, pp. 21271-21284, 2020. https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html

[15] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 589-596. https://doi.org/10.1007/978-3-319-24553-9_72

[16] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[17] A. S. G. Armato III et al., "The Lung Image Database Consortium (LIDC) and Image

Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," Medical Physics, vol. 38, no. 2, pp. 915-931, 2011. https://doi.org/10.1118/1.3528204

[18] J. Zhang, K. Xiang, J. Wang, J. Liu, M. Kang and Z. Pan, "Trans-Inf-Net: COVID-19 Lung Infection Segmentation Based on Transformer," in 2022 8th International Conference on Virtual Reality (ICVR), 2022. https://doi.org/10.1109/ICVR55215.2022.9848136

[19] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou, "Adaptive Consistency Regularization for Semi-Supervised Transfer Learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6923-6932. https://doi.org/10.1109/CVPR46437.2021.00959

[20] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15750-15758. https://doi.org/10.1109/CVPR46437.2021.01549