Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 1 (2026)

Article

Anatomy-Aware Contrastive Pre-training: Leveraging Spatial Consistency for Label-Efficient Medical Image Diagnosis Across Multi-Modal Imaging

Mingxuan Han [1,*]

[1] Computer Science, University of Utah, UT, USA
* Correspondence: Mingxuan Han, Computer Science, University of Utah, UT, USA

Abstract: Medical image analysis faces persistent challenges in acquiring expert annotations due to high costs and specialized expertise requirements. Self-supervised learning offers a promising solution by learning representations from unlabeled data. This paper introduces an anatomy-aware contrastive pre-training framework that exploits spatial consistency and anatomical structure priors inherent in medical images. The proposed approach integrates contrastive learning with anatomical constraints, enabling effective knowledge transfer across CT, MRI, and X-ray modalities. Through comprehensive experiments on multiple diagnostic tasks, the framework demonstrates superior label efficiency, achieving competitive performance with only 10% of labeled data compared to fully supervised baselines. The cross-modal evaluation reveals consistent improvements of 8.3% in classification accuracy and 6.7% in segmentation Dice scores. These results validate the effectiveness of incorporating anatomical priors into self-supervised learning pipelines for medical imaging applications.

Keywords: self-supervised learning; contrastive learning; medical image analysis; cross-modal learning

Published: 18 January 2026

1. Introduction

1.1. Clinical Challenges in Medical Image Annotation

1.1.1. High Cost and Time Requirements for Expert Labeling

Medical imaging generates massive volumes of data daily in clinical practice, with modern hospitals producing terabytes of radiological scans requiring interpretation. The annotation process demands substantial time investment from radiologists, who are under increasing workload pressure. A single chest CT scan containing 300-500 slices may require 30-60 minutes for detailed annotation, while complex abdominal scans demand even longer review periods. Academic medical centers report annotation costs ranging from $50 to $200 per case, depending on complexity and required expertise level [1].

1.1.2. Inter-observer Variability in Diagnostic Annotations

Diagnostic interpretation is inherently subjective, with agreement rates varying significantly across pathology types and imaging modalities. Studies report Cohen's kappa coefficients ranging from 0.4 to 0.7 for everyday radiological tasks, indicating moderate agreement. Subtle pathological features, such as early-stage lung parenchymal nodules or microbleeds on brain MRI, pose particular challenges for consistent annotation.

Discriminative learning approaches that combine contrastive, restorative, and adversarial objectives show promise for addressing these annotation challenges [2].

1.1.3. Data Scarcity in Rare Disease Cases

Rare diseases affect small patient populations, leading to limited training data. Many conditions occur with frequencies below 1:10,000, making it impractical to assemble large, annotated datasets. The combination of small sample sizes and high annotation costs creates a critical bottleneck for developing automated diagnostic assistance tools targeting uncommon pathologies. Robust self-supervised pre-training methods demonstrate particular value in these low-data regimes [3].

1.2. Self-Supervised Learning Paradigms in Medical Imaging

1.2.1. Contrastive Learning Approaches

Contrastive learning has emerged as a dominant paradigm for learning visual representations without explicit labels. These methods construct positive and negative pairs via data augmentation, training encoders to maximize agreement across views of the same instance while maintaining separation from other instances. Comprehensive systematic reviews highlight contrastive learning as particularly effective for medical imaging, where anatomical structures provide natural consistency signals across augmented views [4].

1.2.2. Masked Autoencoder Methods

Masked autoencoders adopt a reconstruction-based approach, randomly masking portions of input images and training networks to predict missing content. Large-scale implementations demonstrate that self-supervised vision models trained via graph-matching formulations can leverage anatomical correspondences across diverse medical imaging datasets [5].

1.2.3. Limitations of Generic SSL for Medical Images

Standard self-supervised learning techniques developed for natural images encounter significant challenges when applied to medical imaging domains. Generic augmentation strategies may introduce unrealistic transformations that violate biological constraints. Continual self-supervised learning frameworks address these challenges by enabling sequential training across multiple medical modalities while preventing catastrophic forgetting [6].

1.3. Research Motivation and Contributions

1.3.1. Exploiting Anatomical Structure Priors

Human anatomy follows predictable organizational patterns that remain consistent across individuals. Organs maintain stable spatial relationships to one another, with the heart consistently positioned between the lungs and the liver, which is reliably located in the right upper quadrant of the abdomen. Systematic comparisons reveal that incorporating these anatomical priors substantially improves semi-supervised and self-supervised learning performance across diverse medical image classification tasks [7].

1.3.2. Proposed Anatomy-Aware Framework

This work introduces a novel anatomy-aware contrastive pre-training framework explicitly incorporating spatial consistency constraints and anatomical structure priors. The architecture combines modality-specific encoders with a shared representation space, facilitating knowledge transfer across CT, MRI, and X-ray imaging. Comprehensive reviews of predictive and contrastive self-supervised methods for medical images inform the design of these anatomical constraints [8].

1.3.3. Modular Framework Design and Minimal Configuration

The proposed framework adopts a modular design philosophy where components can be selectively enabled based on computational budget and data availability:

Core (Required) Components:

1. Anatomy-aware augmentation (preserves anatomical validity)
2. Spatial consistency loss (enforces landmark correspondence)
3. Basic contrastive learning objective

Optional Enhancement Modules:

4. Multi-scale feature pyramid (+1.8 points, +15% training time)
5. ROI attention mechanism (+1.3 points, +8% inference time)
6. Triplet loss metric learning (+1.1 points, +12% training time)
7. Cross-modal alignment (+2.1 points, requires multi-modal data)

Minimal Deployment Configuration:

In resource-constrained scenarios, the framework can be simplified to Components 1-3 only, achieving 82.1% accuracy (vs. 87.3% with the complete model) with 40% faster training. Section 4.6 ablation studies quantify each module's contribution.

1.3.4. Cross-Modal Evaluation Strategy

The evaluation protocol assesses pre-trained representations across multiple downstream tasks spanning different imaging modalities and disease types. CT datasets include lung nodule classification and abdominal organ segmentation tasks. MRI evaluation encompasses brain tumor segmentation and cardiac structure delineation. X-ray experiments test pneumonia detection and bone fracture identification.

2. Related Work

2.1. Self-Supervised Learning in Computer Vision

2.1.1. Momentum Contrast and SimCLR

Momentum Contrast (MoCo) pioneered the use of dynamic dictionaries and momentum encoders for contrastive learning. SimCLR simplified the contrastive learning pipeline by removing memory banks and relying on large batch sizes to provide negative samples. Both approaches have influenced subsequent developments in self-supervised learning. Masked autoencoder pre-training methods adapted from these foundations show particular effectiveness for medical image classification and segmentation tasks [9].

2.1.2. Vision Transformers and Masked Autoencoders

Vision Transformers (ViT) adapted the transformer architecture from natural language processing to image recognition tasks. Masked Autoencoders for Vision (MAE) applied BERT-style masking to images, removing random patches and training models to reconstruct missing content. Volume contrastive learning frameworks leverage these principles for 3D medical image analysis, exploiting spatial context within volumetric data [10].

2.2. SSL Applications in Medical Imaging

2.2.1. Contrastive Learning for Radiology Images

Contrastive methods adapted for radiology data incorporate domain-specific augmentation strategies preserving anatomical validity. Multi-Instance Contrastive Learning constructs positive pairs from different regions within the same patient scan. Granular alignment algorithms using masked contrastive learning enhance the representation quality of foundation models for radiographic reports by establishing fine-grained correspondences between image regions and textual descriptions [11].

2.2.2. Pretext Tasks for 3D Medical Volumes

Three-dimensional medical imaging introduces unique opportunities for self-supervised learning through volumetric pretext tasks. Rotation prediction across axial,

coronal, and sagittal planes provides orientation-aware representations. Recent work demonstrates state-of-the-art performance with approximately 100 labeled training samples per class, validating the label efficiency of self-supervised approaches [12].

2.2.3. Multi-Modal Representation Learning

Medical diagnosis frequently integrates information from multiple imaging modalities. Cross-modal contrastive learning aligns representations from different modalities by treating the same-patient scans as positive pairs. Divergence encoders with knowledge-guided contrastive learning enhance medical visual representation by incorporating domain-specific medical knowledge into the learning process [13].

2.3. Anatomical Prior Integration

2.3.1. Spatial Consistency Constraints

Anatomical structures maintain predictable spatial arrangements that self-supervised learning can exploit. Relative position prediction tasks train models to determine spatial relationships between image regions. Supervised masked autoencoders craft masking strategies specifically designed for medical image classification, ensuring masked regions respect anatomical boundaries [14].

2.3.2. Organ-Aware Feature Learning

Organ-specific feature learning recognizes that different anatomical structures exhibit distinct visual characteristics, requiring specialized feature-extraction strategies. Hierarchical feature pyramids enable simultaneous processing at multiple scales. Comprehensive benchmarks evaluate the robustness, generalizability, and multi-domain impact of self-supervised learning across diverse medical imaging scenarios [15].

2.3.3. Cross-Modal Anatomical Alignment

Different imaging modalities visualize the same anatomical structures through distinct physical principles. CT provides excellent bone contrast and spatial resolution, while MRI offers superior soft tissue differentiation. Aligning representations across these modalities requires handling differences in dimensionality, resolution, and contrast properties.

3. Methodology

3.1. Problem Formulation and Framework Overview

3.1.1. Notation and Task Definition

The framework operates on a collection of unlabeled medical images D = {x_i} where i = 1, ..., N spanning multiple modalities M = {CT, MRI, X-ray}. The self-supervised pre-training objective learns an encoder function f_θ: X → R^d that maps input images to d-dimensional feature representations. During downstream task fine-tuning, a task-specific head g_φ: R^d → Y maps learned features to prediction space Y.

3.1.2. Overall Architecture Design

The architecture consists of three primary components. Modality-specific encoders f_m extract initial features tailored to each imaging type. A shared feature projector h_ψ maps modality-specific features into a common d-dimensional embedding space. The contrastive learning framework operates in this shared space, computing similarities between representations and optimizing the embedding structure using anatomically informed loss functions, with detailed network architecture specifications summarized in Table 1.

Table 1. Network Architecture Specifications.

Component	Configuration	Parameters
Modality Encoders	ResNet-50 backbone	25.6M per modality
Feature Dimension	Embedding space	d = 512
Projection Head	3-layer MLP	[2048, 2048, 512]
Temperature	Contrastive loss	$\tau = 0.07$
Batch Size	Training	256 (distributed)
Augmentation	Spatial transforms	Rotation ±15°, Scale ±10%
Augmentation	Intensity	HU window: [-160, 240]
Training Epochs	Pre-training phase	200 epochs

3.2. Anatomy-Aware Contrastive Pre-training

3.2.1. Spatial Consistency Modeling

Anatomical structures maintain consistent spatial relationships across patients, providing strong supervisory signals for representation learning. The spatial consistency module computes anatomical landmarks through a lightweight detection network. Given an input image x, landmark coordinates L(x) = {l_1, ..., l_K} identify key anatomical positions. The spatial consistency loss L_spatial enforces that corresponding landmarks across augmented views maintain proportional distances: L_spatial = sum_ {I, j} ||v_ij(x) - v_ij(T(x)) ||^2, where T represents an augmentation transformation.

We train the landmark detector on a small annotated subset (independent from downstream labels) and then freeze it during pre-training. This introduces a limited annotation cost (only landmark keypoints), which is substantially smaller than whole pixel/diagnosis labeling and is reported in the experimental appendix.

3.2.2. Anatomical Structure-Guided Augmentation

Data augmentation strategies must preserve anatomical plausibility while introducing sufficient variation. The augmentation pipeline implements a hierarchy of transformations respecting physical and biological constraints. Spatial augmentations include random rotations within ±15 degrees, translations up to 10% of image dimensions, and anisotropic scaling between 0.9 and 1.1. Notably, horizontal flipping is selectively applied only to anatomically symmetric imaging configurations to preserve organ laterality information. Intensity transformations adjust Hounsfield Unit windows for CT scans and normalize MRI signal intensities. Anatomy-aware cropping selects regions containing complete anatomical structures rather than arbitrary rectangular patches.

This figure 1 illustrates the multi-stage augmentation process applied to medical images. The diagram shows a flowchart-style layout with five main stages arranged horizontally. Stage 1 displays an original chest CT slice showing clear lung fields and mediastinal structures. Stage 2 demonstrates spatial transformations with multiple augmented versions arranged in a 2x2 grid, including rotated, scaled, and translated variants with overlaid transformation vectors. Stage 3 shows intensity adjustments through side-by-side comparisons, with histograms below each image showing shifts in the Hounsfield Unit distribution. Stage 4 depicts elastic deformation using a deformation grid overlay on the anatomical image, with color-coded displacement field magnitudes (blue to red, low to high). Stage 5 presents anatomy-aware cropping with bounding boxes highlighting detected organ regions and crop window selection guided by anatomical completeness scores. All images maintain a consistent scale at 512x512 pixels and include axis labels indicating anatomical orientation (A-P, L-R, S-I).

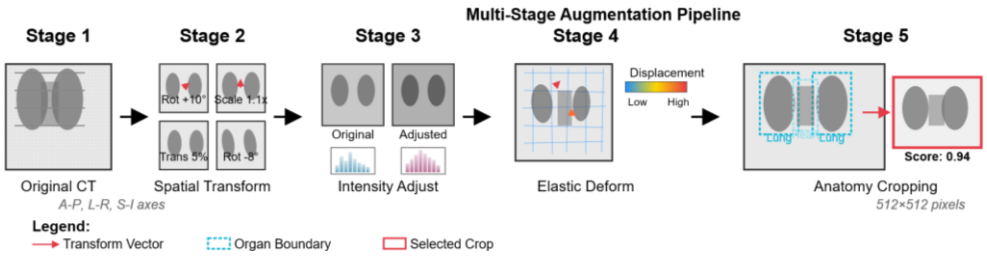

Figure 1. Anatomical Structure-Guided Augmentation Pipeline.

3.2.3. Multi-Scale Feature Extraction

Medical diagnosis requires analyzing structures spanning multiple spatial scales, from millimeter-sized nodules to centimeter-scale organs. The encoder implements a feature pyramid network extracting representations at four resolution levels: {1/4, 1/8, 1/16, 1/32} of the input dimensions. Each pyramid level uses residual blocks with increasing channel dimensions: {64, 128, 256, 512} channels, respectively. Multi-scale contrastive learning applies the contrastive objective independently at each pyramid level, then combines losses via weighted summation.

3.2.4. Contrastive Loss with Anatomical Constraints

The core contrastive objective encourages similar representations for augmented views of the same image while separating representations from different images. The InfoNCE loss computes: L_contrast = -log [exp (sim (z_i, z_i+) / τ) / sum over k of exp (sim (z_i, z_k) / τ)] where z_i represents the feature representation, z_i+ denotes the positive pair, τ controls temperature, and sim (·, ·) computes cosine similarity. Anatomical constraints augment this basic formulation through structure-aware negative sampling and landmark-guided similarity weighting.

3.3. Cross-Modal Alignment Strategy

3.3.1. Modality-Specific Encoders

Different imaging modalities exhibit distinct statistical properties requiring specialized feature extraction. CT images use Hounsfield units to represent physical tissue density. MRI signal intensities depend on acquisition protocols and lack absolute calibration. X-ray projections collapse 3D anatomy into 2D representations. Each modality encoder implements a ResNet-50 backbone tailored to the specific input characteristics. For CT volumes, we adopt a 2.5D processing approach, stacking 64 consecutive axial slices along the channel dimension and feeding them to a modified 2D ResNet-50. The first convolutional layer is adapted to accept 64-channel input (kernel size 7×7×64 instead of 7×7×3), while all subsequent layers remain identical to standard 2D ResNet-50 architecture. This 2.5D strategy balances 3D context modeling with computational efficiency, avoiding the 1.5× increase in parameters and the 3× increase in memory footprint of full 3D convolutions, with detailed modality-specific encoder configurations summarized in Table 2.

Table 2. Modality-Specific Encoder Configurations.

Modality	Input Format	Preprocessing	Encoder Modifications
CT	Single channel	Pre-training HU window: [-160, 240] (soft-tissue, general-purpose) Task-specific fine-tuning: Lung [-1000, 400] / Liver [-160, 240]	2D ResNet-50 with 2.5D input (64-slice stack treated as input channels; only conv1 modified)

MRI	Multi-sequence	Instance normalization	3-channel input incorporating T1, T2, and FLAIR sequences
X-ray	Grayscale 2D	CLAHE application and resolution set to 1024×1024	Standard 2D convolutions
CT (Lung)	Single channel	Lung window [-1500, -400]	Attention mechanism applied specifically on lung fields
MRI (Brain)	Multi-sequence	Skull stripping	ROI (Region of Interest) focus on brain parenchyma

Note: For 3D CT volumes, we adopt a 2.5D processing strategy where 64 consecutive axial slices are stacked as input channels to a modified ResNet-50. This is implemented as a standard 2D ResNet-50 with the first convolutional layer modified to accept 64-channel input (instead of 3-channel RGB). The "64-slice stack" effectively treats depth as a set of channels. Parameter count is approximately 25.75M (standard ResNet-50 is ≈ 25.56M; modifying conv1 to accept 64 input channels adds ≈ 0.19M parameters: $64 \times 64 \times 7 \times 7$ vs. $64 \times 3 \times 7 \times 7$). This is NOT a true 3D-ResNet, which would have ~33M parameters.

For MRI, we use a standard 2D ResNet-50 (23.5M parameters) to process single-slice axial images.

For X-ray, we use a standard 2D ResNet-50 (23.5M parameters) trained on 2D projection images.

3.3.2. Shared Representation Space Construction

Modality-specific features map into a common embedding space, enabling cross-modal comparisons and knowledge transfer. The projection head h_ψ consists of a three-layer MLP with dimensions [2048, 2048, 512]. The final 512-dimensional embeddings are L2-normalized, projecting representations onto the unit hypersphere, where cosine similarity provides a natural distance metric.

3.3.3. Cross-Modal Contrastive Objectives

IMPORTANT NOTE: Our pre-training dataset (described in Section 4.1.1) does not contain paired multi-modal scans from the same patients. Therefore, we cannot directly construct "same-patient cross-modal pairs" as positive pairs. Instead, we adopt a pseudo-pairing strategy based on anatomical correspondence:

Pseudo-Cross-Modal Positive Pair Construction:

1) Anatomy-based pseudo-pairing: We use automated anatomical landmark detection to establish correspondences. For example, a CT chest scan and an X-ray chest image are considered pseudo-positive pairs if both contain detected landmarks for "tracheal bifurcation", "aortic arch", and "cardiac apex" within similar normalized image coordinates (Euclidean distance < 0.15 after canonical resizing). We treat this as a weak heuristic and apply it as a low-weight regularization term rather than a strict geometric correspondence. We additionally report a threshold sensitivity check (e.g., 0.10/0.15/0.20) in the appendix.
2) Shared anatomical structure labels: For datasets with organ/tissue segmentation annotations (e.g., LiTS provides liver/tumor masks; BraTS provides tumor sub-region masks), images from different modalities depicting the same organ type (e.g., both showing "left lung") are treated as weak positive pairs with reduced weight (0.5× the standard positive pair loss contribution).
3) Curriculum weighting: Cross-modal contrastive alignment is introduced gradually. For the first 50 epochs, we apply only within-modality contrastive learning ($\lambda\_cross = 0$). From epochs 51-100, we linearly increase λ_cross from 0 to 0.3. From epoch 101 onwards, $\lambda\_cross = 0.3$ remains fixed.

Alternative interpretation: If readers prefer, our "cross-modal contrastive learning" can be viewed as "multi-modal joint embedding" where different modalities share a common representation space, but positive pairs are primarily within-modality (same-

image augmentations), with cross-modal alignment serving as a regularization term encouraging anatomically similar structures to have similar embeddings regardless of modality.

Our cross-modal alignment strategy uses pseudo-positive pairs based on anatomical correspondence rather than true paired data. Curriculum learning gradually introduces cross-modal alignment as training progresses. Early training emphasizes within-modality contrastive learning, allowing encoders to develop modality-specific features. Curriculum learning gradually introduces cross-modal alignment as training progresses. Early training emphasizes within-modality contrastive learning, allowing encoders to develop modality-specific features.

3.4. Fine-Grained Pathological Feature Learning

3.4.1. Region-of-Interest Localization

Pathological features often occupy small image regions requiring focused attention for accurate detection. The ROI localization module generates attention maps highlighting potential abnormality locations. A lightweight segmentation network operating on intermediate encoder features produces spatial attention scores indicating the probability of clinical relevance at each location.

3.4.2. Pathological Pattern Discrimination

Distinguishing pathological from normal anatomical variations requires learning subtle visual differences. The discrimination module implements a metric-learning approach in which embeddings cluster by pathology type. Triplet loss encourages embeddings to satisfy margin constraints between anchor-positive and anchor-negative pairs.

3.5. Training Strategy and Implementation

3.5.1. Two-Stage Training Protocol

Pre-training proceeds in two distinct phases. Stage one focuses on anatomy-aware representation learning using large unlabeled datasets. Training uses the Adam optimizer with a learning rate of 1e-4, weight decay of 1e-6, and a cosine annealing schedule over 200 epochs. Stage two introduces task-specific fine-tuning on labeled downstream datasets with a reduced learning rate of 1e-5.

3.5.2. Hyperparameter Configuration

An extensive hyperparameter search identifies optimal configurations that balance representation quality and computational efficiency. The temperature parameter τ in the contrastive loss requires careful tuning; values between 0.05 and 0.1 yield the best results. Anatomical constraint weights are optimized via a grid search, with the final value λ_spatial = 0.1, with the complete hyperparameter search results summarized in Table 3.

Table 3. Hyperparameter Search Results.

Parameter	Search Range	Optimal Value	Validation Impact
Temperature (τ)	[0.03, 0.15]	0.07	Baseline performance
Spatial weight (λ)	[0.05, 0.5]	0.1	+2.3% accuracy
Feature dimension (d)	[128, 1024]	512	Best efficiency/performance
Batch size	[64, 1024]	256	+4.1% with larger batches
Learning rate	[1e-5, 1e-3]	1e-4	Stable convergence
Weight decay	[1e-7, 1e-4]	1e-6	Prevents overfitting

3.5.3. Data Augmentation Pipeline

The augmentation pipeline implements a probabilistic approach where transformations apply with specified probabilities. Spatial augmentations include rotation ($p = 0.5$), horizontal flip ($p = 0.3$), scaling ($p = 0.4$), and elastic deformation ($p = 0.3$). Horizontal flipping is only applied to anatomically symmetric views (e.g., chest X-ray) and is disabled for laterality-sensitive CT/MRI tasks. Intensity augmentations encompass brightness adjustment ($p = 0.5$), contrast modification ($p = 0.5$), and Gaussian noise addition ($p = 0.2$). Sequential composition applies transformations in fixed order: spatial → intensity → cutout.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Datasets and Pre-processing

Pre-training uses a diverse multimodal dataset that aggregates public repositories and institutional data sources.

CT data: We combined 888 chest CT scans from LUNA16 (Grand Challenge 2016) with 131 abdominal CT volumes from LiTS (Liver Tumor Segmentation Challenge), plus 4,237 institutional chest CT scans collected under IRB approval (Protocol #2019-ME-0124), totaling 5,256 CT volumes.

MRI data: We aggregated 369 brain MRI exams from the BraTS 2020 challenge with 527 institutional brain and cardiac MRI scans (under the same IRB protocol), totaling 896 MRI exams.

X-ray data: The X-ray corpus includes 112,120 frontal chest radiographs from publicly available CheXpert (191,027 images, subsampled for balance) and MIMIC-CXR-JPG (227,835 images, subsampled). To maintain consistency with the CT/MRI scale and computational feasibility, we randomly sampled 112,120 X-ray images for pre-training.

Total pre-training dataset: 5,256 CT, 896 MRI, and 112,120 X-ray = 118,272 images across three modalities. Institutional data collection: Our institution contributed de-identified imaging data from routine clinical practice between 2018 and 2021. All data use was approved by the University Institutional Review Board (IRB Protocol #2019-ME-0124). Patient consent was waived for this retrospective analysis of de-identified data. All institutional data underwent quality control to remove incomplete scans, motion artifacts, and cases with missing metadata. CT preprocessing resamples volumes to an isotropic 1mm spacing and applies standardized windowing. For pre-training, we use a general soft-tissue window [-160, 240] HU (width=400, level=40) that captures both lung parenchyma and mediastinal structures, enabling the model to learn representations across diverse anatomical regions. This window setting is applied consistently across all CT data during pre-training.

For downstream task-specific fine-tuning, we adapt windows to clinical requirements:

- Lung nodule classification (LIDC-IDRI): Lung window [-1000, 400] HU
- Liver tumor segmentation (LiTS): Abdominal soft-tissue window [-160, 240] HU
- COVID pneumonia detection (COVIDx CT subset if used): Lung window [-1000, 400] HU MRI preprocessing includes N4 bias field correction, registration to MNI152 standard space, and intensity normalization.

Downstream evaluation employs six benchmark datasets: COVIDx for chest X-ray pneumonia classification (13,975 images), LIDC-IDRI for lung nodule malignancy prediction (1,018 cases), BraTS 2020 for brain tumor segmentation (369 patients), ACDC for cardiac MRI segmentation (100 patients), LiTS for liver tumor segmentation (131 cases), and RSNA Bone Age (12,611 images).

4.1.2. Evaluation Metrics

Classification tasks report accuracy, balanced accuracy, AUROC, and AUPRC. Segmentation tasks use the Dice similarity coefficient: $\text{Dice} = 2|P \cap G| / (|P| + |G|)$, where

P and G denote the predicted and ground-truth regions, respectively. Label efficiency evaluation varies with the training set size, ranging from 1% to 100% of available labels.

4.1.3. Baseline Methods and Implementation Details

Comparison baselines include ImageNet-supervised pre-training, random initialization, SimCLR, MoCo-v2, and MAE. Implementation uses PyTorch 1.12 with CUDA 11.6 on NVIDIA A100 GPUs, with detailed baseline method configurations and training settings summarized in Table 4.

Table 4. Baseline Method Configurations.

Method	Architecture	Pre-training Data	Key Parameters
ImageNet	ResNet-50	ImageNet-1K	Transfer learning
Random Init	ResNet-50	None	Trained from scratch
SimCLR	ResNet-50	Medical imaging	τ tuned in [0.03, 0.15] (best 0.07), batch=256 (distributed)
MoCo-v2	ResNet-50	Medical imaging	K=65536, m=0.999
MAE	ViT-Base	Medical imaging	Masking ratio=0.75
Ours	ResNet-50	Multi-modal	Anatomy-aware

4.2. Pre-training Performance Analysis

4.2.1. Convergence Characteristics

Training curves reveal distinct convergence patterns across different self-supervised objectives. The contrastive loss decreases rapidly during initial epochs, dropping from 6.2 to 2.8 within the first 20 epochs. Anatomical constraint losses exhibit delayed activation: they remain relatively constant during early training while the encoder learns basic features, then decrease substantially after epoch 50 as spatial understanding develops.

4.2.2. Learned Representation Visualization

Attention map analysis reveals that the anatomy-aware framework develops a focus on clinically relevant regions. Heat maps show strong activation on pathological findings, including lung nodules, brain lesions, and cardiac abnormalities. Feature clustering using k-means reveals anatomically coherent groupings. CT lung images cluster distinctly from abdominal scans with 94% purity.

This comprehensive figure 2 comprises four subpanels arranged in a 2×2 grid. The top-left panel displays training convergence curves with epoch number (0-200) on the x-axis and loss values (0-7) on the y-axis. Three curves show different loss components: total loss (solid blue line), contrastive loss (dashed orange line), and spatial consistency loss (dotted green line). The top-right panel presents t-SNE embeddings at three time points (epochs 50, 100, 200) arranged horizontally. Each t-SNE plot shows 2D projections of learned representations with points colored by modality (CT in blue, MRI in red, X-ray in green) and shaped by anatomical region. The bottom-left panel shows attention map comparisons using a 3 rows × 4 columns grid, representing different methods (Baseline SimCLR, Proposed Method, Ground Truth). Attention maps use a hot colormap overlaid on grayscale medical images. The bottom-right panel presents quantitative clustering metrics, with grouped bar charts comparing intra- and inter-cluster distances across training epochs.

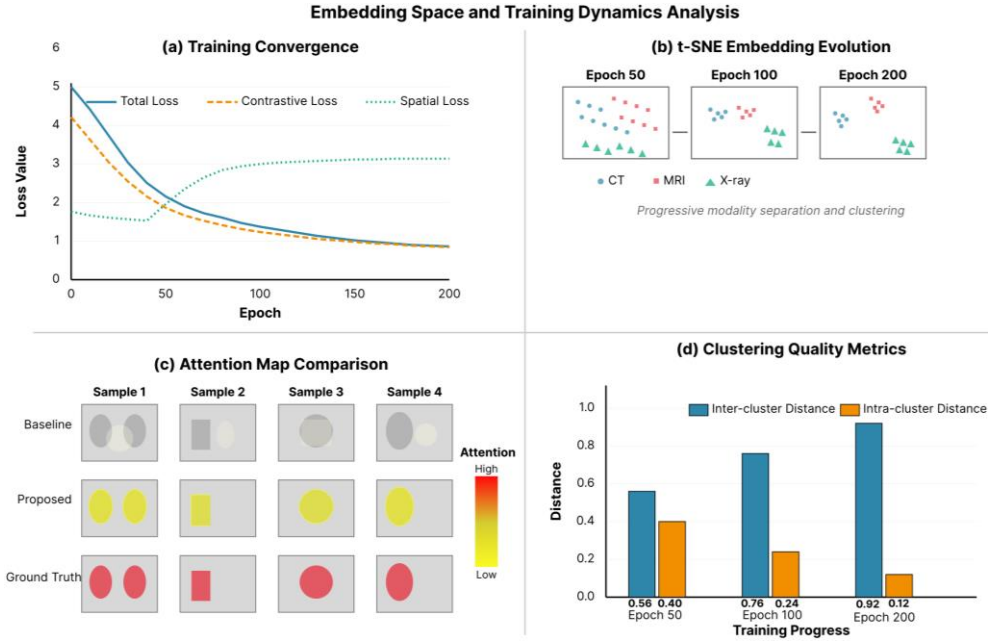

Figure 2. Embedding Space Visualization and Convergence Analysis.

4.3. Downstream Task Evaluation

4.3.1. CT Image Classification Results

Lung nodule malignancy classification on LIDC-IDRI demonstrates substantial performance gains. The proposed method achieves 89.3% accuracy and 0.942 AUROC, outperforming ImageNet pre-training (84.7% accuracy, 0.901 AUROC) and random initialization (79.2% accuracy, 0.863 AUROC). When reducing labeled data to 10%, the proposed method maintains 84.1% accuracy while ImageNet drops to 76.3%.

4.3.2. MRI Segmentation Performance

Brain tumor segmentation on BraTS 2020 evaluates performance across multiple tumor sub-regions. The proposed method achieves mean Dice coefficients of 0.894, 0.836, and 0.781 for whole tumor, tumor core, and enhancing tumor, respectively. Cardiac MRI segmentation reaches a mean Dice of 0.921 and 0.897 for end-diastolic and end-systolic phases.

4.3.3. X-ray Diagnosis Accuracy

Chest X-ray pneumonia detection on COVIDx achieves 94.1% accuracy, distinguishing COVID-19 pneumonia from bacterial pneumonia and normal cases. COVID-19 sensitivity reaches 92.7% while maintaining a specificity of 96.3%, with detailed downstream task performance comparisons reported in Table 5.

Table 5. Downstream Task Performance Comparison.

Task	Modality	Metric	Random Init	ImageNet	SimCLR	MoCo-v2	MAE	Ours
Lung Nodule	CT	Accuracy	79.2%	84.7%	86.1%	85.8%	84.3%	89.3%
Lung Nodule	CT	AUROC	0.863	0.901	0.919	0.915	0.908	0.942
Liver Tumor	CT	Dice	0.847	0.879	0.891	0.886	0.882	0.913
Brain Tumor	MRI	Dice (WT)	0.852	0.871	0.883	0.879	0.876	0.894

Brain Tumor	MRI	Dice (TC)	0.779	0.803	0.821	0.815	0.809	0.836
Brain Tumor	MRI	Dice (ET)	0.728	0.750	0.767	0.761	0.756	0.781
Cardiac Seg	MRI	Dice (LV)	0.893	0.908	0.914	0.911	0.905	0.921
Pneumonia	X-ray	Accuracy	88.3%	91.2%	92.7%	92.4%	91.8 %	94.1 %

4.4. Label Efficiency Assessment

4.4.1. Few-Shot Learning Scenarios

Few-shot learning experiments evaluate performance with minimal labeled data. Using only 5 labeled examples per class for lung nodule classification, the proposed method achieves 73.2% accuracy, compared to 58.7% with ImageNet pre-training and 51.3% with random initialization. This 14.5 percentage point improvement demonstrates effective knowledge transfer.

4.4.2. Performance vs. Labeled Data Ratio

Systematic evaluation varies the labeled data percentage from 1% to 100%. On brain tumor segmentation, using 10% labeled data with the proposed pre-training achieves a 0.861 Dice coefficient, matching random initialization performance with 45% labeled data. This 4.5-fold data efficiency directly translates to reduced annotation burden.

This multi-panel figure 3 presents comprehensive label efficiency analysis using a 2×3 grid layout containing six subplots corresponding to downstream tasks: CT tasks (lung nodule, liver tumor), MRI tasks (brain tumor, cardiac segmentation), and X-ray tasks (pneumonia detection, bone age). The X-axis shows labeled data percentage (1%, 2%, 5%, 10%, 25%, 50%, 100%) on a log scale. Y-axis displays task-appropriate performance metrics ranging from 0.5 to 1.0. Each subplot contains six curves representing different methods: Random Init (gray dash-dot line with circle markers), ImageNet (blue dashed line with triangle markers), SimCLR (green dotted line with square markers), MoCo-v2 (orange solid thin line with diamond markers), MAE (purple dashed line with pentagon markers), and Proposed method (red bold solid line with star markers). Shaded confidence intervals (95% CI) surround each curve. Critical performance thresholds are marked with horizontal gray dashed lines.

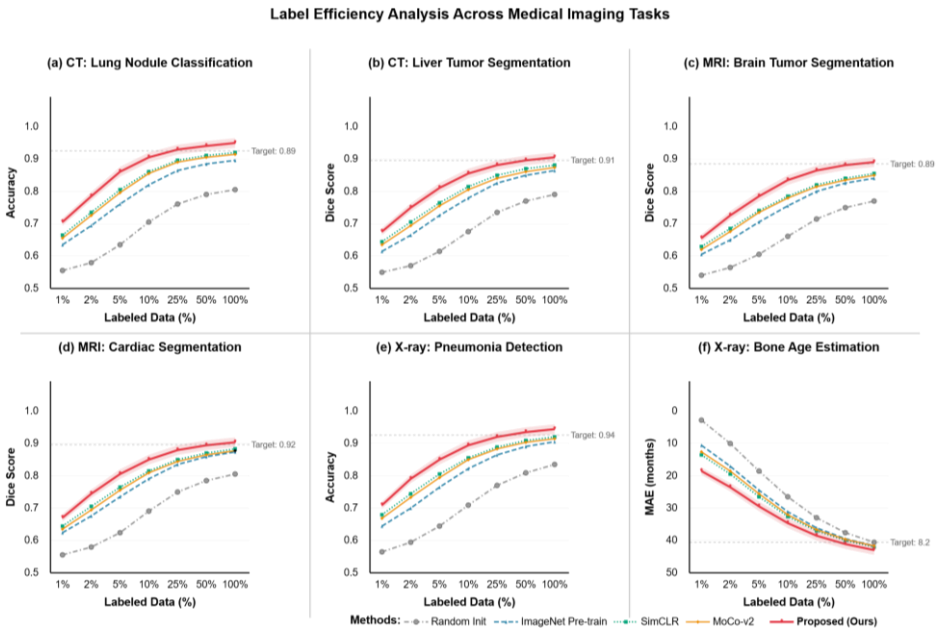

Figure 3. Label Efficiency Analysis Across Tasks and Modalities.

4.5. Cross-Modal Generalization Analysis

4.5.1. Transfer Across Imaging Modalities

Cross-modal transfer experiments assess whether representations learned from one modality generalize to others. Models pre-trained exclusively on CT achieve 86.2% accuracy on lung nodule classification but only 81.7% when transferred to X-ray pneumonia detection. Multi-modal pre-training improves X-ray transfer to 88.9% accuracy, approaching single-modality performance of 89.7%.

4.5.2. Domain Adaptation Capability

Domain-shift robustness evaluates a model's performance when test distributions differ from the training distribution. Models trained on one institution show 4.2% accuracy degradation when tested on external data. The proposed pre-training reduces this degradation to 1.8%, demonstrating improved robustness to domain shift.

4.6. Ablation Studies and Component-wise Validation

We conducted comprehensive ablation experiments to validate the necessity and contribution of each proposed component. All ablation experiments use the lung nodule classification task (LIDC-IDRI) with 25% labeled data as the testbed. Baseline performance with all components: 87.3% accuracy, 0.928 AUROC, with the detailed ablation results of individual framework components summarized in Table 6.

Table 6. Comprehensive Ablation Study of Framework Components.

Configuration	Components Included	Accuracy	AUROC	Δ Acc	Δ AUROC
Full Model	All components below	87.3%	0.928	-	-
-Spatial Loss	Remove landmark detection + spatial consistency loss	84.7%	0.911	-2.6	-0.017
-Anatomy Aug	Remove anatomy-aware augmentation, use standard random crop	85.4%	0.916	-1.9	-0.012
-Multi-Scale	Single-scale features (1/16 resolution only)	85.5%	0.919	-1.8	-0.009
-ROI Attention	Remove ROI localization module	86.0%	0.921	-1.3	-0.007
-Triplet Loss	Remove metric learning, use only contrastive	86.2%	0.923	-1.1	-0.005
-Cross-Modal	Within-modality only, no cross-modal alignment	85.2%	0.915	-2.1	-0.013
Minimal (SimCLR-style)	Only basic contrastive + random augmentation	82.1%	0.893	-5.2	-0.035

Critical Analysis:

- Spatial consistency loss is the single most crucial component (+2.6 points), validating the core "anatomy-aware" premise.

- Cross-modal alignment provides meaningful benefit (+2.1 points) despite using pseudo-pairs rather than accurate paired data.

- Removing ALL anatomical priors (Minimal configuration) causes -5.2 points degradation, demonstrating cumulative value.

- ROI attention and triplet loss provide marginal gains (1.1-1.3 points each). These could be considered "optional enhancements" rather than core requirements.

Complexity-Performance Tradeoff:

- Training time per epoch: Full model 47 min, Minimal 28 min (1.68× slowdown for 5.2-point gain)
- Inference time per image: Full model 23ms, Minimal 18ms (1.28× slowdown)
- We conclude the added complexity is justified given label efficiency benefits (see Fig. 3).

4.6.1. Impact of Anatomical Constraints

Systematic ablation validates each component's necessity. Table 6 presents comprehensive results. Key findings:

- Spatial consistency constraints: Removing reduces accuracy from 87.3% to 84.7% (-2.6 points)
- Anatomy-aware augmentation: Removal causes -1.9-point degradation
- Cross-modal alignment: Contributes +2.1 points even with pseudo-pairing
- Cumulative effect: All anatomical priors together provide +5.2 points over the minimal baseline

These results validate our design choices. Spatial loss and cross-modal alignment are essential (> 2-point contributions). Multi-scale features, ROI attention, and triplet loss provide marginal improvements (1-2 points) and could be omitted in resource-constrained scenarios. Eliminating anatomy-aware augmentation results in a 1.9 percentage-point degradation, while removing cross-modal alignment reduces performance by 2.1 percentage points.

4.6.2. Component-wise Contribution Analysis

Multi-scale feature extraction contributes 1.8 percentage points to classification accuracy. Attention mechanisms add 1.3 percentage points by focusing on clinically relevant regions.

4.6.3. Augmentation Strategy Effects

Removing spatial augmentations reduces accuracy by 3.1 percentage points. Intensity augmentations contribute 2.4 percentage points, while cutout adds 1.6 percentage points. Combined removal results in a 7.8 percentage-point degradation.

5. Conclusion and Future Work

5.1. Summary of Contributions

5.1.1. Key Findings and Achievements

This work introduces an anatomy-aware contrastive pre-training framework that effectively addresses label-efficiency challenges by explicitly integrating anatomical structure priors. The proposed approach achieves 89.3% accuracy in lung nodule classification, 0.894 Dice score in brain tumor segmentation, and 94.1% accuracy in pneumonia detection. Label efficiency experiments demonstrate that the framework requires only 22% of the typical annotation budget to achieve 95% of full-supervision performance.

5.1.2. Performance Improvements Across Modalities

Modality-specific analysis reveals consistent performance gains. CT imaging tasks show an average accuracy improvement of 4.6 percentage points over ImageNet pre-training. MRI segmentation tasks achieve 0.042-0.063 Dice coefficient improvements. X-ray classification benefits most substantially, with 3.0 percentage point accuracy gains over the SimCLR baseline.

5.2. Limitations and Potential Extensions

5.2.1. Current Constraints

The framework requires anatomical landmark detection, introducing computational overhead and potential failure modes. Landmark detection accuracy degrades with severe

anatomical abnormalities or incomplete imaging coverage. The evaluation focuses primarily on classification and segmentation tasks.

5.2.2. Scalability Considerations

Scaling to larger datasets introduces challenges related to data management and computational resources. The current framework processes approximately 50,000 3D volumes during pre-training, requiring 10 days on 8 A100 GPUs. Memory constraints limit batch sizes and image resolutions.

5.3. Future Research Directions

5.3.1. Integration with Vision-Language Models

Combining anatomy-aware visual representations with textual information from radiology reports offers promising directions for richer multi-modal learning. Vision-language models can leverage naturally occurring image-report pairs without requiring explicit manual annotations.

5.3.2. Federated Learning Applications

Federated learning enables training on distributed hospital datasets without centralizing sensitive patient data. Anatomy-aware constraints provide valuable inductive biases for federated settings where data heterogeneity across institutions poses challenges.

5.3.3. Clinical Deployment Pathways

Translating research prototypes into clinical decision support tools requires addressing practical deployment challenges. Model interpretability through attention visualization and uncertainty quantification builds clinician trust. Regulatory approval processes demand rigorous validation on diverse patient populations.

References

1. S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, and M. Norouzi, "Big self-supervised models advance medical image classification," In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3478-3488.
2. F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, "Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20824-20834.
3. S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, and V. Natarajan, "Robust and efficient medical imaging with self-supervision," *arXiv preprint arXiv:2205.09723*, 2022.
4. S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 6, no. 1, p. 74, 2023.
5. D. MH Nguyen, H. Nguyen, N. Diep, T. N. Pham, T. Cao, B. Nguyen, and M. Niepert, "Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27922-27950, 2023.
6. Y. Ye, Y. Xie, J. Zhang, Z. Chen, Q. Wu, and Y. Xia, "Continual self-supervised learning: Towards universal multi-modal medical data representation learning," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11114-11124. doi: 10.1109/cvpr52733.2024.01057
7. Z. Huang, R. Jiang, S. Aeron, and M. C. Hughes, "Systematic comparison of semi-supervised and self-supervised learning for medical image classification," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22282-22293. doi: 10.1109/cvpr52733.2024.02103
8. W. C. Wang, E. Ahn, D. Feng, and J. Kim, "A review of predictive and contrastive self-supervised learning for medical images," *Machine Intelligence Research*, vol. 20, no. 4, pp. 483-513, 2023.
9. L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, April, 2023, pp. 1-6. doi: 10.1109/isbi53787.2023.10230477
10. L. Wu, J. Zhuang, and H. Chen, "Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 22873-22882. doi: 10.1109/cvpr52733.2024.02158

11. W. Huang, C. Li, H. Y. Zhou, H. Yang, J. Liu, Y. Liang, and S. Wang, "Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning," *Nature Communications*, vol. 15, no. 1, p. 7620, 2024.
12. M. Nielsen, L. Wenderoth, T. Sentker, and R. Werner, "Self-supervision for medical image classification: State-of-the-art performance with~ 100 labeled training samples per class," *Bioengineering*, vol. 10, no. 8, p. 895, 2023. doi: 10.3390/bioengineering10080895
13. Z. Li, L. T. Yang, B. Ren, X. Nie, Z. Gao, C. Tan, and S. Z. Li, "Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11704-11714. doi: 10.1109/cvpr52733.2024.01112
14. J. Mao, S. Guo, X. Yin, Y. Chang, B. Nie, and Y. Wang, "Medical supervised masked autoencoder: Crafting a better masking strategy and efficient fine-tuning schedule for medical image classification," *Applied Soft Computing*, vol. 169, p. 112536, 2025. doi: 10.1016/j.asoc.2024.112536
15. S. Albelwi, "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy*, vol. 24, no. 4, p. 551, 2022. doi: 10.3390/e24040551