Medical Image Analysis (2023)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

DrasCLR: A Self-supervised Framework of Learning Disease-related and Anatomy-specific Representation for 3D Lung CT Images

Ke Yu[a,1,*], Li Sun[c,1], Junxiang Chen[b], Max Reynolds[b], Tigmanshu Chaudhary[b], Kayhan Batmanghelich[c]

[a] *School of Computing and Information, University of Pittsburgh, Pittsburgh, USA*
[b] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, USA*
[c] *Department of Electrical and Computer Engineering, Boston University, Boston, USA*

ARTICLE INFO

Article history:

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords:
Self-supervised learning
Contrastive learning
Label-efficient learning
3D Medical imaging data

ABSTRACT

Large-scale volumetric medical images with annotation are rare, costly, and time prohibitive to acquire. Self-supervised learning (SSL) offers a promising pre-training and feature extraction solution for many downstream tasks, as it only uses unlabeled data. Recently, SSL methods based on instance discrimination have gained popularity in the medical imaging domain. However, SSL pre-trained encoders may use many clues in the image to discriminate an instance that are not necessarily *disease-related*. Moreover, pathological patterns are often subtle and heterogeneous, requiring the ability of the desired method to represent *anatomy-specific* features that are sensitive to abnormal changes in different body parts. In this work, we present a novel SSL framework, named DrasCLR, for 3D lung CT images to overcome these challenges. We propose two domain-specific contrastive learning strategies: one aims to capture subtle disease patterns inside a local anatomical region, and the other aims to represent severe disease patterns that span larger regions. We formulate the encoder using conditional hyper-parameterized network, in which the parameters are dependant on the anatomical location, to extract anatomically sensitive features. Extensive experiments on large-scale datasets of lung CT scans show that our method improves the performance of many downstream prediction and segmentation tasks. The patient-level representation improves the performance of the patient survival prediction task. We show how our method can detect emphysema subtypes via dense prediction. We demonstrate that fine-tuning the pre-trained model can significantly reduce annotation efforts without sacrificing emphysema detection accuracy. Our ablation study highlights the importance of incorporating anatomical context into the SSL framework.

1. Introduction

While deep learning approaches have significantly advanced computer vision and many other fields (Voulodimos et al., 2018; Liu et al., 2020; Pouyanfar et al., 2018), efforts to apply these advancements to medical image analysis are still hampered by the scarcity of large-scale annotated datasets. Annotating medical images requires domain expertise and is a laborious and costly process, especially for 3D volumetric medical data. However, massive amounts of unlabeled raw images

*Corresponding author
e-mail: yu.ke@pitt.edu (Ke Yu)
[1] Equal contribution.

have been collected and stored in hospitals' picture archiving and communication systems (PACS) for decades. Recently, self-supervised learning (SSL) has become increasingly popular as a way to alleviate the annotation burden by exploiting the readily available unlabeled data (Jing and Tian, 2020; Ohri and Kumar, 2021; You et al., 2022b,a). However, unlike supervised approaches, which use experts' annotations (*e.g.*, disease labels, lesion segmentation masks) as supervision, self-supervised models are trained with limited supervision derived from the data itself, making it far more difficult to identify disease-related features from the data. Furthermore, certain lesions (*e.g.*, early-stage tumors) may occupy only a small region in high-resolution volumetric medical images, and their visual patterns may vary depending on where they are located in the body. Thus, the desired self-supervised learning algorithm should be sensitive enough to capture local anatomical deformities. In this research, we propose DrasCLR: a novel framework for self-supervised learning of disease-related and anatomy-specific representation of 3D medical imaging. DrasCLR learns a patch-based dense representation that conditionally depends on the anatomical location of the center voxel. We extensively evaluate our method on chest computed tomography (CT) imaging because of its prominent role in the prevention, diagnostics and treatment of lung diseases.

Self-supervised learning methods aim to provide useful feature representations for downstream tasks without human supervision, which is typically achieved by optimizing the model to solve a *proxy* task. When designing a proxy task, the primary consideration is: *what information in the data is important and what is not to the downstream tasks?* Early self-supervised approaches use heuristic-based pretext tasks to learn representations invariant to transformations that do not change the semantic meaning of the target labels (Doersch et al., 2015; Zhang et al., 2016; Gidaris et al., 2018). More recent contrastive learning approaches (Chen et al., 2020a; He et al., 2020) use *instance discrimination task*, which consider every instance as a class of its own and train deep neural networks to discriminate pairs of similar inputs (augmented views of the same instance) from a selection of dissimilar pairs (different instances). In this setting, data augmentation guided by prior knowledge often plays a vital role in preserving task-relevant information (Tian et al., 2020b). The sampling strategy for negative pairs is also crucial for the performance of contrastive learning methods. Recent studies (Jin et al., 2018; Jeon et al., 2021) show that hard negative sampling guided by domain knowledge helps in preventing trivial solutions and improving the alignment of extracted features with human semantics.

Self-supervised representation learning of disease-related features in medical images is particularly challenging for two reasons. First, since disease-related features are often represented through subtle changes, an effective self-supervised learning method should be able to ignore large but irrelevant and non-informative information, such as anatomical differences, and focus on representing fine-grained features, such as small deviations from normal-appearing tissues (Holmberg et al., 2020). Second, because pathological tissues may only scatter in a few small regions, adequately representing local content is crucial for dense (voxel-level) prediction tasks such as anomaly detection and segmentation. Several self-supervised learning methods (Zhou et al., 2019; Chaitanya et al., 2020; Haghighi et al., 2021) have been developed to learn local representations of 3D medical images. These methods use subvolumes sampled from random locations in the image as inputs and train a single encoder with parameters shared across all locations. However, disease types and their visual patterns are often associated with anatomical locations. For example, pulmonary emphysema can be divided into three major subtypes (*i.e.*, centrilobular, paraseptal, and panlobular) based on their visual characteristics and anatomical locations within the lung (Smith et al., 2014). A more sophisticated framework for learning local representations should incorporate anatomical locations as prior information to account for the spatial heterogeneity of anatomical and pathological patterns.

In this research, we take inspiration from the aforementioned challenges and propose a novel contrastive learning framework for 3D lung CT images. In order to represent disease-related

imaging features, we propose to combine two domain-specific contrasting strategies. The first strategy leverages the similarity across patients at the same anatomical location and aims to represent small disease patterns within a local (anchor) region. The second strategy takes advantage of anatomical similarities between the anchor and its nearby anatomical regions, with the goal of complementing the first strategy by learning larger disease patterns that expand beyond the local region. We use small 3D patches to represent local anatomical regions. The effectiveness of both strategies depends on the difficulty of instance discrimination; as the anatomical similarity between the query and negative patches becomes greater, the encoder is forced to rely on subtle and disease-related features rather than normal anatomical features. To that end, we use image registration to obtain *hard negative* patches from different subjects that are anatomically best aligned to the query patch. In particular, we obtain point-by-point correspondence between image pairs by mapping them to the same anatomical atlas. The coordinates in the atlas image can then be viewed as a standard set of anatomical locations. To incorporate anatomical locations into learned representations, we further develop a novel 3D convolutional layer whose kernels are conditionally parameterized through a routing function that takes the coordinates in atlas space as inputs. We call our unified framework **D**isease-**r**elated **a**natomy-**s**pecific **C**ontrastive **L**earning **R**epresentation (DrasCLR). The overview of our proposed approach is illustrated in Fig. 1. We conduct experiments on large-scale lung CT datasets. The results empirically show that our method outperforms baseline methods on both image-level and voxel-level tasks.

In summary, the major contributions of this paper are:

1. We propose a novel framework for contrastive learning of disease-related representation for 3D lung CT images.
2. We propose a novel 3D convolutional layer that encodes anatomical location-dependent information.
3. We extensively validate our model on large-scale lung CT datasets and show that our method outperforms existing baselines for a wide range of image-level tasks.
4. We demonstrate the application of our method for voxel-wise emphysema detection and show that using our pre-trained model can significantly cut annotation costs without compromising detection accuracy.

The paper is organized as follows: We present the details of our proposed methodology in Section 2. Implementation details and experimental results are described in Sections 3 and 4, respectively. We discuss the key findings and limitations of our work in Section 5. We survey the related works and draw comparisons with our preliminary work in Section 6. Finally, we conclude the paper in Section 7.

2. Method

We propose DrasCLR, a novel contrastive learning framework for 3D lung CT images. Our goal is to learn location-specific representations that are sensitive to tissue abnormalities. We start by aligning images to an *anatomical atlas* using image registration and treating the image of each patient as a collection of 3D patches centered at a common set of anatomical locations. Our contrasting strategies are motivated by two domain-specific similarity cues: one leverages the similarity between patients at the same anatomical location, and the second leverages the similarity between nearby anatomical locations on the same image. In the following sections, we explain each component separately. The schematic diagram of the proposed method is shown in Fig. 1. The notations used in this paper are summarized in Table 1.

2.1. Anatomical Alignment via Image Registration

We represent each volumetric image as a collection of 3D patches centered at a standard set of anatomical locations predefined on an anatomical atlas, with each patch corresponding to a distinct anatomical region of the lung. To align anatomical structures among patients, we first choose an image of a healthy subject to serve as the anatomical atlas, and then use image registration to obtain the subject-specific transformations that establish the point-by-point correspondence between the patients' images and the atlas image. Let X_{Atlas} denote the atlas image, x_i denote the image of patient i, the transformation ϕ_i is obtained

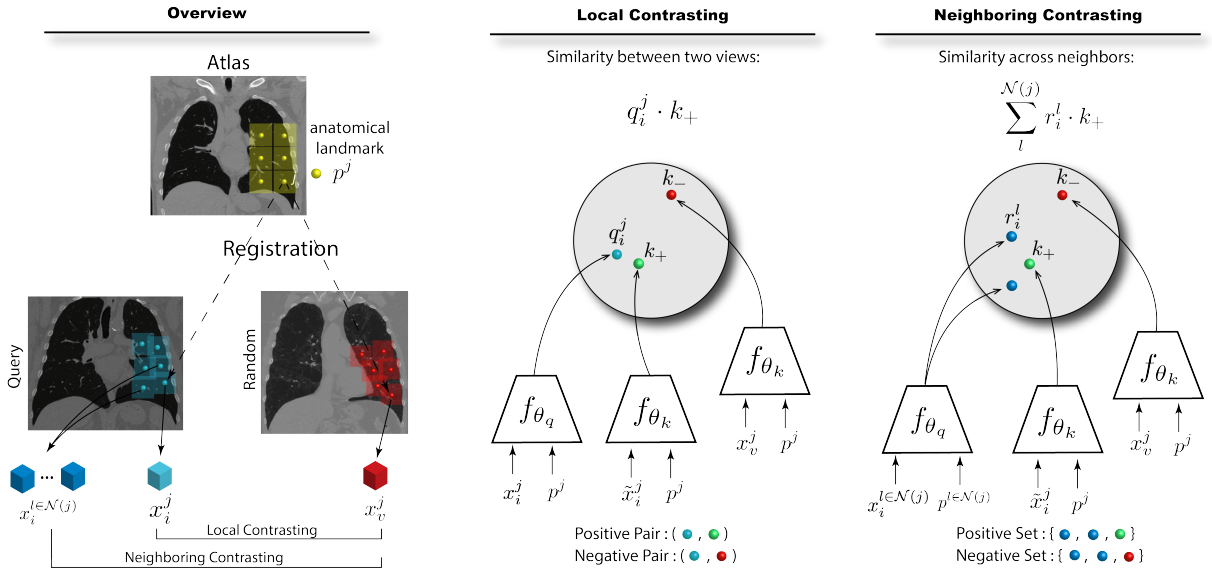

Fig. 1: Schematic diagram of DrasCLR. **Left panel**: We represent a volumetric image with a collection of 3D patches registered with distinctive anatomical landmarks defined in an atlas image. We develop an encoder that generates location-specific representation using the patch and location of associated anatomical landmark as inputs. Our contrastive learning framework comprises two objectives. **Middle panel**: The first one aims to learn local representation from a single patch. **Right panel**: The second one aims to learn representations of larger patterns across neighboring patches. Both contrasting strategies incentivize the encoder to learn disease-related features by using patches of similar anatomy as *hard* negative samples. Notation used in the diagram: i indexes images, j indexes patches; x_i^j and p^j are the query patch and its anatomical location; $x_i^{l\in\mathcal{N}(j)}$ are neighboring patches of x_i^j; x_v^j is a patch with the same anatomical location in a different image; $\tilde{x}_i^j$ represents random transformations of the given patch. q_i^j, r_i^l, k_+, k_- are embeddings of the query patch, neighboring patch, and positive and negative keys, respectively. $f_{\theta_q}, f_{\theta_k}$ are the encoder and momentum-updated encoder, respectively.

by solving the optimization problem as follows:

$$\underset{\phi_i}{\operatorname{argmin}}\ \texttt{Sim}(\phi_i(x_i), X_{\text{Atlas}}) + \texttt{Reg}(\phi_i), \tag{1}$$

where $\texttt{Sim}(\cdot,\cdot)$ is mutual information similarity function and $\texttt{Reg}(\phi_i)$ is a regularization term to ensure the transformation is smooth. We perform the image registration using the Advanced Neuroimaging Tools (ANTs) Tustison et al. (2014).

After registration, we divide the lung region of the atlas image into J evenly spaced three-dimensional patches with some overlap and define the patches' centers as the anatomical landmarks, denoted by $\{p^j\}_{j=1}^J$, where j is the patch index and each $p^j \in \mathbb{R}^3$ is a coordinate in the *atlas space*. We apply the inverse transformation ϕ_i^{-1} to locate the anatomical landmarks on each patient's image and extract the corresponding patches for training. Formally, each patient's image x_i is partitioned into a set of patches $\{x_i^j\}_{j=1}^J$ centered at $\{p_i^j\}_{j=1}^J$, respectively, where $x_i^j \in \mathbb{R}^{d\times d\times d}$, $p_i^j = \phi_i^{-1}(p^j)$ and d is the dimension of patch. It is straightforward to show that patches with the same index across all patients correspond to the same anatomical region on the anatomical atlas:

$$\phi_i(p_i^j) = \phi_i(\phi_i^{-1}(p^j)) = p^j. \tag{2}$$

2.2. *Conditionally Parameterized Convolutional Layer*

Image patches from different anatomical locations have distinctive anatomical features and may be associated with different diseased tissue patterns. Standard convolutional layers that apply the same kernels throughout the entire image may not be sufficient to accommodate spatial heterogeneity among patches at different locations. Inspired by CondConv (Yang et al., 2019), we propose Loc-CondConv, a location-dependent, conditionally parameterized convolutional layer. Instead of using static convolutional kernels, we compute convolutional kernels as a function of the anatomical location. In particular, we parameterize the kernels in Loc-CondConv as a linear combination of n convolutional kernels:

$$W = \alpha_1 W_1 + \cdots + \alpha_N W_N, \tag{3}$$

where $\{W_n\}_{n=1}^N$ are the same-sized convolutional kernels as in the regular convolutional layer and $\{\alpha_n\}_{n=1}^N$ are scalar weights

Table 1: Important notations in this paper.

Models	
$e(\cdot,\cdot;\theta_1)$	Image encoder.
$g(\cdot;\theta_2)$	MLP projection head.
$f(\cdot,\cdot;\theta)$	Network composed of e and g, where $\theta = \{\theta_1, \theta_2\}$.
$r(\cdot)$	Routing function used in Loc-CondConv.
Functions	
$\phi_i(\cdot)$	Transformation from the i-th image to the atlas space.
$\phi_i^{-1}(\cdot)$	Inverse transformation from atlas space to the i-th image.
$\tilde{t}(\cdot)$	Random augmentations.
Variables	
p^j	Location of the j-th anatomical landmark in the atlas space.
p_i^j	Location of the j-th anatomical landmark mapped in the i-th subject.
x_i^j	Patch of the i-th subject centering at the j-th anatomical landmark.
y_i^j	Representation of j-th patch in the i-th image used in downstream tasks.
y_i	Representation of the i-th image used in downstream tasks.
q_i^j	MoCo embedding of j-th patch in the i-th query image.
k_+	MoCo embedding of the positive sample.
k_-	MoCo embedding of the negative sample.
r_i^l	MoCo embedding of l-th neighboring patch in the i-th image.
$\mathcal{N}(j)$	Neighboring patches of the j-th patch.
X_{Atlas}	The atlas image.

computed via a routing function taking anatomical location as input. Specifically, we construct the routing function $r(\cdot)$ using a fully-connected layer followed by a Sigmoid activation function:

$$r(p^j) = \sigma(p^j \times W_r), \tag{4}$$

where p^j is a coordinate in the *atlas space* and W_r is a learnable weight matrix with dimension $3 \times N$, and σ represents the sigmoid function. Fig. 2 illustrates the architecture of Loc-CondConv. In the DrasCLR models, we replace all static convolutional layers with Loc-CondConv layers.

2.3. Local Contrastive Loss

In contrastive learning, the model is trained to discriminate pairs of positive inputs from a selection of negative pairs. Recent studies show that selecting harder negative pairs is critical for the success of contrastive learning (Saunshi et al., 2019; Robinson et al., 2020). The anatomical similarity between patients in the same lung region provides domain-specific cues for selecting hard negatives. More specifically, after registration alignment, any pair of patches centered at the same anatomical landmark, e.g., x_i^j, x_v^j $(i \neq v)$, have highly similar local anatomy, forcing the encoder to discriminate them using more subtle visual features, such as pathological tissues, rather than shortcuts, such as the overall anatomical background or boundaries.

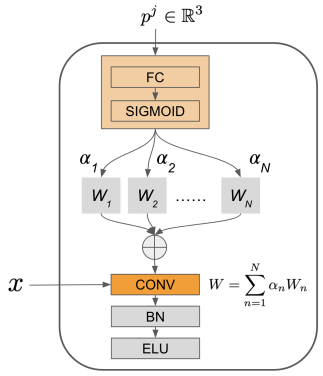

Fig. 2: The architecture of the Loc-CondConv layer. The kernels W are conditionally parameterized for each anatomical location p^j. The symbols α_n denote the routing weights. x denotes the input from the previous layer.

With this motivation, we propose a *local contrasting strategy*. Formally, given a patch x_i^j, we generate two augmented views $\tilde{x}_i^j = \tilde{t}(x_i^j)$, where $\tilde{t}$ is random augmentations sampled from a set of transformations $\mathcal{T}$. These two augmented patches are considered as a positive pair. Each negative sample is generated as $\tilde{x}_v^j = \tilde{t}(x_v^j)$ by randomly sampling a patch in the same anatomical region j from a different patient $(v \neq i)$ and random augmentations $\tilde{t} \sim \mathcal{T}$. For notation simplicity, the tilde symbol that represents random augmentations is omitted for the query and negative sample in subsequent text. We adopt the MoCo (He et al., 2020) as our contrastive learning paradigm, for its capability to efficiently leverage a large number of negative samples. Specifically, we train two networks f_{θ_q}, f_{θ_k} to map the positive pair $(x_i^j, \tilde{x}_i^j)$ and the negative pair (x_i^j, x_v^j) to corresponding embeddings as follows:

$$q_i^j = f(x_i^j, p^j; \theta_q),\ k_+ = f(\tilde{x}_i^j, p^j; \theta_k),\ k_- = f(x_v^j, p^j; \theta_k), \tag{5}$$

where $\theta_k = m\theta_k + (1-m)\theta_q$ and $m \in [0, 1)$ is a momentum coefficient. The network $f(\cdot,\cdot;\theta_q)$ is comprised of a feature extractor function $e(\cdot,\cdot;\theta_1)$, which accepts both patches and their corresponding anatomical landmarks as inputs, and a multilayer perceptron (MLP) projection head $g(\cdot;\theta_2)$, which maps the patch representations to the space where contrastive loss is applied. The equation can be written as $f(x_i^j, p^j; \theta_q) = g(e(x_i^j, p^j; \theta_1); \theta_2)$, where $\theta_q = \{\theta_1, \theta_2\}$. Finally, the *local con-*

trastive loss per location is defined as:

$$\mathcal{L}_l^j = -\log \frac{\exp(q_i^j \cdot k_+/\tau)}{\exp(q_i^j \cdot k_+/\tau) + \sum^{K^-} \exp(q_i^j \cdot k_-/\tau)}, \quad (6)$$

where K^- denotes the number of negative pairs and τ denotes the *temperature* hyperparameter.

2.4. Neighboring Contrastive Loss

The *local contrastive loss* incentivizes representations to be sensitive to tissue abnormalities within local anatomical regions. Pathological tissues, however, may expand beyond the borders of a single patch. We develop a complementary contrasting strategy - *neighboring contrasting* to allow the same encoder to learn disease patterns that may spread across multiple anatomical regions. For a given anatomical region j, we denote the indices of its ℓ nearest neighboring regions by $\mathcal{N}(j)$, its neighboring anatomical landmarks by $\{p^l\}_{l\in\mathcal{N}(j)}^{\ell}$, and the neighboring patches of x_i^j on the same image by $\{x_i^l\}_{l\in\mathcal{N}(j)}^{\ell}$. The corresponding embeddings of the neighboring patches are given by:

$$r_i^l = f(x_i^l, p^l; \theta_q),\ l \in \mathcal{N}(j). \quad (7)$$

Please note that random augmentations were applied to x_i^l in our implementation. For notation brevity, the tilde symbol is omitted.

Instead of constructing positive and negative pairs, we construct positive and negative *sets*, specifically,

$$\text{positive set} : \left\{\{x_i^l\}_{l\in\mathcal{N}(j)}^{\ell},\ x_i^j\right\},$$
$$\text{negative set} : \left\{\{x_i^l\}_{l\in\mathcal{N}(j)}^{\ell},\ x_v^j\right\},\ v \neq i,$$

in which the set of neighboring patches $\{x_i^l\}_{l\in\mathcal{N}(j)}^{\ell}$ serve as query samples, their corresponding central patch x_i^j acts as positive sample, and x_v^j, a patch from a random image at the same central location j, acts as the negative sample.

The *neighboring contrastive loss* per location is define as:

$$\mathcal{L}_n^j = -\log \frac{\sum_l^{\mathcal{N}(j)} \exp(r_i^l \cdot k_+/\tau)}{\sum_l^{\mathcal{N}(j)} \exp(r_i^l \cdot k_+/\tau) + \sum_l^{\mathcal{N}(j)} \sum_{k_-}^{\mathcal{K}^-} \exp(r_i^l \cdot k_-/\tau)}, \quad (8)$$

where k_+ and k_- are the same as defined in Eqn. 5. Minimizing this loss forces the encoder to extract similar visual features of the disease spreading across the patch x_i^j and its neighboring patches $\{x_i^l\}_{l\in\mathcal{N}(j)}^{\ell}$. Additionally, by selecting random patches in the same anatomical region as the hard negatives, the encoder is prevented from using mismatched anatomy as a shortcut to perform this *instance discrimination* task.

2.5. Overall Model

We train our model end-to-end by minimizing the combined *local contrastive loss* and *neighboring contrastive loss* and looping through each anatomical landmark. The overall loss function per location is defined as:

$$\mathcal{L}^j = \mathcal{L}_l^j + \mathcal{L}_n^j. \quad (9)$$

During inference time, the voxel-level representation can be obtained by:

$$y_i^j = e(x_i^j, p^j; \theta_1), \quad (10)$$

where $e(\cdot,\cdot;\theta_1)$ is the trained encoder with Loc-CondConv layers. The image-level representation y_i is obtained by averaging the representations of patches across all the anatomical landmarks. Formally, the representation at the image level is given by:

$$y_i = \frac{1}{J}\sum_{j=1}^{J} e(x_i^j, p^j; \theta_1). \quad (11)$$

Note that, at the time of inference, p^j can be any point inside the atlas space and is not restricted to the predefined anatomical landmarks. In our experiments, we obtain image-level representations using only predetermined anatomical markers for computational efficiency.

3. Implementation Details

We begin by extracting the lung regions from each CT scan using the lung segmentation method proposed by Hofmanninger et al. (2020). We then choose the image of one healthy subject as the anatomical atlas and partition it into a grid of 3D patches with some overlap. This results in 581 patches, each with a size of $32 \times 32 \times 32$, that fully cover the lung in the atlas image. Anatomical landmarks are defined as the centers of these 581 patches on the atlas coordinate system.

For registration, rather than registering raw images, we align the segmentation of the lung in the moving images to the lung

mask of the atlas. A healthy subject's lung, representing common shapes among most subjects, was chosen as the anatomical atlas. We use the image registration toolkit ANTs (Tustison et al., 2014) to obtain the forward and inverse affine transformations between each subject's (moving) lung segmentation and the atlas (fixed) lung segmentation.

We construct the encoder $e(\cdot, \cdot; \theta_1)$ using Loc-CondConv layers as the building blocks. Each Loc-CondConv layer contains N 3D-convolutional kernels with size $3 \times 3 \times 3$ and is zero-padded on each side of the inputs by one pixel. We adopt batch normalization (BN) (Ioffe and Szegedy, 2015) and ELU (Clevert et al., 2015) activation following each Loc-CondConv. For the projection head $g(\cdot; \theta_2)$, we adopt a 2-layer MLP with ReLU activation. We set the number of nearest neighbors used in the neighboring contrastive loss as 2 based on an ablation study (Sec. 4.4.3). We create data augmentations using MONAI (MONAI Consortium, 2020) package. The data augmentation includes random affine transforms (applied in the order of rotation, translation, and scale), Gaussian noise, and random image contrast adjustments. We optimize the networks using SGD with momentum = 0.9 and weight decay = 10^{-4}. The learning rate is set to be 10^{-2} and is updated using a cosine schedule. We choose the batch size of 128. Following the practice in MoCo-v2 (Chen et al., 2020b), we set temperature τ to 0.2 and momentum coefficient to 0.999. Unlike regular MoCo, which uses a single dictionary for negative samples, we develop a conditional memory bank that maintains separate dictionaries for anatomical landmarks, each of which has a size of 4096. For training, we select negative samples from the corresponding dictionary, which stores patch embeddings from the same anatomical location as the query patch. We perform self-supervised pretraining on the full dataset using four NVIDIA Tesla V100 GPUs, each with 32GB memory, for 48 hours or 20 epochs, whichever comes first.

4. Experiments

In this section, we take our DrasCLR pre-trained models and evaluate their performance in medical imaging tasks at both image and voxel levels. At the image level, we evaluate the effectiveness of the learned representation in disease phenotype prediction, disease severity classification, and survival analysis. At the voxel level, we first describe how our model can be used to produce voxel-wise segmentation masks. Using this approach, we then present the quantitative and qualitative results of subtype emphysema detection. Finally, we perform ablation studies to validate the importance of the proposed components in DrasCLR.

4.1. Datasets

We conduct the experiments on two large-scale lung CT datasets, including the COPDGene dataset (Regan et al., 2011) and the MosMed dataset (Morozov et al., 2020). We apply the same data preprocessing procedure for images in both datasets. We begin by re-sampling all images into $1mm^3$ isotropic resolution. We then threshold the Hounsfield Units (HU) to the intensity window of $[-1024, 240]$ and normalize the intensity range to $[-1, 1]$ by linear scaling.

4.1.1. COPDGene Dataset

Chronic Obstructive Pulmonary Disease (COPD) is a chronic inflammatory lung disease that causes obstruction of lung airflow and is one of the leading causes of death worldwide. The COPDGene Study (Regan et al., 2011) is a multi-center observational study that collects imaging data, genetic biomarkers, and relevant phenotypes from a large cohort of subjects. In our study, we use a large set of 3D thorax CT images from 9,180 subjects for self-supervised pre-training. We use the spirometry measures, disease-related phenotypes, and survival status of the same cohort as the image-level labels in our experiments. On a subset of these CT scans, an experienced pulmonologist annotated the bounding boxes of subtypes of emphysema by clicking on locations surrounded by the pathological tissues (Castaldi et al., 2013; Mendoza et al., 2012). This procedure created 696 centrilobular emphysema bounding boxes from 153 subjects, and 243 paraseptal emphysema bounding boxes from 69 subjects. All these bounding boxes are of the same size ($32mm^3$). We use this annotated subset to examine the performance of the DrasCLR pre-trained model for subtype emphysema detection.

4.1.2. MosMed Dataset

The MosMed dataset contains 3D thorax CT images of 1,110 subjects from the municipal hospitals in Moscow, Russia (Morozov et al., 2020). Subjects in this dataset are classified into five grades ("Zero", "Mild", "Moderate", "Severe", and "Critical") based on COVID-19 related CT findings and physiological measures, such as body temperature, respiration rate, blood oxygen saturation level (SpO2) and so on. Triage decisions are made based on the severity levels of the patients. For example, patients in the "Moderate" category only need to be followed up at home by a primary care physician, whereas patients in the "Critical" category are immediately transferred to the intensive care unit. We use the CT images in MosMed for model pre-training and use COVID-19 severity grades as classification labels in downstream analysis.

4.2. Image Level Evaluation

To assess how much disease-related information is preserved by the proposed method, we use the learned image-level representation to predict a wide range of clinical variables measured at the subject level, such as spirometry measurements, disease phenotypes, disease staging, and patients' survival rates.

4.2.1. COPD Phenotype Prediction

We begin by performing self-supervised pre-training with DrasCLR on the COPDGene dataset. Then, we use the learned image-level representations in downstream prediction tasks in a *linear readout* fashion. In particular, we train linear regression models to predict two pulmonary function measures on the log scale, which are percent predicted values of Forced Expiratory Volume in one second (`FEV1pp`) and its ratio with Forced vital capacity ($\texttt{FEV}_1/\texttt{FVC}$). We use R^2 scores as an evaluation metric for the regression analysis. In addition, we train multiclass logistic regression models to predict four categorical outcomes: (1) Global Initiative for Chronic Obstructive Lung Disease (GOLD) spirometric stage, a four-grade categorical variable indicating the severity of airflow limitation, (2) Centrilobular emphysema visual score (CLE), a six-grade categorical variable indicating the extent of emphysema in centrilobular, (3) Paraseptal emphysema visual score (Paraseptal), a three-grade categorical variable indicating the severity of paraseptal emphysema, and (4) Acute Exacerbation history (AE history), a binary variable indicating whether the patient has encountered at least one exacerbation event before enrolling in the study. For all classification tasks, we use accuracy as the evaluation metric. To account for human variability in annotation, for GOLD, CLE, and Paraseptal scores, we also report the proportion of times the predicted class fell within one class of the true score (denoted as *1-off*).

We compare the performance of DrasCLR against both unsupervised and supervised approaches. The unsupervised baselines include: Models Genesis (Zhou et al., 2021), MedicalNet (Chen et al., 2019), MoCo (3D version on the entire volume) (He et al., 2020), Context SSL (Sun et al., 2021), DomainCLR (Chaitanya et al., 2020), SwinUNETR (Tang et al., 2022), and DiRA (Haghighi et al., 2022). To ensure a fair comparion, each method was pre-trained on the corresponding dataset on which DrasCLR was pre-trained. Alongside deep learning-based baselines, we also evaluate methods that rely on expert-designed features. These approaches include the Divergence-based feature extractor (Schabdach et al., 2017), the K-means algorithm applied to features from local lung regions (Schabdach et al., 2017), and the widely-accepted clinical descriptor, Low Attenuation Area (LAA). The supervised baselines include convolutional neural networks (CNN) that were separately trained to predict `FEV1pp`, GOLD and CLE scores using 2D slices as inputs (2D CNN) (González et al., 2018), and Subject2Vec (Singla et al., 2018), where a patch-based CNN model was first trained with `FEV1` and $\texttt{FEV}_1/\texttt{FVC}$ as joint supervised information, and the learned image representations were then used in other prediction tasks. We perform five-fold cross-validation for all experiments and report the average results along with standard deviations. Table 2 shows that the DrasCLR pre-trained model outperforms unsupervised baseline models in all metrics, with the exception of 1-off accuracy for Paraseptal emphysema, where the difference is within one standard deviation. We have also conducted statistical tests (refer to Table A.5 in Appendix) to compare the evaluation outcomes

Table 2: Results of phenotype prediction on the COPDGene dataset. We use R-Square for continuous measurements and accuracy for discrete scores. Results including the mean and standard deviation (mean±s.d.) are derived from 5-fold cross validation. Our DrasCLR model has the best or competitive performance on all phenotype prediction tasks when compared to seven unsupervised methods, and it generalizes better than the supervised method for predicting visual scores and AE history.

Method	Supervised	Spirometry		COPD Staging		Visual scores				Acuity
		logFEV1pp	logFEV$_1$/FVC	GOLD	GOLD *1-off*	CLE	CLE *1-off*	Paraseptal	Paraseptal *1-off*	AE History
Metric		R-Square		% Accuracy						
LAA-950	✗	$0.44_{\pm.02}$	$0.60_{\pm.01}$	55.8	75.7	32.9	77.7	33.3	**87.6**	73.8
K-Means	✗	$0.55_{\pm.03}$	$0.68_{\pm.02}$	57.3	82.3	-	-	-	-	-
Divergence-based	✗	$0.58_{\pm.03}$	$0.70_{\pm.02}$	58.9	84.2	-	-	-	-	-
MedicalNet	✗	$0.47_{\pm.10}$	$0.59_{\pm.06}$	$57.0_{\pm1.3}$	$75.4_{\pm.9}$	$40.3_{\pm1.9}$	$69.6_{\pm1.6}$	$53.1_{\pm0.7}$	$81.8_{\pm0.8}$	$78.7_{\pm1.3}$
ModelsGenesis	✗	$0.58_{\pm.01}$	$0.64_{\pm.01}$	$59.5_{\pm2.3}$	$82.9_{\pm1.3}$	$41.8_{\pm1.4}$	$77.0_{\pm1.5}$	$52.7_{\pm.5}$	$85.3_{\pm1.1}$	$77.8_{\pm.8}$
MoCo	✗	$0.40_{\pm.02}$	$0.49_{\pm.02}$	$52.7_{\pm1.1}$	$67.6_{\pm1.4}$	$36.5_{\pm.7}$	$61.9_{\pm.9}$	$52.5_{\pm1.4}$	$79.7_{\pm1.2}$	$78.6_{\pm.9}$
DomainCLR	✗	$0.39_{\pm.02}$	$0.47_{\pm.01}$	$56.7_{\pm1.0}$	$75.8_{\pm.7}$	$39.9_{\pm.4}$	$71.7_{\pm1.3}$	$53.9_{\pm1.4}$	$82.2_{\pm1.2}$	$78.7_{\pm.6}$
SwinUNETR	✗	$0.54_{\pm.02}$	$0.64_{\pm.02}$	$59.8_{\pm.6}$	$81.0_{\pm.6}$	$42.3_{\pm1.1}$	$76.8_{\pm1.1}$	$52.4_{\pm.5}$	$84.4_{\pm.7}$	$78.3_{\pm1.0}$
DiRA	✗	$0.50_{\pm.03}$	$0.59_{\pm.02}$	$58.8_{\pm1.7}$	$78.3_{\pm1.2}$	$42.0_{\pm0.5}$	$72.0_{\pm0.6}$	$53.7_{\pm0.8}$	$83.2_{\pm0.8}$	$78.9_{\pm0.9}$
Context SSL	✗	$0.62_{\pm.01}$	$0.70_{\pm.01}$	$63.2_{\pm1.1}$	$83.6_{\pm.9}$	$50.4_{\pm1.3}$	$81.5_{\pm1.1}$	$56.2_{\pm1.1}$	$84.9_{\pm1.2}$	$78.8_{\pm1.3}$
2D CNN	✓	0.53	-	51.1	-	-	60.4	-	-	
Subject2Vec	✓	$\mathbf{0.67_{\pm.03}}$	$\mathbf{0.74_{\pm.01}}$	**65.4**	**89.1**	40.6	74.7	52.8	83.0	76.9
Ours	✗	$\underline{0.63_{\pm.01}}$	$\underline{0.71_{\pm.01}}$	$\underline{65.0_{\pm.6}}$	$\underline{85.6_{\pm.6}}$	$\underline{\mathbf{53.9_{\pm.8}}}$	$\underline{\mathbf{86.3_{\pm.7}}}$	$\underline{\mathbf{58.4_{\pm.8}}}$	$87.0_{\pm.8}$	$\underline{\mathbf{78.9_{\pm1.3}}}$

- indicates not reported.
Some baseline methods only report mean value without standard deviation in original manuscript.
The bold font is used to highlight the highest value for each column among *all* methods.
The underline is used to highlight the highest value for each column among *unsupervised* methods.

of DrasCLR with those of the baseline methods. The results show that our DrasCLR significantly outperforms the baseline methods for most of the downstream tasks. Our DrasCLR pre-trained model also outperforms the supervised baseline models, including Subject2Vec and 2D CNN, in terms of CLE, Paraseptal, and AE History predictions. For spirometry and COPD Staging, on which Subject2Vec were trained, the performance gap of our model is smaller compared to other unsupervised baseline models.

Overall, these results suggest that image-level features extracted by the DrasCLR pre-trained model preserve richer information about COPD severity than other unsupervised baselines. When compared to supervised methods, our proposed method learns more generalizable features as it achieves higher predictive performance for a broader range of clinical variables, such as emphysema visual scores and AE history.

4.2.2. Survival Analysis of COPD Patients

We evaluate the effectiveness of DrasCLR in survival analysis for the COPDGene population. We employ the Cox proportional hazards (CPH) model (Cox, 1972) to predict patients' survival using the learned image-level representations while controlling for five potential confounders, including age, gender, race, smoking status, and packyear (calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked). We compare the performance of features extracted by our method against: (1) hand-designed imaging features, (2) imaging features retrieved by other machine learning methods, and (3) relevant clinical features, such as spirometry measures and the BODE index (Celli et al., 2004). The hand-designed imaging features include CT metrics of emphysema, gas trapping, average wall thickness of hypothetical airway, and wall area percentage of segmental airways (Martinez et al., 2006). All comparison baselines use the same CPH model and are controlled by including the same five confounding variables.

We report the results in terms of time-dependent concordance index (C^{td}), which estimates the model's risk ranking ability, at each of the censoring period quantiles. Table 3 shows that the survival model with our imaging features achieved concordance scores of 0.76, 0.75, and 0.74 at the 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ quantiles, respectively, outperforming baselines with imaging-only features retrieved by both deep learning-based models and the hand-designed model. In comparison to clinical features, our method outperformed spirometry measures (0.76 vs 0.74) for risk stratification of near-term events before the 25$^{\text{th}}$ quantile,

Table 3: Time-dependent concordance index on the COPDGene dataset. Results are averages over five runs with bootstrapped standard errors. The highest mean values in each column are highlighted in bold. Our DrasCLR model performs the best when compared to other imaging representation approaches, and it provides incremental predictive value to clinical features.

Feature	**Method**	**Concordance Index**		
		t = 25th	t = 50th	t = 75th
Imaging	Hand-designed	$0.74_{\pm 0.02}$	$0.73_{\pm 0.01}$	$0.74_{\pm 0.01}$
	Models Genesis	$0.72_{\pm 0.02}$	$0.7_{\pm 0.01}$	$0.72_{\pm 0.01}$
	Subject2Vec	$0.72_{\pm 0.02}$	$0.72_{\pm 0.01}$	$0.72_{\pm 0.03}$
	DomainCLR	$0.74_{\pm 0.02}$	$0.72_{\pm 0.02}$	$0.72_{\pm 0.02}$
	SwinUNETR	$0.73_{\pm 0.02}$	$0.73_{\pm 0.01}$	$0.71_{\pm 0.01}$
	DiRA	$0.74_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.74_{\pm 0.02}$
	Context SSL	$0.74_{\pm 0.01}$	$0.74_{\pm 0.02}$	$0.74_{\pm 0.01}$
	Ours	$0.76_{\pm 0.02}$	$0.75_{\pm 0.01}$	$0.74_{\pm 0.01}$
Clinical	Spirometry	$0.74_{\pm 0.02}$	$0.75_{\pm 0.01}$	$0.74_{\pm 0.01}$
	BODE	$0.76_{\pm 0.01}$	$0.75_{\pm 0.01}$	$0.75_{\pm 0.00}$
Imaging + Clinical	Hand-designed + BODE	$0.76_{\pm 0.02}$	$0.76_{\pm 0.01}$	$0.76_{\pm 0.01}$
	Ours + Spirometry	$0.77_{\pm 0.02}$	$0.76_{\pm 0.00}$	$0.76_{\pm 0.01}$
	Ours + BODE	$\mathbf{0.78_{\pm 0.01}}$	$\mathbf{0.77_{\pm 0.01}}$	$\mathbf{0.77_{\pm 0.00}}$

Table 4: Classification of 5-grade COVID-19 severity on the MosMed dataset. The results are the means and standard deviations of accuracy for 5-fold cross validation. The highest mean value is highlighted in bold. Our DrasCLR model leads the best performance over both unsupervised and supervised approaches.

Method	Supervised	% Accuracy
3D CNN	✓	$61.2_{\pm 3.5}$
MedicalNet	✗	$62.1_{\pm 3.3}$
Models Genesis	✗	$62.0_{\pm 3.5}$
MoCo	✗	$62.1_{\pm 3.3}$
DomainCLR	✗	$63.2_{\pm 2.8}$
SwinUNETR	✗	$59.0_{\pm 2.9}$
DiRA	✗	$62.6_{\pm 2.7}$
Context SSL	✗	$65.3_{\pm 3.2}$
Ours w/o Neighbor Contrast	✗	$62.6_{\pm 2.4}$
Ours	✗	$\mathbf{65.4_{\pm 2.5}}$

slightly underperformed the BODE index (0.74 vs 0.75) at the 75th quantile of censoring time, and achieved comparable accuracy otherwise. We also developed survival models with combined imaging and clinical features. The bottom rows of Table 3 show that the model using both our imaging features and BODE index achieved the highest concordance scores of 0.78, 0.77, and 0.77 at the 25th, 50th, and 75th quantiles, respectively, demonstrating that the imaging representation learned by DrasCLR provides incremental predictive value for survival analysis of COPD patients.

4.2.3. COVID-19 Severity Prediction

We first pre-train a model with DrasCLR on the CT scans in the MosMed dataset. Then, we freeze the encoder and train a linear classifier to predict the severity of COVID-19, a categorical variable with five grades. The unsupervised comparison methods, consistent with those used in the COPDGene experiment, are also pre-trained on the MosMed dataset and evaluated using the same linear readout approach. Additionally, a supervised CNN model that uses entire 3D images (referred to as 3D CNN) is incorporated for comparison. To evaluate classification performance, we perform five-fold cross-validation and report the average test accuracy along with standard deviations.

Table 4 shows that the DrasCLR pre-trained model outperforms the unsupervised baseline models. Statistical tests (refer to Table A.7 in Appendix) further indicate DrasCLR's significant outperformance over Models Genesis, DiRA, and SwinUNETR, while its performance is comparable to Context SSL. Both Context SSL and our methods leverage the context between neighboring anatomical regions for representation learning. Context SSL incorporates this information via a graph neural network, whereas our method uses a neighboring contrasting strategy. The ablation study results in Table 4's bottom rows show that leveraging anatomical context from large regions is useful for categorizing COVID-19 severity. With the neighboring contrastive loss, the COVID-19 severity prediction accuracy increases by 2.8%. Interestingly, we found that the supervised 3D CNN model performs the worst, suggesting that directly extracting features from the entire volume may have resulted in the loss of fine-grained information at local anatomy. It is also possible that the supervised model may not converge properly or becomes overfitted due to the small amount of training data.

4.3. Voxel Level Evaluation

To show that the DrasCLR pre-trained model encodes fine-grained information at local anatomy, we demonstrate its ability to detect two subtypes of emphysema (*i.e.*, centrilobular and paraseptal emphysema), which are prevalent in different pulmonary regions. We perform the experiments in three aspects. First, we conduct a quantitative evaluation for emphysema detection via voxel-wise classification. Second, we show the qualitative results of predicted emphysema masks for COPD patients at different stages. Third, we perform emphysema de-

tection on a group of randomly selected subjects and show the relationship between the detected emphysema volume and the patient's COPD stage. In addition, we demonstrate that our method can reduce annotation efforts through transfer learning.

4.3.1. Emphysema Detection via Dense Classification

We propose to use dense or voxel-level classification for emphysema detection. In particular, we first pre-train a model with DrasCLR on all CT scans from the COPDGene dataset. We then fine-tune the model in a binary classification task to discriminate between emphysema-annotated patches and healthy patches. The healthy patches are sampled from random subjects with the criteria that the subjects' GOLD scores are equal to zero, and no centrilobular or paraseptal emphysema is found based on their image-level visual scores. For evaluation, we employ two different fine-tuning schemes: (1) Linear readout, which utilizes the pre-trained encoder as a fixed feature extractor and training a new linear classifier. This is compared to the model employing expert-designed features; and (2) Full fine-tuning, which involves appending a linear classifier to the pre-trained encoder and fine-tuning all network layers. This approach is employed for the unsupervised comparison baselines. During inference time, the fine-tuned model is used to perform voxel-wise classification of emphysema. For a given patch, the detection is positive if 25% voxels within it have a predicted probability of emphysema greater than 0.5; otherwise, the detection is negative.

Table 5 presents the quantitative performance of our method and comparison baselines for emphysema detection. Under the linear readout scheme, our model is compared to a logistic regression model using patch-level LAA features. Under the full fine-tuning scheme, our model is compared against a patch-based CNN, trained from scratch on the same annotated set of patches, and against unsupervised models pre-trained on the entire COPDGene dataset. The results in Table 5 show that, in comparison to the clinical descriptor LAA, the linear model utilizing our learned features achieves superior F1 scores for both subtype emphysema detection. Through full fine-tuning, our model achieves the highest scores across all metrics for paraseptal emphysema detection. For centrilobular emphysema detection, our approach achieves a superior F1 score compared to the comparison methods, while Context SSL exhibits a performance on par with our model.

To qualitatively demonstrate the outcomes of our method, we create voxel-wise emphysema segmentation for subjects at different stages of COPD. In particular, we first use the full fine-tuning model to estimate the probability of emphysema in a sliding-window fashion with a step size of 1 voxel. We then use a 0.5 threshold to map voxels with emphysema probability greater than or equal to the threshold to 1 and all other voxels to 0. Fig. 3 shows the predicted segmentation masks of two subtypes of emphysema of varying COPD stage in coronal and 3D views. We find that as the COPD severity increase (higher GOLD score), the volume of detected emphysema region increases in both subtypes. Furthermore, segmentation masks of GOLD scores 1 and 2 show a clear heterogeneity in the regional distribution of emphysema in the lung between these two subtypes. The regions of predicted segmentation are consistent with the clinical definition of emphysema subtypes, where centrilobular emphysema is commonly described as an abnormal enlargement of airspaces centered on the respiratory bronchiole (Leopold and Gough, 1957) and paraseptal emphysema refers to emphysematous change adjacent to a pleural surface (Heard et al., 1979).

Finally, we analyze the correlation between the total emphysema detected in 3D images and the subjects' COPD stages. In particular, we randomly select 500 subjects from the COPDGene dataset and use the fully fine-tuned model to make voxel-wise emphysema classification on their CT scans. Then, we aggregate all voxels in a CT scan to determine the fraction of voxels with a predicted probability of emphysema greater than 0.5. The box plots in Fig. 4 represent the distributions of detected emphysema proportion against GOLD scores, as well as the group with preserved ratio impaired spirometry (PRISm). Subplots of both centrilobular and paraseptal emphysema show positive correlations between the detected emphysema and patient's COPD severity.

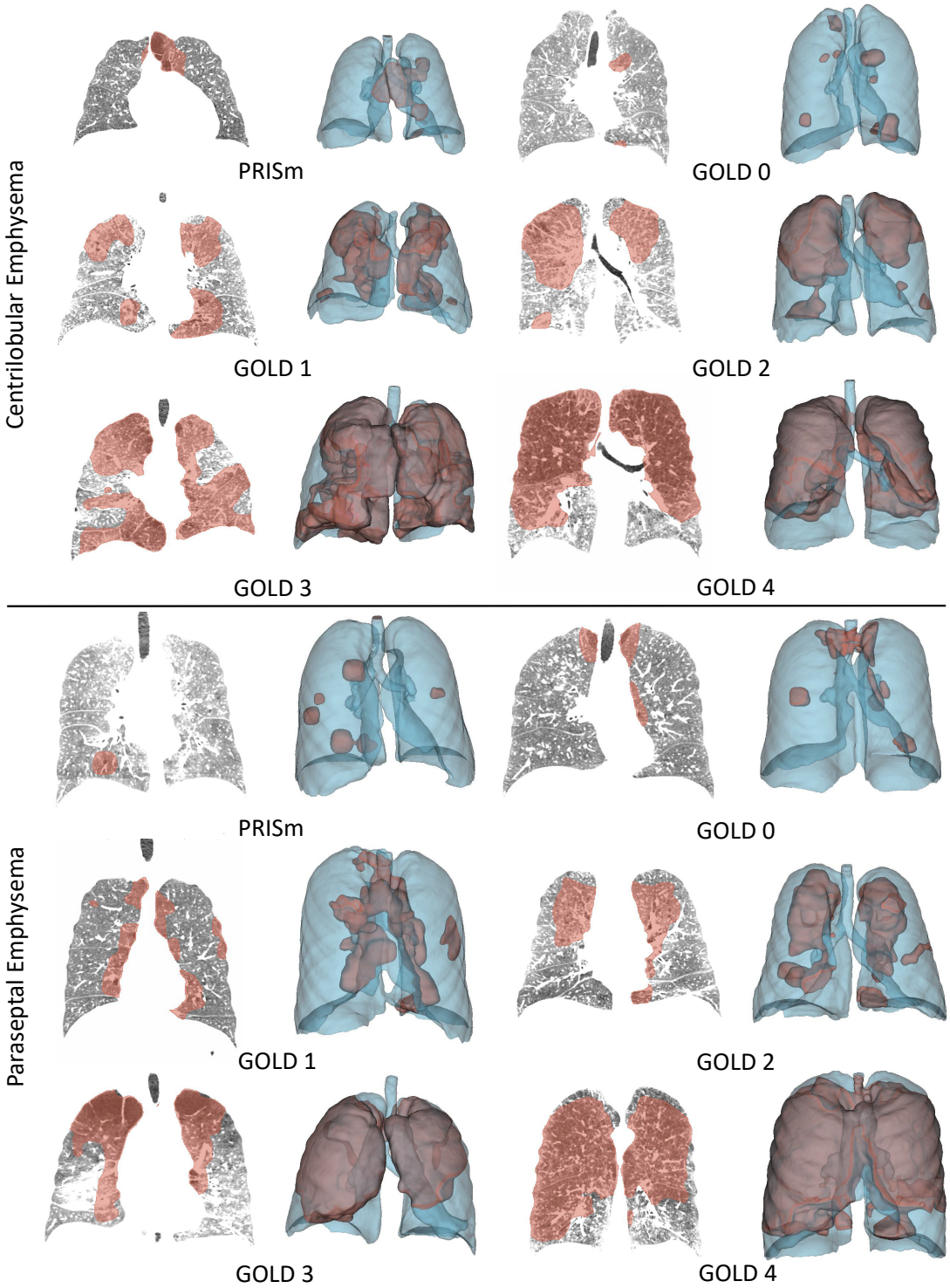

Fig. 3: Examples of predicted dense emphysema binary masks for subjects with different GOLD scores. The top three rows show the predicted regions of centrilobular emphysema, and the bottom three rows show the predicted regions of paraseptal emphysema. The intensity range is set as [-1060, -825] to better illustrate the emphysema. The predicted emphysema regions are plotted in red, and the lung regions are plotted in blue. As the severity of COPD increases (higher GOLD score), the detected region increases in both subtypes of emphysema. In addition, the predicted emphysema regions correspond to the clinical description of their subtypes.

Table 5: Evaluation for subtype emphysema detection. Results are the means and standard deviations (mean±s.d.) of F1, precision and recall scores for 5-fold cross validation. The highest mean values in each column are highlighted in bold. Our DrasCLR surpasses the clinical descriptor LAA in F1 score when using a linear readout, and achieves the top scores across all metrics for both subtype emphysema detections with full fine-tuning.

Scheme	Model	Centrilobular			Paraseptal		
		F1	Precision	Recall	F1	Precision	Recall
Linear Readout	Patch LAA	$0.77_{\pm.03}$	$0.94_{\pm.02}$	$0.65_{\pm.04}$	$0.70_{\pm.06}$	$0.83_{\pm.06}$	$0.61_{\pm.09}$
	Ours	$0.82_{\pm.03}$	$0.75_{\pm.04}$	$0.91_{\pm.02}$	$0.82_{\pm.03}$	$0.73_{\pm.05}$	$0.95_{\pm.01}$
Full Fine-tuning	Patch-based CNN	$0.91_{\pm.03}$	$0.92_{\pm.01}$	$0.89_{\pm.05}$	$0.82_{\pm.04}$	$0.77_{\pm.05}$	$0.88_{\pm.02}$
	Models Genesis	$0.81_{\pm.02}$	$0.71_{\pm.03}$	$0.95_{\pm.01}$	$0.82_{\pm.01}$	$0.76_{\pm.01}$	$0.90_{\pm.01}$
	SwinUNETR	$0.96_{\pm.03}$	$\mathbf{0.97_{\pm.03}}$	$0.94_{\pm.04}$	$0.93_{\pm.01}$	$0.94_{\pm.01}$	$0.91_{\pm.01}$
	Context SSL	$\mathbf{0.98_{\pm.01}}$	$\mathbf{0.97_{\pm.01}}$	$\mathbf{0.99_{\pm.01}}$	$0.98_{\pm.01}$	$0.96_{\pm.02}$	$0.99_{\pm.01}$
	Ours	$\mathbf{0.98_{\pm.01}}$	$\mathbf{0.97_{\pm.01}}$	$\mathbf{0.99_{\pm.01}}$	$\mathbf{0.99_{\pm.01}}$	$\mathbf{0.99_{\pm.02}}$	$\mathbf{1.00_{\pm.00}}$

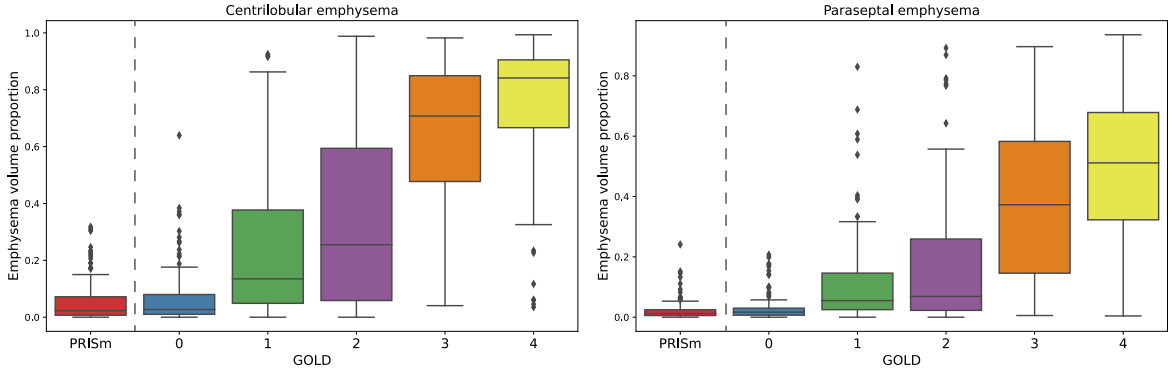

Fig. 4: Comparison of predicted volume proportion of centrilobular (left) and paraseptal (right) emphysema for subjects with different GOLD scores. A higher GOLD score indicates a more severe stage of COPD. The results show that with an increasing GOLD score, the predicted emphysema volume proportion in CT scan becomes higher.

4.3.2. Improve Annotation Efficiency via Transfer Learning

Transfer learning makes use of the knowledge of underlying data structure learned by the pre-trained models and has been demonstrated to be beneficial in medical imaging analysis, where the amount of annotated data is often limited. We simulate the scenarios of using a subset of annotated data to investigate the power of our method in transfer learning. Specifically, we fine-tune the DrasCLR pre-trained model by starting with 10% annotated emphysema patches and gradually increasing the amount of annotations by 10% in subsequent experiments. Fig. 5 shows the results of transfer learning on two target tasks. The performance of centrilobular detection learning from scratch with the entire dataset can be surpassed using DrasCLR with only 50% of the dataset, hence doubling the annotation efficiency. The performance of paraseptal detection learning from scratch with the entire dataset can be surpassed using DrasCLR with only 20% of the dataset, thus improving the annotation efficiency by five times. These results demonstrate how DrasCLR can significantly reduce the cost of manual image annotation, ultimately leading to more label-efficient deep learning.

4.4. Ablation Study

In this section, we conduct ablation experiments to validate the effects of several DrasCLR components.

4.4.1. Design Choices for Incorporating Anatomical Location

We demonstrate the effects of various designs for incorporating anatomical location information into DrasCLR. Specifically, we compare the following four approaches: (1) No conditioning, which uses a standard CNN to extract features from image patches without taking their anatomical locations into account; (2) Concatenation (Sun et al., 2021), which concatenates the features from the last layer of the standard CNN with the

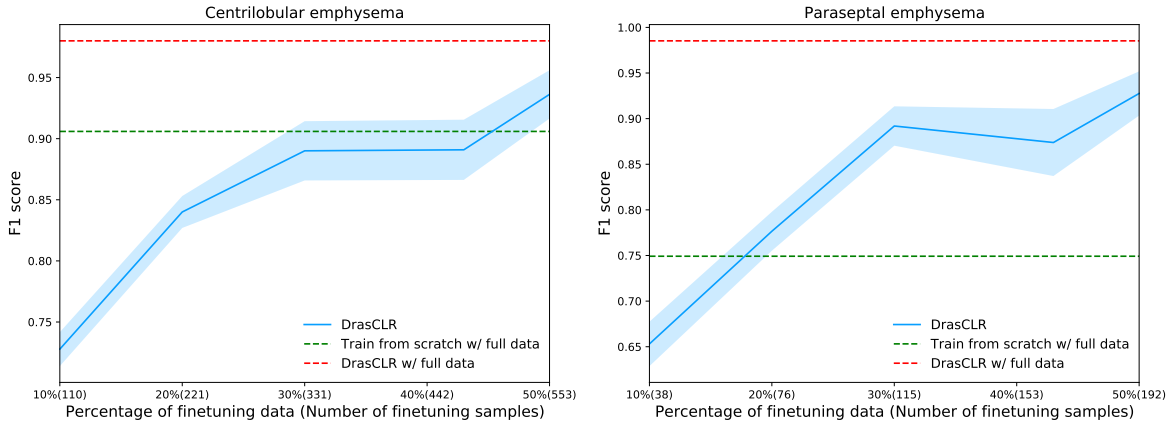

Fig. 5: Results of fine-tuning with different amounts of data. We perform evaluations for centrilobular (left) and paraseptal (right) emphysema detection. Compared to the model fine-tuned with full data from scratch (random initialization), the DrasCLR pre-trained model only needs 50% and 20% annotated data to achieve the same performance for centrilobular and paraseptal emphysema detection, respectively.

Table 6: Ablation study for how to incorporate anatomical location. We report R-Square for log`FEV1pp`, accuracy scores for the GOLD and CLE scores. The mean and standard deviation values are calculated via 5-fold cross-validation. The highest mean values in each column are highlighted in bold.

Method	log`FEV1pp`	GOLD	CLE
No conditioning	$0.57_{\pm.04}$	$61.8_{\pm1.1}$	$48.0_{\pm.9}$
Concatenation	$0.60_{\pm.01}$	$62.5_{\pm1.0}$	$49.2_{\pm1.1}$
HyperNetwork	$0.60_{\pm.01}$	$58.6_{\pm1.7}$	$44.1_{\pm1.3}$
Loc-CondConv (Ours)	$\mathbf{0.62_{\pm.02}}$	$\mathbf{63.4_{\pm1.0}}$	$\mathbf{50.3_{\pm.9}}$

anatomical coordinate and then fuses them through fully connected layers; (3) HyperNetwork (Ha et al., 2016), which uses a separate fully-connected network that takes an anatomical location as input to produce weights in the standard CNN; (4) A CNN consists of our proposed Loc-CondConv layers. We benchmark these different ways to incorporate anatomical locations on COPDGene dataset. We use the same network backbone as our DrasCLR, and train the model with local contrastive loss. As shown in Table. 6, the simple concatenation approach outperforms the standard CNN in three image-level tasks, suggesting that including anatomical location enriches the learned representations and enhances the performance of the downstream analysis. Furthermore, Table. 6 shows that the CNN with Loc-CondConv layers achieves the best performances in all target tasks, demonstrating it is a superior design for incorporating anatomical location information.

4.4.2. Impact of Anatomical Location for Disease Detection

We have thus far validated the importance of anatomical location as well as the effectiveness of Loc-CondConv layer for image-level prediction tasks. Our DrasCLR features Loc-CondConv layers, which use anatomical location as the condition to control the parameter of convolutional kernel for feature extraction. To investigate whether anatomical location is an important factor in representing fine-grained local features with DrasCLR, we explore the impact of perturbing the input locations of DrasCLR on the performance of emphysema detection. In particular, instead of extracting features from a given patch using its corresponding anatomical location used as the condition, we use random coordinate sampled in the lung as the condition. In order to evaluate the quality of extracted features, We train a linear classifier for emphysema classification using the extracted representation. We use the same dataset as described in Section 4.3.1. As shown in Table 7, by using random anatomical locations as inputs, the detection accuracy drops by 6% and 16% for centrilobular and paraseptal emphysema, respectively, showing statistically significant decreases (p-value$<$ 0.005, one-sided two sample t-test). The results indicate that the DrasCLR pre-trained encoder is sensitive to anatomical locations and captures anatomy-specific features.

Table 7: Sensitivity of the pre-trained DrasCLR model to anatomical location perturbation. The results are the means and standard deviations of emphysema detection accuracy for 5-fold cross validation. The highest mean values in each column are highlighted in bold.

Subtype	Random location	Patch location
Centrilobular emphysema	$73.3_{\pm 2.6}$	$\mathbf{79.2_{\pm 1.4}}$
Paraseptal emphysema	$56.4_{\pm 4.1}$	$\mathbf{72.1_{\pm 4.3}}$

Table 8: Ablation study for neighboring contrastive loss. We report R-Square for log`FEV1pp`, accuracy scores for GOLD and CLE scores. The mean and standard deviation values are calculated via 5-fold cross-validation. The highest mean values in each column are highlighted in bold.

Method	log`FEV1pp`	log`FEV1pp/FVC`	CLE
No Neighboring Contrast	$0.62_{\pm .02}$	$0.69_{\pm .02}$	$50.3_{\pm 0.9}$
# Neighbors = 1	$0.61_{\pm .01}$	$0.70_{\pm .01}$	$53.0_{\pm 1.3}$
# Neighbors = 2 (Ours)	$\mathbf{0.63_{\pm .01}}$	$\mathbf{0.71_{\pm .01}}$	$\mathbf{53.9_{\pm 0.8}}$
# Neighbors = 3	$\mathbf{0.63_{\pm .01}}$	$\mathbf{0.71_{\pm .01}}$	$52.5_{\pm 0.5}$

4.4.3. Neighboring Contrastive Loss

In this experiment, we investigate the effect of neighboring contrastive loss as well as the impact of the number of neighbors used. We pre-train DrasCLR models with no neighboring contrast as well as with different numbers of neighbors. Using a linear readout scheme, we benchmark the performance of the pre-trained models in downstream image-level tasks. As shown in Table 8, the incorporation of spatial context from neighboring patches enhances the performance of image-level tasks in most situations. The improvement is particularly notable in the prediction of the centrilobular visual score, which ranges from 5% to 8% depending on the number of neighbors. This is likely due to the fact that CLE grades are determined by the extent to which the lung's center is damaged by the disease, and the neighboring contrasting strategy encourages the learning of common disease patterns that span multiple anatomical regions.

4.4.4. Study the Robustness of DrasCLR to the Selection of atlas

Our DrasCLR framework requires a registration process to align anatomical landmarks between subjects, a step in the data preprocessing phase that involves selecting an atlas subject. In the experiments reported in the prior section, a healthy individual was chosen to serve this purpose. To assess the sensitivity of our method to the choice of atlas, we conduct an ablation analysis by pre-training the model using an unhealthy subject, specifically a subject with GOLD score of 3, as the atlas. We then replicate the experiments of COPD phenotype prediction with the same settings. The results in Table 9 show that the performance in most downstream tasks is consistent, regardless of whether the atlas subject is healthy or unhealthy. Statistical analyses (refer to Table A.6 in Appendix) confirm that there is no significant difference in outcomes when leveraging the original or the new atlas. These results suggest that our DrasCLR is robust against the choice of the atlas subject.

5. Discussions

5.1. Does DrasCLR learn disease-related features?

In this paper, we propose a novel self-supervised method, DrasCLR to learn *disease-related* representation of 3D lung CT images. Medical images have recurring and similar anatomy across patients. While previous self-supervised methods Zhou et al. (2021); Haghighi et al. (2021, 2022) used this knowledge to learn common anatomical representation, our method uses this knowledge to create hard negative samples in contrastive learning. As a result, our method is more sensitive to tissue abnormalities and can encode more disease-related information. As seen in Table 2, our method demonstrates excellent performance in predicting a wide range of clinical variables, such as spirometry measures and COPD phenotypes, which are closely associated with the degree of lung impairments in COPD patients. Emphysema is a hallmark of COPD. However, visually assessing emphysema at CT is time-consuming and subject to human variability. Our method demonstrates superior performance for this task and may provide a data-driven way of quantifying the visual score of emphysema from CT imaging.

Moreover, as evidenced by Table 3, imaging features learned by our method add incremental value to the BODE index for survival analysis in COPD patients, suggesting that our method is capable of capturing complementary risk factors from CT imaging. In addition to the evaluation on the COPD cohort, we demonstrate, as shown in Table 4, that DrasCLR can learn robust features associated with COVID-19 severity. The ability to represent disease-related information from medical images in

Table 9: Results of phenotype prediction on the COPDGene dataset with different atlas. The highest mean values in each column are highlighted in bold.

Method	Supervised	Spirometry		COPD Staging		Visual scores				Acuity
		logFEV1pp	logFEV$_1$/FVC	GOLD	GOLD *1-off*	CLE	CLE *1-off*	Paraseptal	Paraseptal *1-off*	AE History
Metric		R-Square		% Accuracy						
Ours (unhealthy atlas)	✗	$0.62_{\pm.02}$	$\mathbf{0.71_{\pm.02}}$	$64.6_{\pm1.4}$	$84.9_{\pm.8}$	$52.6_{\pm.8}$	$84.2_{\pm.4}$	$57.8_{\pm.9}$	$\mathbf{87.3_{\pm.8}}$	$\mathbf{79.2_{\pm1.5}}$
Ours (healthy atlas)	✗	$\mathbf{0.63_{\pm.01}}$	$\mathbf{0.71_{\pm.01}}$	$\mathbf{65.0_{\pm.6}}$	$\mathbf{85.6_{\pm.6}}$	$\mathbf{53.9_{\pm.8}}$	$\mathbf{86.3_{\pm.7}}$	$\mathbf{58.4_{\pm.8}}$	$87.0_{\pm.8}$	$78.9_{\pm1.3}$

an unsupervised manner is particularly useful during pandemic outbreaks, when labeled data is rarely available.

5.2. Does DrasCLR extract location-specific features?

A conventional convolutional layer is designed to be transitionally invariant. However, pathological patterns tend to be heterogeneous across locations in the human body, and an *one-size-fits-all* design may not be sufficient to learn the variety of tissue abnormalities at different anatomical locations. In this paper, we incorporate anatomical context into representation learning via two components in DrasCLR. First, image registration is used to provide a unified anatomical coordinate system, and all images are aligned to the atlas. Second, a novel Loc-CondConv layer is introduced to have modifiable weights that are conditionally dependent on anatomical location.

We have explored different ways to incorporate anatomical location into representation learning. As seen in Table 6, the Loc-CondConv layer is superior to simple concatenation and hypernetwork. To further validate the effect of our design, we investigate the impact of perturbing the input locations of DrasCLR on emphysema detection. As shown in Table. 7, the emphysema detection accuracy decreases when random location is used, demonstrating that lack of correct anatomical context is detrimental.

5.3. Why do we choose sliding-window based approach for emphysema segmentation?

The most well-known architecture for medical image segmentation is U-Net (Ronneberger et al., 2015), which is composed of opposing convolution and deconvolution layers, and spatial information is provided through skip connections to each decoder layer to recover fine-grained details. Traditionally, U-Net training relies on complete pixel/voxel-level annotation, a resource-intensive requirement that can be challenging to meet. Recent developments in weakly-supervised UNet methods (Dubost et al., 2017; Liu et al., 2022) have sought to mitigate this requirement using image-level labels. In our scenario, emphysema annotation in the COPDGene dataset was collected as bounding boxes by a physician clicking on regions surrounding by pathological tissues, thus the annotation is not complete at the voxel level. Our method can leverage these partially annotated images and produce voxel-wise emphysema classification in a sliding-window fashion. Moreover, the DrasCLR pre-trained model is capable of utilizing spatial information by modifying kernels' weights based on the input anatomical location. As illustrated by Fig. 3, our method can produce high-quality emphysema segmentation in which the regional distribution of detected emphysema matches with the clinical description of emphysema subtypes. The box plots in Fig. 4 show quantitative evidence that the detected emphysema volume correlates with the subjects' COPD stages.

5.4. Scope and limitations of the study

This study is specifically focused on lung CT scans. Our evaluation predominantly utilizes the COPDGene dataset, which is one of the largest publicly available 3D medical datasets, making it an excellent benchmark for evaluating the efficacy of self-supervised learning methods. While we have showcased the robustness of DrasCLR in a range of downstream tasks related to pulmonary diseases, its potential applications to other anatomical structures remain unexplored in our current research. Future directions could involve expanding the applicability of DrasCLR to interpret scans of other body regions or exploring its adaptability across different imaging modalities.

While in our methodology, image registration provides the dual advantages of aligning anatomical structures and estab-

lishing a consistent coordinate system within the atlas, this process is not without its challenges. The computational demands of image registration are considerable, often resulting in prolonged processing time and increased resource requirements. Specifically, within our implementation, applying the image registration process to the COPDGene dataset took approximately two days. Additionally, the quality of learned representations depends on the registration algorithm's accuracy and can be sensitive to factors like initialization, optimization strategy, and parameter settings. Inaccurate registration can lead to misalignment of structures, which can adversely affect subsequent analysis.

It is important to note that our contrastive learning strategy may not be universally optimal for all medical imaging analysis tasks. Specifically, our method emphasizes disease-related features by employing *hard* negative samples with aligned anatomies. While effective in enhancing sensitivity to deviations caused by diseased tissues, this approach may deprioritize features related to anatomical differences. In contrast, Chaitanya et al. (2020) specifically targeted organ segmentation as the primary task and employed negative samples derived from different anatomies to learn anatomical features. As organ segmentation is beyond the scope of this study, we did not assess our method's performance on this particular task.

6. Related Work

In the following sections, we review related works in four areas: (1) self-supervised learning approaches, including pretext task-based methods and contrastive learning methods, (2) applications of self-supervised learning in medical image analysis, (3) self-supervised learning methods that exploit anatomical context in medical images, and (4) conditionally parameterized networks. Following this review, we highlight the specific improvements of this paper in comparison to our preceding work.

6.1. Self-supervised Learning

Self-supervised learning has been shown to be an effective approach for learning semantically useful representations from large-scale unlabeled data without requiring human annotation (Jing and Tian, 2020; Ohri and Kumar, 2021). To generate supervisory signals from the data itself, a popular strategy is to present the model with various pretext tasks to solve. Commonly used pretext tasks include image inpainting (Pathak et al., 2016), image colorization (Zhang et al., 2016), relative position prediction (Doersch et al., 2015), image jigsaw puzzle (Noroozi and Favaro, 2016), patch cut and paste (Li et al., 2021a), temporal order verification (Misra et al., 2016), geometric transformation recognition (Gidaris et al., 2018), cross-modal correspondence (Korbar et al., 2018; Arandjelovic and Zisserman, 2017), and so on. These methods all have one thing in common: they build predictive-based pretext tasks using data's inherent structures, such as context similarity, spatial correlation, and temporal order. High-level semantic features are extracted in the process of accomplishing these tasks.

More recently, contrastive learning methods have emerged as one of the most popular self-supervised approaches due to their empirical success in computer vision (Misra and Maaten, 2020; Lee et al., 2021; Caron et al., 2021). The objective of contrastive training is to push learned representations to be similar for positive (similar) pairs and dissimilar for negative (dissimilar) pairs. This task is called *instance discrimination* (Wu et al., 2018) and is often formulated using the InfoNCE loss (Gutmann and Hyvärinen, 2010; Van den Oord et al., 2018). A variety of contrastive learning frameworks have been proposed, such as SimCLR (Chen et al., 2020a), which uses the augmented view of the same input as positive samples and the augmented views of other samples in a minibatch as negative samples, and MoCo (He et al., 2020), which uses a slow moving average (*momentum*) encoder and a dictionary that stores old negative representations to enable constructing very large batches of negative pairs. We extend MoCo as our contrastive learning paradigm. Rather than using a global dictionary, we develop a conditional memory bank that maintains distinct dictionaries of negative representations for each anatomical location. The design of sampling strategies for positive and negative pairs is a key driver to the success of contrastive learning (Saunshi et al.,

2019; Tian et al., 2020a). Previous studies have demonstrated that encoders trained with harder negative pairs can represent more challenging features (Jin et al., 2018; Jeon et al., 2021; Robinson et al., 2020; Kalantidis et al., 2020; Robinson et al., 2021). We create negative pairs from examples with highly similar local anatomy to force the model to solve instance discrimination using more subtle visual features (*e.g.* deviation from normal appealing tissues).

6.2. SSL Applications in Medical Imaging

Self-supervised learning is particularly useful for medical imaging analysis, in which labels are expensive to collect. Several studies have shown the effectiveness of self-supervised approaches in a variety of medical imaging analysis tasks such as disease diagnosis (Shurrab and Duwairi, 2021; Li et al., 2021b; Azizi et al., 2021), detection and localization (Tajbakhsh et al., 2019; Jiao et al., 2020; Lei et al., 2021), image segmentation (Taleb et al., 2020; Ross et al., 2018; Ye et al., 2022; He et al., 2022; Zeng et al., 2021; Zhang et al., 2022), and image registration (Li and Fan, 2018). Contrastive learning frameworks have been employed to leverage large-scale, unlabeled medical imaging data to produce pre-trained models. For example, Sowrirajan et al. (2021) adopted MoCo as a pre-training approach to obtain high-quality representations for detecting diseases in chest X-rays; Azizi et al. (2021) extended SimCLR to train robust representations for dermatology condition classification and thoracic disease classification by using the availability of multiple views of the same pathology from the same patients; Tang et al. (2022) combined multiple proxy tasks including volume inpainting, image rotation and SimCLR to pre-train a 3D Swin Transformer encoder. Our contrastive learning framework is built upon MoCo and is specifically designed to learn voxel-level representations that are sensitive to local anatomical deformities.

6.3. Leveraging Anatomical Structure in SSL

Medical images have consistent anatomy across patients, providing domain-specific cues for self-supervised representation learning. Zhou et al. (2021) introduced the Models Genesis, which learns image representation by recovering anatomical patterns from transformed sub-volumes extracted from CT images. Haghighi et al. (2021) extended the Models Genesis framework by adding a self-classification objective to enable the encoder to learn common anatomical semantics at similar body locations across patients. In a subsequent study, Haghighi et al. (2022) further enhanced their self-supervised learning framework by integrating an adversarial learning component. Bai et al. (2019) presented an anatomical position prediction task for learning segmentation features from cardiac magnetic resonance images. Chaitanya et al. (2020) enhanced SimCLR by integrating two domain-specific contrasting strategies: incentivizing similar representations for volumetric image slices coming from similar anatomical areas, and incentivizing distinctive local representations for different anatomical regions coming from the same image, both of which focus on learning features that represent anatomical differences, and the learned features are subsequently employed for organ segmentation. Our approach differs from the reviewed methods in two aspects: (1) We use image registration to improve the alignment of anatomical structures across patients, and (2) We leverage local anatomical similarity to create *hard negative* samples, thereby blocking shortcuts caused by anatomical differences and prioritizing sublet differences arising from tissue abnormalities.

6.4. Conditionally Parameterized Networks

In comparison to conventional neural networks with fixed weights, conditionally parameterized networks have weights computed as a function of the inputs. One such network is hypernetwork (Ha et al., 2016), which is typically a small network that outputs weights for a primary network. It has been used in functional image representation (Klocek et al., 2019) and hyperparameter optimization (Brock et al., 2017; Lorraine and Duvenaud, 2018; Hoopes et al., 2021). Another prominent conditional parameterization technique is CondConv (Yang et al., 2019), which creates convolutional kernels as a linear combination of experts with scalar weights dependent on input images. In our framework, the CondConv is modified to compute convolutional kernels as a function of the input anatomical locations.

As a result, the learned representation at a given voxel depends on both its surrounding patch and its anatomical location.

6.5. Our previous work

Sun et al. (2021) first presented a contrastive learning method (*i.e.*, Context SSL) that leverages anatomical context to detect deviation from normal-appearing tissues. This paper extends the preliminary version substantially with the following improvements:

1. We have introduced a novel neighboring contrastive loss, replacing the graph-level loss employed in Context SSL. This change offers two primary benefits: (1) Improved memory efficiency and the facilitation of end-to-end training, resolving previous memory limitations associated with using entire volumetric images; (2) Focused on a broader local context as opposed to the entire image, thereby mitigating shortcuts and fostering the model's ability to identify anomalies.
2. We have introduced the Loc-CondConv layer to represent anatomy-specific features. The ablation study, presented in Section 4.4.1, demonstrates the effectiveness of Loc-CondConv over the simple concatenation approach employed in Context SSL.
3. We have broadened our experimental scope to include survival analysis, subtype emphysema detection, and extensive ablation studies. The results, as shown in Tables 2, 3, 4, and 5, illustrate that DrasCLR surpasses Context SSL in mean evaluation performance. Hypothesis testing confirms that DrasCLR significantly outperforms in emphysema visual score classifications.
4. We have presented experimental results for two subtypes of emphysema detection tasks using limited amounts of annotated data in Section 4.3.2, demonstrating that transfer learning from DrasCLR could significantly reduce annotation efforts.
5. We've included qualitative examples of two emphysema subtypes segmented by DrasCLR in Figure 3, illustrating how the detected emphysema damage aligns with COPD severity and clinically-defined locations of each emphysema subtype.

7. Conclusion

In this paper, we present a novel method for anatomy-specific self-supervised representation learning on 3D lung CT images. We propose two domain-specific contrasting strategies to learn disease-related representations, including a local contrasting loss to capture small disease patterns and a neighboring contrasting loss to learn anomalies spanning across larger anatomical regions. In addition, we introduce a novel conditional encoder for location-specific feature extraction. The experiments on multiple datasets demonstrate that our proposed method is effective, generalizable, and can be used to improve annotation efficiency for supervised learning.

Acknowledgments

This work was partially supported by NIH Award Number 1R01HL141813-01 and NSF 1839332 Tripod+X. We are grateful for the computational resources provided by Pittsburgh SuperComputing grant number TG-ASC170024.

References

Arandjelovic, R., Zisserman, A., 2017. Look, listen and learn, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 609–617.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M., 2021. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3478–3488.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 541–549.

Brock, A., Lim, T., Ritchie, J.M., Weston, N., 2017. Smash: one-shot model architecture search through hypernetworks. arXiv preprint arXiv:1708.05344 .

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.

Castaldi, P.J., San José Estépar, R., Mendoza, C.S., Hersh, C.P., Laird, N., Crapo, J.D., Lynch, D.A., Silverman, E.K., Washko, G.R., 2013. Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. American journal of respiratory and critical care medicine 188, 1083–1090.

Celli, B.R., Cote, C.G., Marin, J.M., Casanova, C., Montes de Oca, M., Mendez, R.A., Pinto Plata, V., Cabral, H.J., 2004. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. New England Journal of Medicine 350, 1005–1012.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems 33, 12546–12558.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 .

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 .

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34, 187–202.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, pp. 1422–1430.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W.J., Vernooij, M., De Bruijne, M., 2017. Gp-unet: Lesion detection from weak labels with a 3d regression network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 214–221.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 .

González, G., Ash, S.Y., Vegas-Sánchez-Ferrero, G., Onieva Onieva, J., Rahaghi, F.N., Ross, J.C., Díaz, A., San José Estépar, R., Washko, G.R., 2018. Disease staging and prognosis in smokers using deep learning in chest computed tomography. American journal of respiratory and critical care medicine 197, 193–203.

Gutmann, M., Hyvärinen, A., 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings. pp. 297–304.

Ha, D., Dai, A., Le, Q.V., 2016. Hypernetworks. arXiv preprint arXiv:1609.09106 .

Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J., 2022. 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, pp. 20824–20834.

Haghighi, F., Taher, M.R.H., Zhou, Z., Gotway, M.B., Liang, J., 2021. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. IEEE transactions on medical imaging 40, 2857–2868.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738.

He, X., Fang, L., Tan, M., Chen, X., 2022. Intra-and inter-slice contrastive learning for point supervised oct fluid segmentation. IEEE Transactions on Image Processing 31, 1870–1881.

Heard, B., Khatchatourov, V., Otto, H., Putov, N., Sobin, L., 1979. The morphology of emphysema, chronic bronchitis, and bronchiectasis: definition, nomenclature, and classification. Journal of clinical pathology 32, 882.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G., 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental 4, 1–13.

Holmberg, O.G., Köhler, N.D., Martins, T., Siedlecki, J., Herold, T., Keidel, L., Asani, B., Schiefelbein, J., Priglinger, S., Kortuem, K.U., Theis, F.J., 2020. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nature Machine Intelligence 2, 719–726.

Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.V., 2021. Hypermorph: amortized hyperparameter learning for image registration, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 3–17.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448–456.

Jeon, S., Min, D., Kim, S., Sohn, K., 2021. Mining better samples for contrastive learning of temporal correspondence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1034–1044.

Jiao, J., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Self-supervised representation learning for ultrasound video, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1847–1850.

Jin, S., RoyChowdhury, A., Jiang, H., Singh, A., Prasad, A., Chakraborty, D., Learned-Miller, E., 2018. Unsupervised hard example mining from videos for improved object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 307–324.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence 43, 4037–4058.

Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D., 2020. Hard negative mixing for contrastive learning. Advances in Neural Information Processing Systems 33, 21798–21809.

Klocek, S., Maziarka, Ł., Wołczyk, M., Tabor, J., Nowak, J., Śmieja, M., 2019. Hypernetwork functional image representation, in: International Conference on Artificial Neural Networks, Springer. pp. 496–510.

Korbar, B., Tran, D., Torresani, L., 2018. Cooperative learning of audio and video models from self-supervised synchronization. Advances in Neural Information Processing Systems 31.

Lee, K.H., Arnab, A., Guadarrama, S., Canny, J., Fischer, I., 2021. Compressive visual representations. Advances in Neural Information Processing Systems 34.

Lei, W., Xu, W., Gu, R., Fu, H., Zhang, S., Zhang, S., Wang, G., 2021. Contrastive learning of relative position regression for one-shot object localization in 3d medical images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 155–165.

Leopold, J., Gough, J., 1957. The centrilobular form of hypertrophic emphysema and its relation to chronic bronchitis. Thorax 12, 219.

Li, C.L., Sohn, K., Yoon, J., Pfister, T., 2021a. Cutpaste: Self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664–9674.

Li, H., Fan, Y., 2018. Non-rigid image registration using self-supervised fully convolutional networks without training data, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1075–1078.

Li, X., Hu, X., Qi, X., Yu, L., Zhao, W., Heng, P.A., Xing, L., 2021b. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. IEEE Transactions on Medical Imaging 40, 2284–2294.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020. Deep learning for generic object detection: A survey. International journal of computer vision 128, 261–318.

Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D., 2022. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. Pattern recognition 122, 108341.

Lorraine, J., Duvenaud, D., 2018. Stochastic hyperparameter optimization through hypernetworks. arXiv preprint arXiv:1802.09419 .

Martinez, F.J., Foster, G., Curtis, J.L., Criner, G., Weinmann, G., Fishman, A., DeCamp, M.M., Benditt, J., Sciurba, F., Make, B., Mohsenifar, Z., Diaz, P., Hoffman, E., Wise, R., Group, N.R., 2006. Predictors of mortality in patients with emphysema and severe airflow obstruction. American journal of respiratory and critical care medicine 173, 1326–1334.

Mendoza, C.S., Washko, G.R., Ross, J.C., Diaz, A.A., Lynch, D.A., Crapo, J.D., Silverman, E.K., Acha, B., Serrano, C., Estépar, R.S.J., 2012. Emphysema quantification in a multi-scanner hrct cohort using local intensity distributions, in: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 474–477.

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717.

Misra, I., Zitnick, C.L., Hebert, M., 2016. Shuffle and learn: unsupervised learning using temporal order verification, in: European Conference on Computer Vision, Springer. pp. 527–544.

MONAI Consortium, 2020. MONAI: Medical Open Network for AI. URL: `https://github.com/Project-MONAI/MONAI`, doi:`10.5281/zenodo.4323058`.

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I.A., Gelezhe, P., Gonchar, A., Chernina,

V.Y., 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 .

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, Springer. pp. 69–84.

Ohri, K., Kumar, M., 2021. Review on self-supervised image recognition using deep neural networks. Knowledge-Based Systems 224, 107090.

Van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv e-prints , arXiv–1807.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S.S., 2018. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR) 51, 1–36.

Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D., 2011. Genetic epidemiology of copd (copdgene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease 7, 32–43.

Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S., 2020. Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 .

Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S., 2021. Can contrastive learning avoid shortcut solutions? Advances in Neural Information Processing Systems 34.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., Maier-Hein, L., 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery 13, 925–933.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., Khandeparkar, H., 2019. A theoretical analysis of contrastive unsupervised representation learning, in: International Conference on Machine Learning, PMLR. pp. 5628–5637.

Schabdach, J., Wells, W.M., Cho, M., Batmanghelich, K.N., 2017. A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 170–183.

Shurrab, S., Duwairi, R., 2021. Self-supervised learning methods and applications in medical imaging analysis: A survey. arXiv preprint arXiv:2109.08685 .

Singla, S., Gong, M., Ravanbakhsh, S., Sciurba, F., Poczos, B., Batmanghelich, K.N., 2018. Subject2vec: generative-discriminative approach from a set of image patches to a vector, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 502–510.

Smith, B.M., Austin, J.H., Newell Jr, J.D., D'Souza, B.M., Rozenshtein, A., Hoffman, E.A., Ahmed, F., Barr, R.G., 2014. Pulmonary emphysema subtypes on computed tomography: the mesa copd study. The American journal of medicine 127, 94–e7.

Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P., 2021. Moco pretraining improves representation and transferability of chest x-ray models, in: Medical Imaging with Deep Learning, PMLR. pp. 728–744.

Sun, L., Yu, K., Batmanghelich, K., 2021. Context matters: Graph-based self-supervised representation learning for medical images, in: Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, NIH Public Access. pp. 4874–4882.

Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., Ding, X., 2019. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 1251–1255.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3d self-supervised methods for medical imaging. Advances in Neural Information Processing Systems 33, 18158–18172.

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

Tian, Y., Krishnan, D., Isola, P., 2020a. Contrastive multiview coding, in: European conference on computer vision, Springer. pp. 776–794.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P., 2020b. What makes for good views for contrastive learning? Advances in Neural Information Processing Systems 33, 6827–6839.

Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., Avants, B.B., 2014. Large-scale evaluation of ants and freesurfer cortical thickness measurements. Neuroimage 99, 166–179.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience 2018.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742.

Yang, B., Bender, G., Le, Q.V., Ngiam, J., 2019. Condconv: Conditionally parameterized convolutions for efficient inference. Advances in Neural Information Processing Systems 32.

Ye, Y., Zhang, J., Chen, Z., Xia, Y., 2022. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 545–555.

You, C., Dai, W., Liu, F., Su, H., Zhang, X., Staib, L., Duncan, J.S., 2022a. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. arXiv preprint arXiv:2209.13476 .

You, C., Dai, W., Staib, L., Duncan, J.S., 2022b. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. arXiv preprint arXiv:2206.02307 .

Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y., 2021. Positional contrastive learning for volumetric medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 221–230.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: European conference on computer vision, Springer. pp. 649–666.

Zhang, Y., Sapkota, N., Gu, P., Peng, Y., Zheng, H., Chen, D.Z., 2022. Keep your friends close & enemies farther: Debiasing contrastive learning with spatial priors in 3d radiology images, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 1824–1829.

Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021. Models genesis. Medical image analysis 67, 101840.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 384–393.

Highlights

- Self-supervised method extracting disease-specific features from 3D medical data.
- Two domain-specific contrastive learning cues leverage anatomical similarities.
- Novel 3D convolutional layer with anatomical location-dependent kernels.
- Pretrained model generalizes well across downstream prediction tasks.
- Improved label-efficiency in lung CT image segmentation.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: