Learning Low-Rank Feature for Thorax Disease Classification

Rajeev Goel[*1] **Utkarsh Nath**[*1] **Yancheng Wang**[*1] **Alvin C. Silva**[2] **Teresa Wu**[1] **Yingzhen Yang**[1]

[1]School of Computing and Augmented Intelligence , Arizona State Univerisity, Tempe, AZ 85281, USA , {rgoel15,unath,ywan1053,Teresa.Wu,yingzhen.yang}@asu.edu
[2]Department of Radiology, Mayo Clinic Arizona Scottsdale, AZ 85259, USA silva.alvin@mayo.edu

Abstract

Deep neural networks, including Convolutional Neural Networks (CNNs) and Visual Transformers (ViT), have achieved stunning success in medical image domain. We study thorax disease classification in this paper. Effective extraction of features for the disease areas is crucial for disease classification on radiographic images. While various neural architectures and training techniques, such as self-supervised learning with contrastive/restorative learning, have been employed for disease classification on radiographic images, there are no principled methods which can effectively reduce the adverse effect of noise and background, or non-disease areas, on the radiographic images for disease classification. To address this challenge, we propose a novel Low-Rank Feature Learning (LRFL) method in this paper, which is universally applicable to the training of all neural networks. The LRFL method is both empirically motivated by the low frequency property observed on all the medical datasets in this paper, and theoretically motivated by our sharp generalization bound for neural networks with low-rank features. In the empirical study, using a neural network such as a ViT or a CNN pre-trained on unlabeled chest X-rays by Masked Autoencoders (MAE), our novel LRFL method is applied on the pre-trained neural network and demonstrate better classification results in terms of both multiclass area under the receiver operating curve (mAUC) and classification accuracy.

[*]Indicates equal contribution.

1 INTRODUCTION

A chest radiograph [Van Ginneken et al., 2001], commonly known as a chest X-ray, is the predominant diagnostic imaging method in diagnosing abnormal conditions in the airways, blood vessels, bones, heart, lungs, and other structures within the chest cavity. Following the huge success of deep learning in computer vision [He et al., 2016, Lin et al., 2017b,a], there is a growing interest in developing deep neural networks (DNNs) to detect abnormalities in anatomy in chest X-rays [Guendel et al., 2018, Xiao et al., 2023, Çallı et al., 2021, Guan and Huang, 2018]. The integration of DNNs into radiography practices can assist radiologists in providing more accurate and timely diagnoses. Accurate clinical decision-making with DNNs heavily relies on learning informative medical feature representation. Early works usually adopt convolutional neural networks (CNNs) such as U-Net [Ronneberger et al., 2015] and its variants [Falk et al., 2019, Zhou et al., 2018, Cui et al., 2019] for representation learning on radiography images. However, CNNs are biased in detecting local patterns in images, such as edges, shapes, and textures [Chefer et al., 2021, Bai et al., 2021, Dosovitskiy et al., 2020, Xiao et al., 2023]. Different from photograph image processing, long-range feature dependencies are crucial indicators of abnormalities in radiography images [Xie et al., 2021, Hatamizadeh et al., 2022, Shamshad et al., 2022, Xiao et al., 2023]. To address that issue, Visual Transformers (ViTs) are adopted to learn more informative medical representations from radiography images [Ma et al., 2022, Chen et al., 2021b, Xiao et al., 2023], utilizing their capabilities in capturing long-range feature dependencies via self-attention. Albeit the success of CNNs and ViTs in analyzing radiography images, their accuracy heavily relies on the quality and quantity of data and annotations [Feng et al., 2020]. However, the collection of large amounts of training data and high-quality annotations in the medical imaging domain is extremely hard [Zhou, 2021]. To tackle this problem, self-supervised learning (SSL) has been employed as a promising solution for acquiring representations from unlabeled data. Given the greater availability of unla-

beled medical images [Azizi et al., 2022], SSL proves to be an efficient approach for obtaining generalizable representations, which can subsequently be adapted to downstream tasks, even when labeled data is limited. SSL employs a range of pretext tasks to acquire transferable representations without manual annotations. Over recent years, numerous variations of self-supervised learning have surfaced using contrastive learning [Chen et al., 2020c,b, Grill et al., 2020, Caron et al., 2020] and restorative learning [Zhou et al., 2021b, Tang et al., 2022, Feng et al., 2020, Xiao et al., 2023].

Challenges in the Current Literature for Disease Classification. We study thorax disease classification in this paper. As detailed in Section 2.1 about the background for radiographic imaging, the disease areas on radiographic images can be subtle which exhibit localized variations, and such conditions are further complicated by the inevitable noise which is ubiquitously on radiographic images. Effective and robust extraction of features for the disease areas is crucial for disease classification on radiographic images. Although various neural architectures, such as CNNs and ViTs, and different training techniques, such as self-supervised learning with contrastive/restorative learning, have been employed for disease classification on radiographic images, there have been no principled methods which can effectively reduce the adverse effect of noise and background, or non-disease areas, for disease classification on radiographic images.

Our Contributions. The contributions of this paper are presented as follows. First, in order to address the aforementioned challenge, we propose a novel Low-Rank Feature learning (LRFL) method in this paper, which is universally applicable to the training of all neural networks with the application for thorax disease classification. Our LRFL method employs low-rank features for disease classification. The usage of low-rank features are empirically motivated by the low frequency property as shown in Figure 2. That is, the low-rank projection of the ground truth class labels possesses the majority of the information of the class labels. Inspired by this observation, our LRFL method adds the truncated nuclear norm as a low-rank regularization term to the training loss of a neural network so as to promote low-rank features. Because the actual features used for classification are approximately low-rank and the high-frequency features are significantly truncated, all the noise and the information about the background, or the non-disease areas on radiographic images in the high-frequency features are largely discarded and not learned in a neural network. As a result, the adverse effect of such noise and background is considerably reduced in a network trained by our LRFL method. Furthermore, we propose a new separable approximation to the truncated nuclear norm, so that standard SGD can be used to optimize the training loss with the approximate truncated nuclear norm. Extensive experimental results demonstrate that our LRFL method renders new record mAUC on three standard thorax disease datasets, NIH-ChestX-ray [Wang et al., 2017], COVIDx [Pavlova et al., 2022], and CheXpert [Irvin et al., 2019], surpassing the current state-of-the-art [Xiao et al., 2023] with the same pre-training setup.

Second, we provide theoretical result on the sharp generalization bound for our LRFL method, justifying the promising benefit of low-rank learning method in a representation learning framework. Due to our theoretical result and the fact that LRFL can be applied to the training of all neural networks, we expect that our LRFL method can be applied to classification of diseases other than thorax diseases, and generate even broader impact on general classification problems with radiographic images.

1.1 NOTATIONS

We use bold letters to denote matrices or vectors. $[\cdot]_i$ stands for the i-th row of a matrix. $\|\cdot\|_p$ denotes the p-norm of a vector or a matrix. $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm of a matrix. We use $[m \dots n]$ to indicate numbers between m and n inclusively, and $[n]$ denotes the natural numbers between 1 and n inclusively.

2 RELATED WORKS

2.1 RADIOGRAPHIC IMAGING

Radiographic imaging [Li et al., 2023] has long stood as a cornerstone in medical image analysis. Different from photographic images, where large objects usually lie in the center of images with diverse backgrounds [Deng et al., 2009], radiography images are generated based on fixed medical imaging protocols [Zhou, 2021, Li et al., 2022, Shamshad et al., 2022, Xiao et al., 2023]. As a result, the backgrounds of different radiography images usually exhibit consistent anatomy [Zhou et al., 2012]. In radiography images, some vital clinical details spread across their expanse. Meanwhile, areas indicating illness, which stand out in the foreground, frequently show more nuanced, detailed, and localized variations [Xiao et al., 2023, Suetens, 2017, Zhou et al., 2022c]. Such differences make radiographic imaging analysis much more challenging than photographic imaging analysis.

Noise is ubiquitous and inevitable in radiography images generated by medical imaging devices, which stems from various sources, including quantum fluctuations, electronic system interference, scatter radiation, motion blur, and overlapping anatomical structures [Siewerdsen et al., 1998, Manson et al., 2019, Chandra and Verma, 2020]. Among these, quantum noise is often identified as the primary source of noise in radiographic imaging [Chandra and Verma, 2020]. Quantum noise originates from the inherent statistical fluc-

tuations in the number of X-ray photons absorbed or transmitted through the object and detected by the imaging system [Sprawls, 1993, Shung et al., 2012, Suetens, 2017, Chandra and Verma, 2020]. These variations introduce graininess or mottling to the image, potentially obscuring fine details and diminishing tissue contrast. The level of quantum noise on a radiography image is affected by factors such as the X-ray dose, the detector sensitivity, and the thickness of the imaged object [Sprawls, 1993]. As the quantum noise is a result of the random manner in which X-ray photons are emitted and absorbed, its presence can be modeled as a Poisson process [Suetens, 2017, Chandra and Verma, 2020]. In addition, under conditions of high photon flux, the behavior of quantum noise can be approximated by a Gaussian distribution [Lee et al., 2018, Ding et al., 2018]. This approximation allows for the use of various image processing techniques, such as filtering [Ding et al., 2018], aimed at reducing noise and improving image quality.

2.2 MEDICAL IMAGE ANALYSIS WITH DEEP LEARNING

Following the remarkable achievements of deep learning in photographic image processing [He et al., 2016, Lin et al., 2017b,a], there has been a growing interest in harnessing deep neural networks to enhance medical image analysis due to their ability to learn complex representations. Ever since U-Net [Ronneberger et al., 2015, Falk et al., 2018, Zhou et al., 2018] wihch first shows the power of Convolutional Neural Networks (CNNs) in medical imaging, methods based on CNNs have demonstrated dominated performance in almost every field of medical imaging, including image classification [Shen and Gao, 2018, Wang et al., 2019, Ma et al., 2020], object detection [Falk et al., 2019, Zhou et al., 2018, Yang and Yu, 2021], and semantic segmentation [Yang and Yu, 2021, Yao et al., 2021, Zhou et al., 2018, Simpson et al., 2019, Sourati et al., 2019]. In addition, methods based on other techniques such as Recurrent Neural Networks (RNNs) [Zhou et al., 2019a, Gao et al., 2019] and Reinforcement Learning (RL) [Zhou et al., 2021a, Xu et al., 2022, Hu et al., 2023] have also been developed for medical imaging.

More recently, following the success of Transformer in natural language processing [Vaswani et al., 2017], visual transformers have demonstrated remarkable performance compared to state-of-the-art CNNs across a wide range of computer vision tasks, including image classification [Yuan et al., 2021, Dosovitskiy et al., 2020], object detection [Liu et al., 2021, Zhu et al., 2021], and semantic segmentation [Cai et al., 2023]. Despite the debate over the adoption of Transformers and CNNs in visual domains, regarding generalization ability [Liu et al., 2022b, Zhou et al., 2022b, Bao et al., 2021, Xiao et al., 2022, Touvron et al., 2021, Ding et al., 2022, Bai et al., 2021, Mao et al., 2022, Zhang et al., 2022, Zhou et al., 2022a], training data requirements [Dosovitskiy et al., 2020, Steiner et al., 2021, Tay et al., 2022], computational costs [Paul and Chen, 2022], visual transformers have shown the potential to achieve even better performance than CNNs in medical imaging analysis [Xiao et al., 2023, Chen et al., 2021a,b]. For example, the TransUNet [Chen et al., 2021b] follows the 2D UNet [Falk et al., 2019] design and incorporates the Transformer blocks in building the encoder and decoder, which achieves promising results in CT segmentation tasks. Given that most state-of-the-art CNNs use small-sized convolution kernels, such locality predisposes CNNs towards local spatial patterns. In contrast, the global self-attention mechanism in visual transformers greatly boosts their ability to model the long-range dependencies in medical images [Li et al., 2023].

Self-supervised contrastive learning [Chen et al., 2020c,b, Grill et al., 2020, Caron et al., 2020, Xiao et al., 2023] has demonstrated significant promise in dealing with the scarcity of high-quality annotations in the medical imaging domain [Zhou, 2021, Xiao et al., 2023, Chen et al., 2021a]. Contrastive learning approaches treat each image as a separate class, aiming to reduce the similarity between augmented views of distinct images while increasing the similarity between views of the same image during neural network pre-training. However, since radiography images are obtained following standard radiography imaging protocols [Xiang et al., 2021, Haghighi et al., 2022], different images show much higher similarity compared to photographic images [He et al., 2020, Chen et al., 2020c]. To avoid such problems in pre-training neural networks for medical imaging, recent works adopt restorative strategies [Alex et al., 2017, Chen et al., 2019, Zhou et al., 2019b, Zhu et al., 2020, Chen et al., 2020a, Xie et al., 2022, Xiao et al., 2023], which perform pixel-wise image reconstruction. For instance, [Xiao et al., 2023] adopts masked autoencoders (MAE) [He et al., 2022] to pre-train both CNNs and ViTs and achieve state-of-the-art performance in radiography image classification. In our work, we also adopt the MAE method in [Xiao et al., 2023] to pre-train our neural networks before learning low-rank features.

3 FORMULATION

3.1 PIPELINE FOR THORAX DISEASE CLASSIFICATION

We follow [Xiao et al., 2023] and use masked MAE [He et al., 2022] to pre-train CNNs or ViTs, and then perform our novel LRFL. The full training pipeline of learning low-rank features for disease classification can be described in three steps. In the first step, we pre-train the networks with the self-supervised restorative learning method masked MAE [He et al., 2022] on the pre-training dataset such as ImageNet-1k [Krizhevsky et al., 2012] and X-rays (0.5M) [Xiao et al.,

2023]. We randomly mask patches on input images and optimize the networks for pixel-wise image reconstruction on the masked patches. In the second step, we finetune the pre-trained networks with the cross-entropy loss for image classification on the target datasets such as NIH-ChestX-ray [Wang et al., 2017], COVIDx [Pavlova et al., 2022], and CheXpert [Irvin et al., 2019]. In the last step, we fix the backbones of the network and finetune the linear classifier with our novel LRFL method, as illustrated in Figure 1.

Figure 1: Training Pipeline for Thorax Disease Classification.

3.2 PROBLEM SETUP FOR LRFL

We now introduce the problem setup for LRFL with training details. Suppose the training data are given as $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ where $\mathbf{x}_i$ and $\mathbf{y}_i \in \mathbb{R}^C$ are the i-th training data point and its corresponding class label vector respectively, and C is the number of classes. Each element $\mathbf{y}_i$ is binary with $\mathbf{y}_i = 1$ indicating the i-th disease is present in $\mathbf{x}_i$, otherwise $\mathbf{y}_i = 0$. Given the feature $\mathbf{F} \in \mathbb{R}^{n\times d}$ of all the training data where d is the dimension of the feature and $\mathbf{F}$ is the feature produced by the neural network obtained by the regular finetuning as step two of the pipeline in Figure 1, we can train a linear neural network as a linear classifier by optimizing

$$\min_{\mathbf{W}^{(\mathrm{lin})}\in\mathbb{R}^{d\times C}} L(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(\mathbf{y}_i, \left[\sigma\left(\mathbf{F}\mathbf{W}^{(\mathrm{lin})}\right)\right]_i\right). \tag{1}$$

Here $[\sigma(\mathbf{A})]_{ic} = 1/(1+\exp(-\mathbf{A}_{ic}))$ is the element-wise sigmod function for $\mathbf{A} \in \mathbb{R}^{n\times C}$, $i \in [n], c \in [C]$. When σ is applied on a matrix $\mathbf{A}$, it returns a matrix of the same size as $\mathbf{A}$ and where the sigmod function is applied on every element of $\mathbf{A}$. $[\cdot]_i$ stands for the i-th row of a matrix. KL stands for the element-wise binary cross-entropy function. Given two nonnegative vectors $\mathbf{u} = [u_1, \ldots, u_d] \in \mathbb{R}^d, \mathbf{v} = [v_1, \ldots, v_d] \in \mathbb{R}^d$ where $u_i \in \{0, 1\}$ for all $i \in [d]$ and $\|\mathbf{v}\|_\infty \le 1$, $\mathrm{KL}(\mathbf{u}, \mathbf{v}) := \sum_{j=1}^d -u_i \log v_i - (1-u_i)\log(1-v_i)$. We use $\mathbf{Y} = [\mathbf{y}_1; \mathbf{y}_2; \ldots; \mathbf{y}_n] \in \mathbb{R}^{n\times C}$ to denote the training label matrix by stacking the label vectors of all the training data.

3.3 GENERALIZATION BOUND FOR LOW-RANK FEATURE LEARNING

We define the loss function $\ell(\mathrm{NN}(\mathbf{x}), \mathbf{y}) := \|\mathrm{NN}(\mathbf{x}) - \mathbf{y}\|_2^2$, and the generalization error of the network NN is the expected risk of the loss ℓ, which is denoted by $L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell(\mathrm{NN}_{\mathbf{W}}(\mathbf{x}), \mathbf{y})\right]$. The kernel gram matrix for the feature $\mathbf{F}$ is $\mathbf{K}_n = \frac{1}{n}\mathbf{F}\mathbf{F}^\top$. We let $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \ldots \ge \widehat{\lambda}_r > 0$ where $\bar{r} \le \min\{n, d\}$ is the rank of $\mathbf{K}_n$. Suppose the Singular Value Decomposition of $\mathbf{F}$ is $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$ has orthogonal columns, $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal elements being the singular values of $\mathbf{F}$, and $\mathbf{V} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. The columns of $\mathbf{U}$ and $\mathbf{V}$ are also called the left eigenvectors and the right eigenvectors of $\mathbf{F}$ respectively. Let $\sigma_1 \ge \sigma_2 \ldots \ge \sigma_d$ be the singular values of $\mathbf{F}$, and $\bar{\mathbf{Y}} = \mathbf{U}^{\bar{r}}\mathbf{U}^{\bar{r}\top}\mathbf{Y}$ be the projection of the training label matrix $\mathbf{Y}$ onto the subspace spanned by the top-$\bar{r}$ left eigenvectors of $\mathbf{F}$, where $\mathbf{U}^{\bar{r}} \in \mathbb{R}^{n\times\bar{r}}$ is formed by the top $\bar{r}$ eigenvectors in $\mathbf{U}$. Then we have the following theorem giving the sharp generalization error bound for the linear neural network in (1).

Theorem 3.1. For every $x > 0$, with probability at least $1 - \exp(-x)$, after the t-th iteration of gradient descent for all $t \ge 1$, we have

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \le \left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}} + c_1\left(1 - \eta\widehat{\lambda}_r\right)^{2t}\|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2 \min_{h\in[0,r]}\left(\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i=h+1}^{r}\widehat{\lambda}_i}\right) + \frac{c_3 x}{n}, \tag{2}$$

where c_1, c_2, c_3 are positive constants.

Remark 3.2. The RHS of (2) is the generalization error bound for the linear neural network used in LRFL as step three of the pipeline in Figure 1. Moreover, let $\sigma_1 \ge \sigma_2 \ldots \ge \sigma_d$ be the singular values of $\mathbf{F}$. Due to the fact that $\sqrt{\frac{1}{n}\sum_{i=h+1}^{r}\widehat{\lambda}_i} \le \frac{1}{n}\sum_{i=h+1}^{r}\sigma_i$, it follows by (2) that

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \le c_1\left(1 - \eta\widehat{\lambda}_r\right)^{2t}\|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2\left(\frac{h}{n} + \frac{1}{n}\sum_{i=T+1}^{d}\sigma_i\right) + \frac{c_3 x}{n}, \tag{3}$$

which holds for all $T \in [0, d]$. (3) motivates the reduction of the truncated nuclear norm of the feature $\mathbf{F}$, as detailed in the next subsection.

3.4 OPTIMIZATION OF THE TRUNCATED NUCLEAR NORM IN SGD

Using notations in Section 3.3, the truncated nuclear norm of $\mathbf{F}$ is $\|\mathbf{F}\|_T := \sum_{i=T+1}^{d}\sigma_i$ where $T \in [0, d]$. It can be observed by the generalization error bound (3) of Remark 3.2 that a smaller $\|\mathbf{F}\|_T$ renders a tighter upper bound for the generalization error of the linear neural network used for

LRFL. This observation gives a strong theoretical motivation for us to add the truncated nuclear norm $\|\mathbf{F}\|_T$ to the training loss (1). However, $\|\mathbf{F}\|_T$ is not separable, so the training loss with $\|\mathbf{F}\|_T$ cannot be directly optimized by the standard SGD. To address this problem, we propose an approximation $\overline{\|\mathbf{K}\|_T}$ to $\|\mathbf{K}\|_T$ which is separable so that $\overline{\|\mathbf{K}\|_T}$ can be optimized by standard SGD.

First, we note that if $\mathbf{U}, \mathbf{V}$ are known, then $\mathbf{\Sigma} = \mathbf{U}^\top \mathbf{F} \mathbf{V}$. If we have an approximation $\overline{\mathbf{U}}$ to $\mathbf{U}$ and an approximation $\overline{\mathbf{V}}$ to $\mathbf{V}$, then $\mathbf{\Sigma}$ can be approximated by

$$\overline{\mathbf{\Sigma}} = \overline{\mathbf{U}}^\top \mathbf{F} \overline{\mathbf{V}}.$$

As a result, the approximation $\overline{\|\mathbf{K}\|_T}$ to the truncated nuclear norm is

$$\begin{aligned}\overline{\|\mathbf{K}\|_T} &= \sum_{s=T+1}^{d} \overline{\mathbf{\Sigma}}_{ss} = \sum_{s=T+1}^{d} \left[\overline{\mathbf{U}}^\top \mathbf{F} \overline{\mathbf{V}}\right]_{ss} \\ &= \sum_{i=1}^{n} \left(\sum_{s=T+1}^{d} \sum_{k=1}^{d} \overline{\mathbf{U}}_{si}^\top \mathbf{F}_{ik} \overline{\mathbf{V}}_{ks} \right). \end{aligned} \tag{4}$$

Due to the above discussions, the loss function of LRFL with the approximate truncated nuclear norm $\overline{\|\mathbf{K}\|_T}$ is

$$\begin{aligned}\mathcal{L}(\mathbf{W}) = &\frac{1}{m} \sum_{v_i \in \mathcal{V}_\mathcal{L}} \mathrm{KL}\left(\mathbf{y}_i, \left[\sigma\left(\mathbf{F}\mathbf{W}^{(\mathrm{lin})}\right)\right]_i\right) \\ &+ \eta \sum_{i=1}^{n} \left(\sum_{s=T+1}^{d} \sum_{k=1}^{d} \overline{\mathbf{U}}_{si}^\top \mathbf{F}_{ik} \overline{\mathbf{V}}_{ks} \right), \end{aligned} \tag{5}$$

where $\eta > 0$ is the weighting parameter for the truncated nuclear norm. Because (5) is to be optimized by the standard SGD, we have the loss function of LRFL for the j-th minibatch $\mathcal{B}_j \subseteq [n]$ as follows:

$$\begin{aligned}\mathcal{L}_j(\mathbf{W}) = &\frac{1}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} \mathrm{KL}\left(\mathbf{y}_i, \left[\sigma\left(\mathbf{F}\mathbf{W}^{(\mathrm{lin})}\right)\right]_i\right) \\ &+ \frac{\eta}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} \left(\sum_{s=T+1}^{d} \sum_{k=1}^{d} \overline{\mathbf{U}}_{si}^\top \mathbf{F}_{ik} \overline{\mathbf{V}}_{ks} \right). \end{aligned} \tag{6}$$

The approximation $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ can be computed as the left and right eigenvectors of the feature $\mathbf{F}$ computed at earlier epochs. In order to save computation and avoiding performing SVD for $\mathbf{F}$ at every epoch, we propose to update $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ only after certain epochs. Algorithm 1 describes the training algorithm for our LRFL that uses standard SGD to optimize the loss function (5). Before the first epoch, we compute $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ as the left and right eigenvectors of the feature $\mathbf{F}$ at the initialization of the neural network. After every t_0 epochs for t_0 being a constant integer, we update $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ as the left and right eigenvectors of the feature $\mathbf{F}$ of the neural network righter after t_0-th epoch. The algorithm is described in Algorithm 1.

Algorithm 1 Training Algorithm with the Approximate Truncated Nuclear Norm by SGD

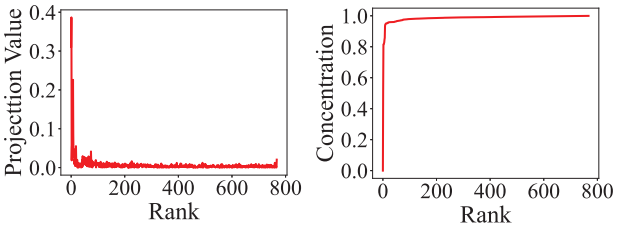
Initialize the weights of the network by $\mathcal{W} = \mathcal{W}(0)$ through random initialization
Compute feature $\mathbf{F}$ by the neural network, and its SVD as $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$
Update $\overline{\mathbf{U}} = \mathbf{U}$, $\overline{\mathbf{V}} = \mathbf{V}$
for $t = 1, 2, \ldots, t_{\max}$ **do**
  if $t \equiv 0 \pmod{t_0}$ **then**
    Compute feature $\mathbf{F}$ of the neural network, and its SVD $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$.
    Update $\overline{\mathbf{U}} = \mathbf{U}$, $\overline{\mathbf{V}} = \mathbf{V}$
  end if
  for $b = 1, 2, \ldots, B$ **do**
    Update $\mathcal{W}$ by applying gradient descent on batch $\mathcal{B}_j$ using the gradient of the loss $\mathcal{L}_j$ in Eq.(6)
  end for
end for
return The trained weights $\mathcal{W}$ of the network

4 EXPERIMENTAL RESULTS

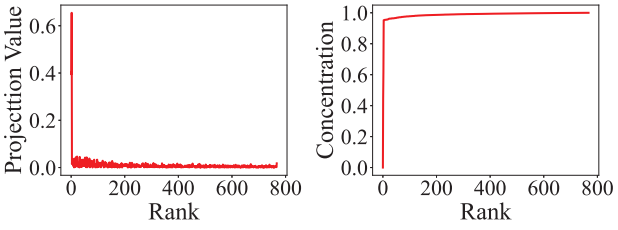
In this section, we conduct experiments on medical datasets to show the effectiveness of the proposed LRFL. The experiments section is organized as follows. In Section 4.1, we discuss our experimental setup and implementation details. In Section 4.2, Section 4.3 and Section 4.4, we compare LRFL against various architectures on three medical datasets. In Section 4.5, we investigate the performance of LRFL on small data regimes. In Section 4.6, we empirically show the effectiveness of our loss function and investigate the significance of the low-rank regularization.

4.1 IMPLEMENTATION DETAILS

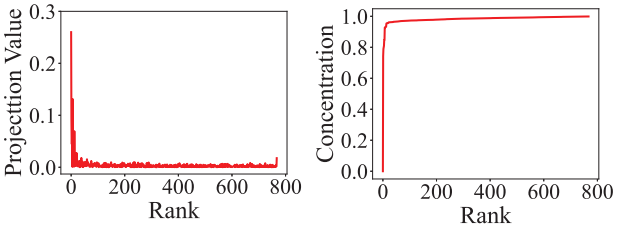
In this section, we evaluate the proposed LRFL for thorax disease classification. We utilize networks pre-trained on ImageNet [Russakovsky et al., 2015] or chest X-rays in [Xiao et al., 2023] with MAE, which adopts the self-supervised learning strategy by reconstructing missing pixels from patches of input images. We fine-tune the pre-trained networks with low-rank regularization for classification on three public X-ray datasets, namely (1) NIH-Chest Xray 14 [Wang et al., 2017], (2) Stanford CheXpert [Irvin et al., 2019] and (3) COVIDx [Pavlova et al., 2022]. We use ADAM optimizer in the fine-tuning process. The batch size is set to 2048 for all datasets. We first fine-tune the entire networks for 75 epochs with Adam following the settings in [Xiao et al., 2023]. Next, we fine-tune the network with low-rank regularization for another 75 epochs. The learning rate is initialized as 2.5×10^{-5} and annealed down to 1×10^{-7} following a cosine schedule. The default values of momentum and weight decay are set to 0.9 and 0. We use standard data augmentation techniques for both datasets, including random-resize cropping, random rotation, and ran-



(a) NiH-ChestXray-14



(b) COVIDx



(c) CheXpert

Figure 2: Eigen-projection (first column) and signal concentration ratio (second column) of Vit-Base/16 on NiH-ChestXray-14, COVIDx, and CheXpert. To compute the eigen-projection, we first calculate the eigenvectors $\mathbf{U}$ of the kernel gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ computed by a feature matrix $\mathbf{F} \in \mathbb{R}^{n\times d}$, then the projection value is computed by $\mathbf{p} = \frac{1}{C}\sum_{c=1}^{C} \left\|\mathbf{U}^\top \mathbf{Y}^{(c)}\right\|_2^2 / \left\|\mathbf{Y}^{(c)}\right\|_2^2 \in \mathbb{R}^n$, where C is the number of classes, and $\mathbf{Y} \in \{0,1\}^{n\times C}$ is the one-hot labels of all the training data, $\mathbf{Y}^{(c)}$ is the c-th column of $\mathbf{Y}$. The eigen-projection $\mathbf{p}_r$ for $r \in [\min(n,d)]$ reflects the amount of the signal projected onto the r-th eigenvector of $\mathbf{K}$, and the signal concentration ratio of a rank r reflects the proportion of signal projected onto the top r eigenvectors of $\mathbf{K}$. The signal concentration ratio for rank r is computed by $\left\|\mathbf{p}^{(1:r)}\right\|_2$, where $\mathbf{p}^{(1:r)}$ contains the first r elements of $\mathbf{p}$. For example, by the rank $r = 38$, the signal concentration ratio of $\mathbf{Y}$ on NIH ChestX-ray14, COVIDx, and CheXpert are 0.959, 0.964, and 0.962 respectively.

dom horizontal flipping. Throughout the paper, we evaluate our LRFL method on both CNN and visual transformer architectures including ResNet-50, DenseNet, ViT-S, and ViT-B. We refer to our model by 'X-LR', where X is the base model. For example, a ResNet-50 model trained using low-rank features is referred to as ResNet-50-LR.

Tuning the T and η by Cross-Validation. We tune the optimal values of feature rank T and weighting parameter for the truncated nuclear norm η on each dataset. Let $T = \lceil \gamma \min(n,d) \rceil$, where γ is the rank ratio. We select the values of γ and η by performing 5-fold cross-validation on 20% of the training data in each dataset. The value of γ is selected from $\{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2\}$. The value of η is selected from $\{5\times10^{-4}, 1\times10^{-3}, 2.5\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}\}$. The optimal values of η and γ selected by cross-validation on each dataset are shown in Table 1.

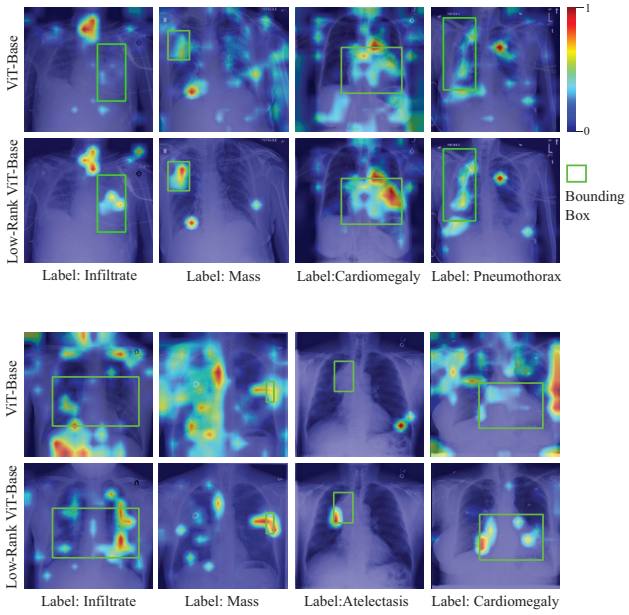


Figure 3: Grad-CAM visualization results on NIH ChestX-ray 14. The figures in the first row are the visualization results of ViT-Base, and the figures in the second row are the visualization results of Low-Rank ViT-Base. Ground-truth bounding box for each disease is shown in green. Although both the base model and its corresponding low-rank model predict the correct disease label, the low-rank model pays more attention to the disease location than the base model. More Grad-CAM visualization results are deferred to Figure 6 of the supplementary.

Parameters	NIH-ChestX-ray	COVIDx	CheXpert
γ	0.05	0.003	0.05
η	5×10^{-4}	1×10^{-3}	1×10^{-3}

Table 1: Selected rank ratio γ and weighting parameter for the truncated nuclear norm η on each dataset.

4.2 NIH CHESTX-RAY14

Experimental setup. NIH-ChestX-ray14 [Wang et al., 2017] consists of 112,120 X-rays collected from 30,805 unique patients. Each X-ray has up to 14 associated labels,

Pre-training Dataset	Model	Rank	mAUC
Imagenet-1k	ResNet-50	-	81.78
	ResNet-50-LR	0.05r	82.18
Imagenet-1k	DenseNet-121	-	82.02
	DenseNet-121-LR	0.05r	82.35
X-rays(0.3M)	ViT-S [Xiao et al., 2023]	-	82.30
	ViT-S-LR	0.05r	82.70
X-rays(0.5M)	ViT-B [Xiao et al., 2023]	-	<u>83.00</u>
	ViT-B-LR	0.05r	**83.40**

Table 2: The table compares the performance of various models and their low-rank counterparts on NIH ChestX-ray14 dataset. The best result is presented in bold, and the second-best result is underlined. This convention applies to all the tables in our paper.

with the possibility of multiple labels per image. Following the official data split in [Wang et al., 2017], we use $75,312$ images for training and $25,596$ images for testing. Raw images from the dataset are in the size of 1024×1024. In our experiments, we scale down the input images to the size of 224×224. We report the mean AUC (Area Under the Curve) for 14 distinct classes and conduct a comprehensive comparison with 18 widely recognized and influential baseline methods.

Method	Architecture	Pre-training	mAUC
Wang et al. [Wang et al., 2017]	RN50		74.5
Li et al.[Li et al., 2018]	RN50		75.5
Yao et al. [Yao et al., 2018]	RN&DN		76.1
Wang et al.[Wang et al., 2019]	R152		78.8
Ma et al.[Ma et al., 2019]	R101		79.4
Tang et al.[Tang et al., 2018]	RN50		80.3
Baltruschat et al.[Baltruschat et al., 2019]	RN50		80.6
Guendel et al.[Guendel et al., 2018]	DN121		80.7
Guan et al.[Guan and Huang, 2018]	DN121	ImageNet-1K	81.6
Seyyed et al.[Seyyed-Kalantari et al., 2020]	DN121		81.2
Ma et al.[Ma et al., 2020]	DN121(×2)		81.7
Hermoza et al.[Hermoza et al., 2020]	DN121		82.1
Kim et al.[Kim et al., 2021]	DN121		82.2
Haghighi et al.[Haghighi et al., 2022]	DN121		81.7
Liu et al.[Liu et al., 2022a]	DN121		81.8
Taslimi et al.[Taslimi et al., 2022]	SwinT		81.0
MoCo v2 [Xiao et al., 2023]	DN121		80.6
MAE [Xiao et al., 2023]	DN121	X-rays (0.3M)	81.2
MAE [Xiao et al., 2023]	ViT-S/16		82.3
MAE [Xiao et al., 2023]	ViT-B/16	X-rays (0.5M)	<u>83.0</u>
RN-50-LR (Ours)	RN50	ImageNet-1K	82.2
DN-121-LR (Ours)	DN121		82.4
ViT-S-LR (Ours)	ViT-S/16	X-rays (0.3M)	82.7
ViT-B-LR (Ours)	ViT-B/16	X-rays (0.5M)	**83.4**

Table 3: The table shows the performance of various state-of-the-art (SOTA) CNN-based and Transformer-based methods on ChestX-ray14. With the same pre-training settings as ViT-B in [Xiao et al., 2023], our ViT-B-LR achieves the new record high of 83.4 mAUC. RN, DN, and SwinT represent ResNet, DenseNet, and Swin Transformer.

Results and analysis. Table 2 presents the performance comparisons between several top-performing baseline models and their corresponding low-rank models on the NIH ChestX-ray14 dataset. Throughout this section we use postfix "-LR" to indicate a neural network trained with our LRFL. For example, we use ViT-B model pre-trained on $266,340$ chest X-rays with Masked Autoencoders (MAE) [Xiao et al., 2023]. The pre-trained ViT-B network is fine-tuned on the NIH ChestX-ray14 dataset and achieves a mean AUC of 83.0. Next, we fine-tune ViT-B with low-rank regularization for another 75 epochs. The low-rank model, denoted as ViT-B-LR, achieves a mean AUC of 83.4. It is observed that all low-rank models achieve improvement in mean AUC compared to the corresponding base models. ViT-S-LR improves its base model by a mean AUC of 0.4%. Similar improvements are observed for CNN-based models as well. For example, ResNet-50-LR improves its base model by a mean AUC of 0.40%.

Table 3 shows the performance comparisons of models trained with LRFL against state-of-the-art CNN and Transformer models on NIH ChestX-ray14. In our experiments, ViT-B-LR achieves the new state-of-the-art performance with a mean AUC of 83.4%. It is important to highlight that the research community dedicated four years to enhancing the AUC score for CNN-type architectures, advancing it from 74.5 to 82.2. This improvement was primarily attributed to the challenging nature of the training process.

To study how LRFL improves the performance of base models in disease detection, we use the Grad-CAM [Selvaraju et al., 2017] to visualize the parts in the input images that are responsible for the predictions of the base models and low-rank models. Examples of visualization results in Figure 3 show that our LRFL models usually focus more on the areas inside the bounding box associated with the labeled disease. In contrast, the base models also focus on the areas outside the bounding box or even areas in the background. More Grad-CAM visualization results are deferred to Figure 6 of the supplementary.

4.3 COVIDX

Experimental setup. COVIDx (Version 9A) [Pavlova et al., 2022] consists of 30,386 chest X-rays collected from 17,026 unique patients. We follow the previous works[Pavlova et al., 2022, Xiao et al., 2023] in splitting the dataset into 29,986 training images with four different classes and 400 testing images with three classes. For fair comparisons with the previous methods, we report Top-1 accuracy on the test set (3 classes).

Results and analysis. Table 4 compares the performance of SOTA transformer-based models and the LRFL models on the COVIDx dataset. Similar to Section 4.2, the base ViTs are first pre-trained on 266,340 chest X-rays using Masked Autoencoders (MAE) and next the pre-trained model is fine-tuned on COVIDx dataset. It can be observed by Table 4 that both ViT-S-LR and ViT-B-LR outperform their corresponding base models ViT-S and ViT-B, achieving an increase in accuracy of 1.6% and 1.7%, respectively. Table 4 also compares the performance of our LRFL models

Method	Architecture	Rank	Accuracy	Covid-19 Senstitvity
COVIDNet-CXR Small [Wang et al., 2020]	-	-	92.6	87.1
COVIDNet-CXR Large [Wang et al., 2020]	-	-	94.4	96.8
DN121 (MoCo v2) [Xiao et al., 2023]	DN121	-	94.0	94.5
DN121 [Xiao et al., 2023]	DN121	-	93.5	97.0
ViT-S [Xiao et al., 2023]	ViT-S/16	-	95.2	94.5
ViT-S-LR (Ours)	ViT-S/16	0.01r	96.8	97.5
ViT-B [Xiao et al., 2023]	ViT-B/16	-	95.3	95.5
ViT-B-LR (Ours)	ViT-B/16	0.003r	**97.0**	**98.5**

Table 4: The table shows the performance of various state-of-the-art (SOTA) CNN-based and Transformer- based methods on COVIDx. With the same pre-training settings as ViT-S and ViT-B in [Xiao et al., 2023], our ViT-S-LR and ViT-B-LR achieve 97.0 % Accuracy. DN represents DenseNet.

Method	Architecture	Rank	Atelectasis	Cardiomegaly	mAUC (%)
Allaouzi et al.[Allaouzi and Ahmed, 2019]	DN121	-	72.0	88.0	82.8
Irvin et al.[Irvin et al., 2019]		-	81.8	82.8	88.9
Seyyedkalantari et al.[Seyyed-Kalantari et al., 2020]		-	81.2	83.0	87.3
Pham et al.[Pham et al., 2021]		-	82.5	85.5	89.4
Hosseinzadeh et al.[Hosseinzadeh Taher et al., 2021]		-	-	-	87.1
Haghighi et al.[Haghighi et al., 2022]		-	-	-	87.6
Kang et al.[Kang et al., 2021]		-	82.1	85.9	89.0
DN121 (MoCo v2) [Xiao et al., 2023]		-	78.5	77.9	88.7
DN121 [Xiao et al., 2023]		-	81.5	77.6	88.7
ViT-S [Xiao et al., 2023]	ViT-S/16	-	83.5	81.8	89.2
ViT-S-LR (Ours)	ViT-S/16	0.05r	**86.3**	93.7	89.6
ViT-B [Xiao et al., 2023]	ViT-B/16	-	82.7	83.5	89.3
ViT-B-LR (Ours)	ViT-B/16	0.05r	85.4	**94.6**	**89.8**

Table 5: The table shows the performance of various state-of-the-art (SOTA) CNN-based and Transformer- based methods on CheXpert.

against the state-of-the-art models on the COVIDx dataset. LRFL models achieve much higher accuracy as compared to CNN-based models such as DenseNet-121. ViT-B-LR achieves the new SOTA performance of 97% top-1 accuracy with input resolution set to 224×224, which exceeds the previous SOTA performance [Xiao et al., 2023] by 1.7 % in top-1 accuracy.

4.4 STANFORD CHEXPERT

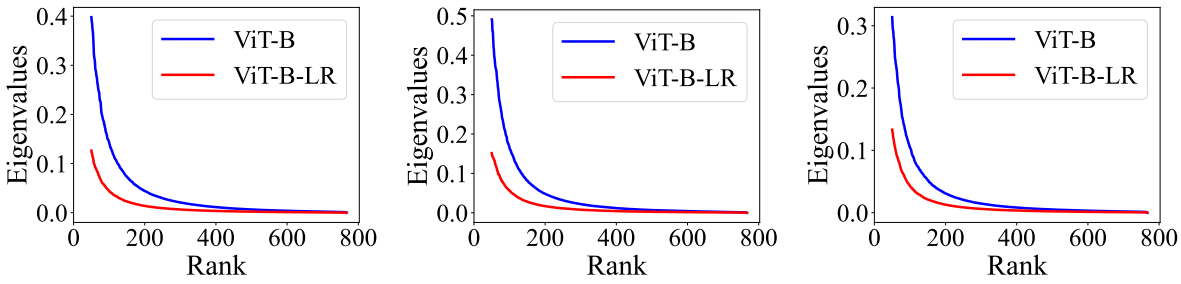
Experimental setup. CheXpert [Irvin et al., 2019] consists of 224,316 chest X-rays collected from 65,240 patients, where 191,028 chest X-rays are used for training. Each X-ray in the dataset has radiology reports indicating the presence of 14 diseases. Following the protocol in [Xiao et al., 2023], all images are resized into 224 × 224. We also report the mean AUC (Area Under the Curve) for the 5 distinct classes and conduct a comprehensive comparison with 9 widely recognized and influential baseline methods.

Results and analysis. Table 5 presents the performance comparisons between the baseline models and the LRFL models on the CheXpert dataset. It is observed that ViT-B-LR achieves state-of-the-art performance of 89.8% in mAUC, and improves the performance of ViT-B by 0.5% in mAUC. ViT-S-LR also improves the performance of ViT-S by 0.4% in mAUC, which demonstrates the power of LRFL. We also show the classification accuracy on Atelectasis and Cardiomegaly in Table 5, where our method exhibits much better performance than baseline methods. For example, ViT-S-LR achieves an mAUC of 86.3% on Atelectasis, with a 2.8% improvement over ViT-S trained with MAE. Such improvements demonstrate the power of LRFL in detecting distinct diseases.

4.5 EXPERIMENTS IN SMALL DATA REGIMES

Experimental setup. We explore the effectiveness of low-rank features learned in scenarios with limited data availability, which is particularly significant given the challenges in acquiring high-quality data annotations in the medical imaging domain. We expect that LRFL models can demonstrate improved performance in such situations due to our theoretical guarantee of the better generalization capability of LRFL. We randomly select 5%, 10%, 15%, 20%, 25%, and 50% of training data from the NIH ChestX-ray14 dataset and then fine-tune the base model using its default training configurations. We then train LRFL models for 20 epochs.

Results and analysis. As depicted in Table 7, our LRFL models consistently outperform their corresponding base methods across all data subsets, including 5%, 10%, 15%,



(a) ChestX-ray14 (b) COVIDx (c) CheXpert

Figure 4: Eigenvalues comparison between ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert.

20%, 25%, and 50% on the NIH ChestX-ray14 dataset. Notably, the average improvement in performance is more substantial for the 5% data subset compared to the remaining subsets. For instance, ViT-B-LR exhibits a remarkable improvement of 1.05% for the 5% data subset, which significantly surpasses the improvements of 0.15%, 0.06%, 0.06%, 0.09% and 0.11% observed for the 10%, 15%, 20%, 25% and 50% training data subsets, respectively. These findings are consistent with our expectations, showcasing the strong generalization capability of LRFL models in mitigating overfitting issues with limited data. In conclusion, our findings in the low-data regimes demonstrate the superiority of our LRFL in delivering more generalizable and robust representations for tasks with limited data availability, thereby contributing to the reduction of annotation costs. Models of this nature hold substantial value across various medical tasks, particularly because annotating medical images tends to be challenging and necessitates specialized expertise.

4.6 ABLATION STUDY

4.6.1 Study on the Kernel Eigenvalues and Kernel Complexity

In this section, we compare the kernel eigenvalues and kernel complexity of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert. The eigenvalues of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert are shown in Figure 4. The kernel complexity of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert are shown in Table 6.

Method	ChestX-ray14		COVIDx		CheXpert	
	Complexity	h	Complexity	h	Complexity	h
ViT-B	0.0101	465	0.0207	303	0.0040	766
ViT-B-LR	0.0076	262	0.0155	187	0.0038	389

Table 6: Kernel complexity comparison between ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert.

4.6.2 Study on the Rank in Low-rank Feature Learning

In this section, we conduct an ablation study to investigate the impact of the rank T on the performance of our LRFL model on the NIH-ChestX-ray14 dataset for the ViT-B/16 model. In Figure 5, we present the performance change with respect to different values for the rank T, with T varying from $0.01r$ to $0.3r$, where $r = \min\{n, d\}$. It can be observed from this figure that our LRFL model often delivers aAUC higher than 83% for most values for T.

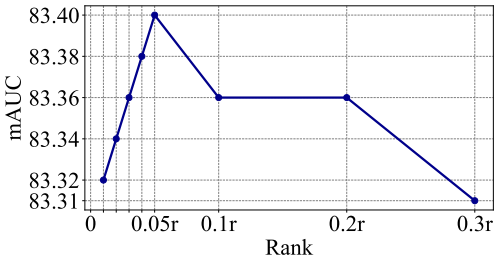


Figure 5: The relationship between mAUC and rank T of ViT-B-LR on ChestX-ray14.

4.6.3 Exploring Fine-tuning Strategies

Our LRFL method learns low-rank features by leveraging models pre-trained on the target dataset. In this section, we conduct an ablation study to investigate the significance of low-rank regularization in the fine-tuning process. In Table 8, we present a comparative analysis of low-rank regularization compared with several performance-enhancing techniques, including mix-up [Zhang et al., 2018], label smoothing [Müller et al., 2019], and EMA [Wightman, 2019]. We also perform an experiment by fine-tuning without low-rank regularization and other tricks, which serves as a baseline for studying the effects of fine-tuning strategies. All models undergo equivalent training epochs to ensure a fair comparison. It can be observed that the LRFL models achieve

Pre-training Dataset	Model	Label Fractions											
		5%		10%		15%		20%		25%		50%	
		Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC
X-rays(0.3M)	ViT-S	-	61.22	-	73.19	-	76.99	-	78.65	-	79.57	-	81.20
	ViT-S-LR(Ours)	$0.05r$	61.81	$0.2r$	73.84	$0.04r$	77.21	$0.04r$	78.86	$0.05r$	79.65	$0.05r$	81.35
X-rays(0.5M)	ViT-B	-	70.71	-	78.67	-	79.99	-	80.59	-	81.13	-	82.19
	ViT-B-LR (Ours)	$0.05r$	71.76	$0.2r$	78.82	$0.2r$	80.05	0.1r	80.65	$0.05r$	81.22	$0.05r$	82.30

Table 7: The table evaluates the performance of various models under low data regimes on the NIH ChestX-rays14 dataset. Models trained with low-rank features effectively combat overfitting in scenarios with limited data availability, thereby enhancing the quality of representations for downstream tasks.

Model	mAUC					
	Base Model	Fine-tuning	Mix-up [Zhang et al., 2018]	Label Smoothing [Müller et al., 2019]	EMA [Wightman, 2019]	LRFL
ViT-S	82.27	82.26	82.09	82.24	82.26	82.70
ViT-B	83.00	83.00	82.37	82.99	82.98	**83.40**

Table 8: Comparison of fine-tuning strategies on NIH ChestX-ray14.

the highest performance improvement compared to all other approaches. Unlike natural images, results in Table 8 show that applying mix-up, label smoothing, or EMA to the NIH ChestX-ray dataset lead to performance drops. In addition, it is observed that fine-tuning models pre-trained on the target dataset without low-rank regularization does not lead to performance improvements compared to fine-tuning with low-rank regularization. For example, the original ViT-S [Xiao et al., 2023] achieves a Mean AUC of 82.27% on NIH Chest Xray-14. Fine-tuning this model for 20 epochs without low-rank regularization leads to a mean AUC of 82.26%. In contrast, fine-tuning with low-rank regularization for 75 epochs leads to a mean AUC of 83.40%. We observe similar results for all models based on low-rank features, which demonstrates the significance of LRFL.

5 CONCLUSION

In this paper, we propose a novel Low-Rank Feature Learning (LRFL) method for thorax disease classification, which can effectively reduce the adverse effect of noise and background, or non-disease areas, on the radiographic images for disease classification. Being universally applicable to the training of all neural networks, LRFL is both empirically motivated by the low frequency property and theoretically motivated by our sharp generalization bound for neural networks with low-rank features. Extensive experimental results on thorax disease datasets, including NIH-ChestX-ray, COVIDx, and CheXpert, demonstrate the superior performance of LRFL in terms of mAUC and classification accuracy.

References

Varghese Alex, Kiran Vaidhya, Subramaniam Thirunavukkarasu, Chandrasekharan Kesavadas, and Ganapathy Krishnamurthi. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. *Journal of Medical Imaging*, 4(4): 041311, 2017.

Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Tej Bahadur Chandra and Kesari Verma. Analysis of quantum noise-reducing filters on chest x-ray images: A review. *Measurement*, 153:107426, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021a.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021b.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58: 101539, 2019.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels. *Advances in Neural Information Processing Systems*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Hejie Cui, Xinglong Liu, and Ning Huang. Pulmonary vessel segmentation based on orthogonal fused u-net++ of chest ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–300. Springer, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

Qiaoqiao Ding, Yong Long, Xiaoqun Zhang, and Jeffrey A Fessler. Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct. *arXiv preprint arXiv:1801.09533*, 2018.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, page 1, 2018.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

Ruibin Feng, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Parts2whole: Self-supervised contrastive learning via reconstruction. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 85–95. Springer, 2020.

Riqiang Gao, Yuankai Huo, Shunxing Bao, Yucheng Tang, Sanja L Antic, Emily S Epstein, Aneri B Balar, Steve Deppen, Alexis B Paulson, Kim L Sandler, et al. Distanced lstm: time-distanced gates in long short-term memory models for lung cancer detection. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*, pages 310–318. Springer, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018.

Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning

to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican Congress on Pattern Recognition*, pages 757–765. Springer, 2018.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Renato Hermoza, Gabriel Maicas, Jacinto C Nascimento, and Gustavo Carneiro. Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–549. Springer, 2020.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13. Springer, 2021.

Mingzhe Hu, Jiahan Zhang, Luke Matkovic, Tian Liu, and Xiaofeng Yang. Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions. *Journal of Applied Clinical Medical Physics*, 24(2):e13898, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

Mintong Kang, Yongyi Lu, Alan L Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. *In Submission: Thirty-Sixth Conference on Neural Information Processing Systems*, 2021. URL https://arxiv.org/pdf/2109.12265.pdf.

Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15719–15728, June 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Sangyoon Lee, Min Seok Lee, and Moon Gi Kang. Poisson–gaussian noise analysis and estimation for low-dose x-ray images in the nsct domain. *Sensors*, 18(4):1019, 2018.

Jun Li, Junyu Chen, Yucheng Tang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *arXiv preprint arXiv:2206.01136*, 2022.

Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, page 102762, 2023.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8290–8299, 2018.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017b.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20706, 2022a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022b.

Congbo Ma, Hu Wang, and Steven C. H. Hoi. Multi-label thoracic disease image classification with cross-attention networks, 2020.

DongAo Ma, Mohammad Reza Hosseinzadeh Taher, Jiaxuan Pang, Nahid UI Islam, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Benchmarking and boosting transformers for medical image classification. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 12–22. Springer, 2022.

Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1378–1382. IEEE, 2019.

EN Manson, V Atuwo Ampoh, E Fiagbedzi, JH Amuasi, JJ Flether, and C Schandorf. Image noise in radiography and tomography: Causes, effects and reduction techniques. *Curr. Trends Clin. Med. Imaging*, 2(5):555620, 2019.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

Maya Pavlova, Tia Tuinstra, Hossein Aboutalebi, Andy Zhao, Hayden Gunraj, and Alexander Wong. Covidx cxr-3: a large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2206.03671*, 2022.

Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 232–243. World Scientific, 2020.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

Yan Shen and Mingchen Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In *International Workshop on Machine Learning in Medical Imaging*, pages 389–397. Springer, 2018.

K Kirk Shung, Michael Smith, and Benjamin MW Tsui. *Principles of medical imaging*. Academic Press, 2012.

JH Siewerdsen, LE Antonuk, Y El-Mohri, J Yorkston, W Huang, JM Boudry, and IA Cunningham. Empirical and theoretical investigation of the noise performance of indirect detection, active matrix flat-panel imagers (amfpis) for diagnostic radiology. *Medical physics*.

JH Siewerdsen, LE Antonuk, Y El-Mohri, J Yorkston, W Huang, and IA Cunningham. Signal, noise power spectrum, and detective quantum efficiency of indirect-detection flat-panel imagers for diagnostic radiology. *Medical physics*, 25(5):614–628, 1998.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical

image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Xavier Tomas-Fernandez, Sila Kurugol, and Simon K Warfield. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, 38(11):2642–2653, 2019.

Perry Sprawls. Physical principles of medical imaging. 1993.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022.

Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018.

Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

Bram Van Ginneken, BM Ter Haar Romeny, and Max A Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on medical imaging*, 20(12):1228–1241, 2001.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. Thoraxnet: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE journal of biomedical and health informatics*, 24(2):475–485, 2019.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76550-z. URL https://doi.org/10.1038/s41598-020-76550-z.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Tiange Xiang, Yongyi Liu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. In-painting radiography images for unsupervised anomaly detection. *arXiv preprint arXiv:2111.13495*, 2021.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Transforming radiograph imaging with transformers: Comparing vision transformers with convolutional neural networks in multi-label thorax disease classification. In *Radiological Society of North America (RSNA)*, 2022.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, June 2022.

Lanyu Xu, Simeng Zhu, and Ning Wen. Deep reinforcement learning and its applications in medical imaging and radiation therapy: a survey. *Physics in Medicine & Biology*, 67(22):22TR02, 2022.

Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11: 638182, 2021.

Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

Yuan Yao, Fengze Liu, Zongwei Zhou, Yan Wang, Wei Shen, Alan Yuille, and Yongyi Lu. Unsupervised domain adaptation through shape modeling for medical image segmentation. In *Medical Imaging with Deep Learning*, 2021.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7286, 2022.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022a.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022b.

S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019a.

S Kevin Zhou, Hoang Ngan Le, Khoa Luu, Hien V Nguyen, and Nicholas Ayache. Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis*, 73:102193, 2021a.

Xiang Sean Zhou, Yiqiang Zhan, Vikas C Raykar, Gerardo Hermosillo, Luca Bogoni, and Zhipang Peng. Mining anatomical, physiological and pathological information from medical images. *ACM SIGKDD Explorations Newsletter*, 14(1):25–34, 2012.

Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019b.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021b.

Zongwei Zhou, Michael Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health: The Role of AI*. Springer, 2022c.

Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021.

A PROOFS

Proof of Theorem 3.1. It can be verified that at the t-th iteration of gradient descent for $t \geq 1$, we have

$$\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \frac{\eta}{n}\mathbf{F}^\top \left(\mathbf{F}\mathbf{W}^{(t-1)} - \mathbf{Y}\right). \tag{7}$$

It follows by (7) that

$$\begin{aligned}\mathbf{F}\mathbf{W}^{(t)} &= \mathbf{F}\mathbf{W}^{(t-1)} - \eta\mathbf{K}_n\left(\mathbf{F}\mathbf{W}^{(t-1)} - \mathbf{Y}\right)\\ &= \mathbf{F}\mathbf{W}^{(t-1)} - \eta\mathbf{K}_n\left(\mathbf{F}\mathbf{W}^{(t-1)} - \bar{\mathbf{Y}}\right),\end{aligned} \tag{8}$$

where $\mathbf{K} = \mathbf{F}\mathbf{F}^\top$, $\bar{\mathbf{Y}} = \mathbf{U}^{\bar{r}}{\mathbf{U}^{\bar{r}}}^\top\mathbf{Y}$.

We define $\mathbf{F}(\mathbf{W}, t) \coloneqq \mathbf{F}\mathbf{W}^{(t)}$, then it follows by (8) that

$$\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}} = (\mathbf{I}_n - \eta\mathbf{K}_n)\left(\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}}\right),$$

which indicates that

$$\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}} = (\mathbf{I}_n - \eta\mathbf{K}_n)^t\left(\mathbf{F}(\mathbf{W}, 0) - \bar{\mathbf{Y}}\right) = -(\mathbf{I}_n - \eta\mathbf{K}_n)^t\bar{\mathbf{Y}},$$

and

$$\|\mathbf{F}(\mathbf{W}, t) - \mathbf{Y}\|_{\mathrm{F}} \leq \left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}} + \left(1 - \eta\widehat{\lambda}_r\right)^t\|\mathbf{Y}\|_{\mathrm{F}} \leq \left(1 - \eta\widehat{\lambda}_r\right)^t\left\|\bar{\mathbf{Y}}\right\|_{\mathrm{F}} \leq \left(1 - \eta\widehat{\lambda}_r\right)^t\|\mathbf{Y}\|_{\mathrm{F}}. \tag{9}$$

As a result of (9), by using the proof of [Bartlett et al., 2005, Theorem 3.3], for every $x > 0$, with probability at least $1 - \exp(-x)$,

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \leq \left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}} + c_1\left(1 - \eta\widehat{\lambda}_r\right)^{2t}\|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2\min_{h\in[0,r]}\left(\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i=h+1}^{r}\widehat{\lambda}_i}\right) + \frac{c_3 x}{n}. \tag{10}$$

□

B MORE EXPERIMENT RESULTS

B.1 ADDITIONAL GRAD-CAM VISUALIZATION RESULTS

In this section, we show more grad-cam visualization results. We visualize the parts in the input images that are responsible for the predictions of the ground-truth disease label for base models and low-rank models. Examples of visualization results in Figure 6 show that our low-rank models usually focus more on the areas inside the bounding box associated with the labeled disease. In contrast, the base models also focus on the areas outside the bounding box or even areas in the background.

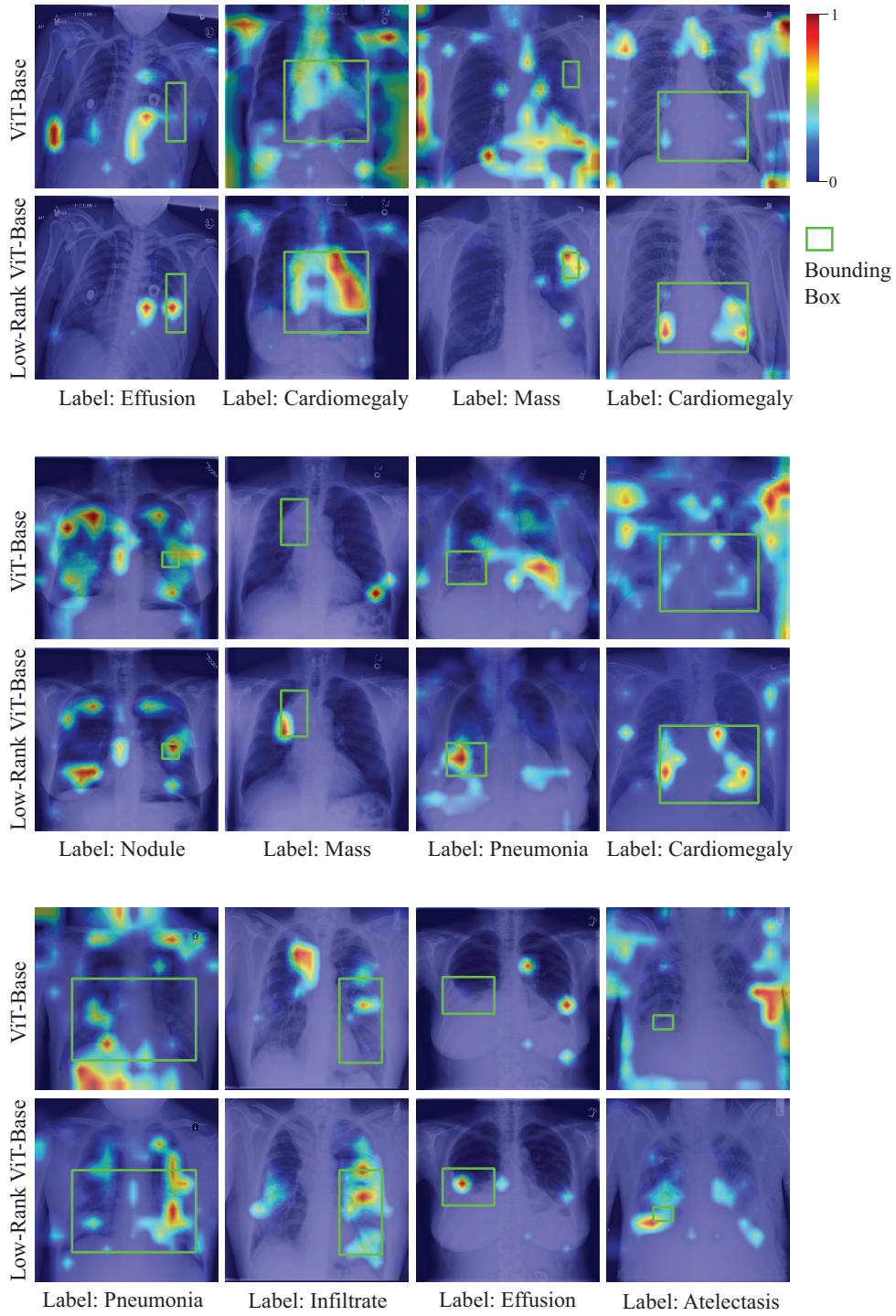


Figure 6: Grad-CAM visualization results. The figures in the first row are the visualization results of ViT-Base, and the figures in the second row are the visualization results of Low-Rank ViT-Base.