Learning Anatomy-Disease Entangled Representation

Fatemeh Haghighi[1] Michael B. Gotway[2]
Jianming Liang[1]
[1]Arizona State University, USA [2]Mayo Clinic, USA
{fhaghigh,jianming.liang}@asu.edu Gotway.Michael@mayo.edu

Abstract

Human experts demonstrate proficiency not only in disentangling anatomical structures from disease conditions but also in intertwining anatomical and disease information to accurately diagnose a variety of disorders. However, deep learning models, despite their prowess in acquiring intricate representation, often struggle to learn representation where distinct semantic aspects of the data (both anatomy and pathology) are entangled, particularly in medical images, which present a rich array of anatomical structures and potential pathological conditions. We envision that a deep model, when trained to comprehend medical images akin to human perception, would offer powerful representation with higher generalizability, robustness, and interpretability. To realize this vision, we have developed ***LeADER****, a framework for* ***le****arning* ***a****natomy-****d****isease* ***e****ntangled* ***r****epresentation from medical images. As a proof of concept, we have trained LeADER on ≈1M chest radiographs gathered from 10 public datasets. Experimental results across 11 medical tasks, compared to 8 baselines in zero-shot, linear probing, limited data regimes, and full fine-tuning settings, demonstrate LeADER's superior performance over the Google CXR Foundation Model, large-scale medical models, and fully/self-supervised baselines across diverse downstream tasks. This enhanced performance is attributed to the significance of entangling anatomy-specific and disease-specific representations via our framework, which enables the simultaneous acquisition of both anatomical and disease knowledge, yet overlooked in existing supervised/self-supervised learning methods. All code and models are available at GitHub.com/JLiangLab/LeADER.*

1. Introduction

Medical images contain diverse anatomical structures and potential pathological conditions. Human experts excel not only in separating (disentangling) anatomical structures from disease conditions but also in combining (entangling) anatomical and disease information for diagnosing various disorders. However, current deep learning models fall short of effectively learning representation where distinct semantic aspects of the data (both anatomy and pathology) are entangled [59, 62, 79]. We hypothesize that if deep neural networks can comprehend medical images akin to human perception—entangling anatomy-specific and disease-specific visual information—their learned representation would exhibit increased generalizability, robustness, and interpretability. To test this hypothesis, we have chosen chest radiography (CXR) because chest radiograph is the most widely used imaging modality worldwide, and there is a pressing need to develop robust CXR models [34, 64, 65]. Therefore, this paper seeks to address a critical question:

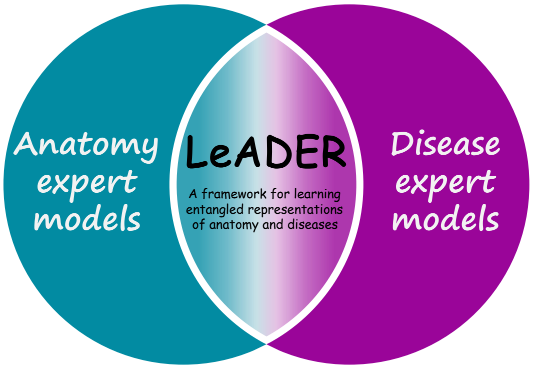

Anatomy expert models
LeADER
A framework for learning entangled representations of anatomy and diseases
Disease expert models

Figure 1. Human experts demonstrate proficiency in intertwining anatomical and disease information to accurately diagnose a variety of disorders. However, current representation learning approaches either focus on disease-related learning through utilization of expert disease labels, or focus on learning anatomy-specific features, disregarding disease-related features due to lack of appropriate supervision in their learning objectives. To emulate humans' ability to intertwine anatomical and disease information, we introduce a framework that explicitly learns to encode both pathological and anatomical information from medical images, leading to the the development of a powerful model (LeADER) that yields not only discriminative disease features but also semantically rich anatomical features.

How to learn entangled representations of diseases and anatomy from medical images, yielding more powerful models for a broad range of applications?

In answering this question, we have developed a framework, called **LeADER**, for its ability of **le**arning **a**natomy-**d**isease **e**ntangled **r**epresentation from medical images, as depicted in Fig. 2 and illustrated in Algorithm 1. Our extensive experiments on across 11 medical tasks, compared with 8 baselines in (i) zero-shot (Figs. 3 and 4), (ii) linear probing (Fig. 5), (iii) limited data transfer learning (Tab. 3), and (iv) full fine-tuning settings (Tab. 2), showcase LeADER's superior performance over Google's proprietary Foundation Model (CXR-FM) [65], large-scale medical models, and existing SSL methods across diverse downstream tasks.

LeADER's superior performance is attributed to the entanglement of anatomy-specific and disease-specific representations, which enables the simultaneous learning of both anatomical and disease knowledge; thereby, offering superior performance for both disease identification and anatomy understanding. LeADER differs fundamentally from existing representation learning methods, including supervised learning [54, 55, 65] which focus solely on disease features through the utilization of disease labels, and self-supervised learning [3, 26, 31, 60, 83, 94, 95] which are limited in capturing discriminative disease-relevant representation [45]. In contrast to both, LeADER exploits both pathological and anatomical cues as supervisory signals for learning more comprehensive representation from medical images.

In summary, we have made the following contributions:

- A framework that learns entangled representations of diseases and anatomy, yielding discriminative representations enriched with the semantics of anatomical and pathological knowledge.
- A set of empirical analyses in zero-shot settings that highlights the effectiveness of LeADER 's representations for both disease identification and anatomy understanding compared with existing disease expert foundation models, including Google CXR-FM.
- A comprehensive set of experiments that demonstrates the generalizability and robustness of LeADER 's representations across diverse tasks compared with fully-supervised and self-supervised learning baselines.

2. Related works

2.1. Supervised representation learning

Supervised representation learning focuses on pretraining deep models using expert-provided labels. In this paradigm, deep models are trained by minimizing the objective to align the model's predictions with expert labels [6, 42, 91]. In the context of medical image analysis, a substantial body of work has developed supervised pretrained models generalizable to various medical applications by assembling large-scale labeled medical datasets [34, 39, 41, 48, 54, 55, 58, 65]. Notably, the RadImageNet [55] model was developed using a large corpus of 1.35 million radiology images in a fully supervised manner, demonstrating the significance of pretraining with millions of radiology images compared to the ImageNet dataset. Additionally, Google's proprietary CXR Foundation Model (CXR-FM) [65] was trained on 821,544 labeled chest X-ray scans, demonstrating generalizability to a range of medical tasks. Despite the success of these models, they tend to retain more disease-specific information due to the use of disease labels as their supervisory signals, while overlooking anatomy-related features. By contrast, our LeADER not only leverages the power of existing disease expert models to learn discriminative disease features but also concurrently learns anatomy-related features, leading to more comprehensive representation for a diverse range of medical tasks.

2.2. Self-supervised representation learning

Self-supervised learning (SSL) focuses on pretraining deep models without expert-provided labels. Instance discrimination SSL [5, 7, 8, 14–18, 21, 29, 44, 75, 82, 85, 88, 90, 93] has emerged as a prominent approach for visual representation learning, where the pivotal idea is to consider each image as a unique class and train a model to align the representations of the augmented views from the same image. On the other hand, reconstruction-based pretext tasks, particularly masked image modeling methods [4, 12, 19, 28, 38, 43, 47, 57, 73, 76, 77, 86, 87], mask or perturb random parts of the input image and reconstruct the missing parts at the pixel level. In the context of medical imaging, both instance discrimination and masked image modeling have been widely studied [2, 3, 13, 24, 26, 33, 40, 60, 72, 94, 95], while recent SSL methods seek to learn consistent anatomical representations through carefully designed pretext tasks [9, 25, 27, 31, 37, 71]. Among them, Adam [31, 71] has recently shown potential in learning anatomy-specific representations by exploiting the hierarchical nature of anatomical structures in its learning objective. Despite their success, SSL methods are limited in capturing discriminative disease-relevant representations due to the lack of explicit disease learning supervision in their training objectives. Our LeADER addresses this limitation by incorporating disease learning, enforcing the model to capture both anatomy and disease features.

2.3. Knowledge distillation

Knowledge distillation (KD) methods focus on training a student network to mimic the output of a teacher network,

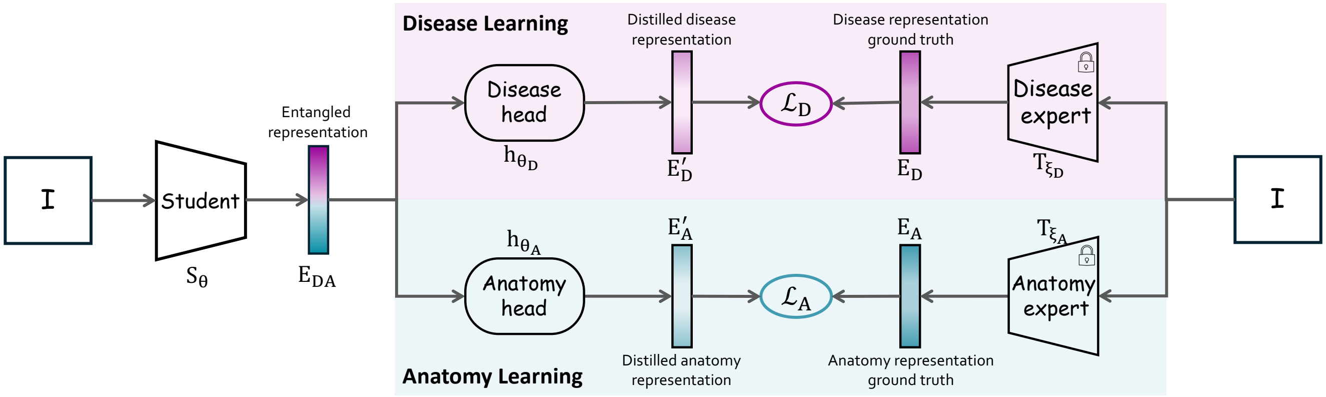


Figure 2. LeADER learns entangled representation of diseases and anatomy by distilling anatomical and pathological information from two expert teachers via two branches: disease learning and anatomy learning. Given input image/patch I, we pass it to student S_θ to get its representation E_{DA}. The disease learning branch extracts disease-specific representation E'_D from E_{DA} via disease head h_{θ_D} and maximizes the consistency between E'_D and embedding E_D of I generated by a disease expert teacher, such as Google CXR-FM [65], known for its proficiency in generating disease-related features. The anatomy branch extracts anatomy-specific representation E'_A from E_{DA} via anatomy head h_{θ_A} and maximizes the alignment between E'_A and embedding E_A of I generated by an anatomy expert teacher, such as Adam [31], which excels in generating anatomy-related features. By joint training of the disease and anatomy branches, student S_θ is compelled to capture entangled representation, while the disease/anatomy heads encode distilled disease/anatomy representation. We ablate LeADER in zero-shot, linear-probing, and fine-tuning settings and show its superiority for both anatomy understanding (Fig. 3) and disease identification (Figs. 4 to 6, Tab. 2, and Tab. 3).

aiming to transfer the knowledge of the teacher model to the student model [30]. KD has been widely explored in a supervised manner. Numerous works [10, 11, 23, 30, 36, 56, 63, 96] minimize the difference between the outputs of teacher and student networks at different stages via different similarity objectives, in addition to utilizing class label supervision. Moreover, recent works [8, 15, 20, 69, 74, 81, 84] have extended KD to the self-supervised paradigm by designing various pretext tasks to align the outputs of student and teacher networks without relying on human supervision, demonstrating promising outcomes for a range of vision tasks. Additionally, a line of work has adopted multiple teachers to enhance KD with more robust features [22, 50, 68, 92]. In contrast to all existing KD methods, our LeADER simultaneously learns pathological and anatomical information by distilling knowledge from both disease and anatomy expert teachers, leading to entangled representations of diseases and anatomy.

3. Method

Our framework, depicted in Fig. 2, aims to learn entangled representations of diseases and anatomy. To do so, our framework employs a student network (S_θ) that simultaneously learns anatomical and diseases information from two teacher models—one specializing in diseases and the other in anatomy—via two key learning branches: (1) anatomy learning, aiming to encode semantically rich anatomical features, and (2) disease learning, aiming to acquire discriminative disease-related features. By integrating these learning branches into a unified framework, our method captures comprehensive anatomy-related and disease-related information, providing more powerful representations for various downstream tasks. In the following, we first introduce each branch and then describe the joint training loss.

3.1. Anatomy learning

Anatomy learning branch aims to empower the student S_θ with semantics-rich anatomical features by distilling knowledge from an anatomy expert teacher T_{ξ_A}, which is proficient in understanding anatomy. This branch takes the input I and generate its latent embedding E_{DA} with the student S_θ network. The embedding E_{DA} is then processed by the anatomy head h_{θ_A}, which extracts anatomy-related features from the entangled features E_{DA} and outputs the distilled anatomy embedding $E'_A = h_{\theta_A}(E_{DA})$. The input I is also fed to the anatomy expert model T_{ξ_A} to generate anatomy embedding $E_A = T_{\xi_A}(I)$, serving as targets for training the anatomy branch. The objective of the anatomy branch is to maximize the consistency between the anatomy expert's embeddings and those generated by the anatomy head:

$$\mathcal{L}_A = \ell_s(E_A, E'_A) \quad (1)$$

where $\ell_s(.)$ is a function that measures the similarity between E_A and E'_A, and can be any suitable function such as cross-entropy, etc.

Code	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
Dataset	PadChest [1]	ChestX-ray14 [80]	CheXpert [34]	Shenzhen [35]	RSNA Pneumonia [70]	MIMIC-CXR [39]	VinDR-CXR [61]	ChestX-Det [46]	Node21 [67]	TBX-11K [49]
#Samples	160,828	75,312	223,414	463	21,295	368,879	36,096	15,810	2,952	2,422

Table 1. **Pretraining datasets:** LeADER is trained on around 900K samples collected from the training sets of 10 public datasets. Alongside the images in D_1-D_{10}, disease bounding box labels from D_7-D_{10} are utilized to extract diseased patches for training, enriching LeADER with nuanced disease/anatomy features. Specifically, for D_7-D_{10}, the provided bounding box labels are employed to extract diseased patches from diseased images, while random normal patches are sampled from healthy images.

Algorithm 1: A round of training LeADER

```
Data: Image Datasets $\mathcal{D}_I = \{D_1, D_2, D_3, D_4, D_5, D_6, D_7\}$
      Images with Bounding Boxes Datasets
      $\mathcal{D}_{IB} = \{D_7, D_8, D_9, D_{10}\}$
Trainable Parameters: Student $S_\theta$, Disease/Anatomy heads $h_{\theta_D}, h_{\theta_A}$
Frozen Parameters: Disease expert $T_{\xi_D}$, Anatomy expert $T_{\xi_A}$
$\mathcal{D}_{IP} = \{\}$
// Extract patches from datasets with disease bounding box labels
for $D_{ib} \in \mathcal{D}_{IB}$ do
    Extract diseased patches using the provided bounding box labels for
      diseased images
    Extract random healthy patches from normal images
    Create a new dataset $D_{ip}$ with pairs $(I, P)$ of patches and their
      corresponding images
    $\mathcal{D}_{IP} \leftarrow D_{ip}$
// Train disease branch with images as input
for $D_i \in \mathcal{D}_I$ do
    for $I \in D_i$ do
        $E_D, E'_D = T_{\xi_D}(I), h_{\theta_D}(S_\theta(I))$
        Loss = $\mathcal{L}_D(E_D, E'_D)$
        // Optimize trainable parameters with back-propagation
        Update($\{S_\theta, h_{\theta_D}\}, Loss$)
// Train disease and anatomy branches jointly with images/patches as input
for $D_{ip} \in \mathcal{D}_{IP}$ do
    // (I,P): pair of image/patch
    for $(I, P) \in D_{ip}$ do
        $E_{DA_I}, E_{DA_P} = S_\theta(I),\ S_\theta(P)$
        $E_{D_I}, E_{D_P} = T_{\xi_D}(I),\ T_{\xi_D}(P)$
        $E'_{D_I}, E'_{D_P} = h_{\theta_D}(E_{DA_I}),\ h_{\theta_D}(E_{DA_P})$
        $E_{A_I}, E_{A_P} = T_{\xi_A}(I),\ T_{\xi_A}(P)$
        $E'_{A_I}, E'_{A_P} = h_{\theta_A}(E_{DA_I}),\ h_{\theta_A}(E_{DA_P})$
        $Loss_{Disease}$ = $\mathcal{L}_D(E_{D_I}, E'_{D_I}) + \mathcal{L}_D(E_{D_P}, E'_{D_P})$
        $Loss_{Anatomy}$ = $\mathcal{L}_A(E_{A_I}, E'_{A_I}) + \mathcal{L}_A(E_{A_P}, E'_{A_P})$
        $Loss = Loss_{Disease} + Loss_{Anatomy}$
        Update($\{S_\theta, h_{\theta_D}, h_{\theta_A}\}, Loss$)
```

3.2. Disease learning

Disease learning branch aims to equip the student S_θ with discriminative disease-related features by distilling knowledge from a disease expert teacher T_{ξ_D}, which specializes in generating disease-related features. This branch takes the input I, which can be a healthy/diseased image or patch, and process it with the student S_θ to generate its latent embedding E_{DA}. The embedding E_{DA} is then passed to the disease head h_{θ_D}, which extracts disease-related features from the entangled features E_{DA} and outputs the distilled disease embedding $E'_D = h_{\theta_D}(E_{DA})$. The input I is also passed to the disease expert teacher T_{ξ_D}, which generates disease embedding $E_D = T_{\xi_D}(I)$, serving as the ground truth for the disease learning branch. The disease branch's objective is to maximize the alignment between the disease expert's embeddings and those generated by the disease head using the following general loss function:

$$\mathcal{L}_D = \ell_s(E_D, E'_D) \tag{2}$$

where $\ell_s(.)$ is a function that measures similarity between E_D and E'_D, and can be cross-entropy, mean squared error (MSE), or any other sophisticated measures.

3.3. Training pipeline

LeADER is a general framework that allows for various choices of disease and anatomy expert models without any constraints. Moreover, LeADER works with different types of inputs, including whole images as well as patches. Thus, if a dataset includes disease bounding boxes, patches can be optionally extracted and used alongside images to distill local disease and anatomy features for specific regions, thereby enriching representation learning. To enable end-to-end representation learning from both disease and anatomy experts, LeADER integrates disease and anatomy branches and jointly train them with one single objective:

$$\mathcal{L}_{LeADER} = \mathcal{L}_D + \mathcal{L}_A \tag{3}$$

Through our unified training scheme, our framework yields both entangled and distilled representations of diseases and anatomy. In particular, the joint optimization of $\mathcal{L}_A$ and $\mathcal{L}_D$ enforces the student S_θ to simultaneously encode anatomy- and disease-related information from the input, resulting in entangled anatomy-disease representations, denoted as E_{DA}. Moreover, the anatomy and disease heads (h_{θ_A} and h_{θ_D}) extract exclusive anatomy-specific and disease-specific factors from the entangled representation E_{DA}, mapping the input into distilled anatomy and disease representations E'_A and E'_D , respectively.

4. Implementation details

Pretraining settings. We use Swin transformer base (Swin-B) [51] as the backbone of student S_θ, and two-layers MLP heads for h_{θ_D} and h_{θ_A}. For the disease expert T_{ξ_D}, we employ Google CXR-FM [65], known for its proficiency in generating disease-related features. For the anatomy expert T_{ξ_A}, we employ Adam [31, 71], which excels in generating anatomy-related features. Other suitable disease/anatomy expert models can also be integrated into our framework

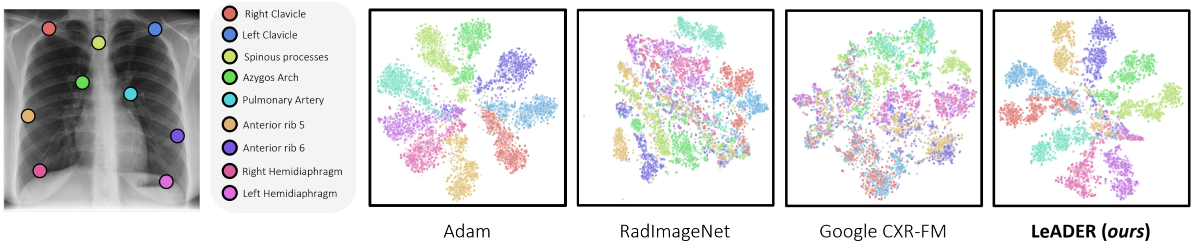


Figure 3. **Zero-shot anatomy understanding:** Compared with CXR-FM and RadImageNet models, which solely focus on disease learning, LeADER excels in anatomy understanding, demonstrating a strong capability in discriminating different anatomical landmarks in its embedding space.

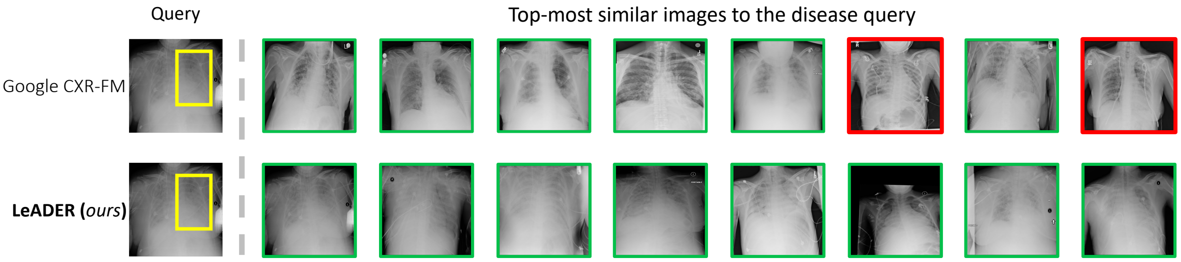


Figure 4. **Samples of top retrieved images:** LeADER outperforms CXR-FM in zero-shot disease retrieval, showcasing the efficacy of its entangled representation of anatomy and diseases. While CXR-FM exhibits errors in its top similar images (red boxes) to a sample disease query (yellow box), LeADER accurately identifies diseased images (green boxes) that share the same abnormality with the query.

without constraint. During training, we optimize S_θ, h_{θ_D}, and h_{θ_A} using the AdamW optimizer with a learning rate of $2e-4$, while keeping the T_{ξ_D} and T_{ξ_A} frozen. MSE is used as $\mathcal{L}_D$ and $\mathcal{L}_A$. Random affine transformation, horizontal flip, and color jitter are used as data augmentation. Detailed information of LeADER's training data and procedure is provided in Tab. 1 and Algorithm 1 and Appendix.

Evaluations. We extensively evaluate LeADER in (1) Zero-shot anatomy understanding, (2) disease retrieval, (3) linear probing, and (4) full transfer settings. We consider 11 downstream tasks on nine publicly available datasets, including ChestX-ray14 [80], CheXpert [34], ChestX-Det [46], VinDr-CXR [61], NIH Shenzhen [35], RSNA Pneumonia [70], SIIM-ACR [89], COVIDx [78], and JSRT [66]. These tasks rigorously examine the generalizability of our LeADER across a diverse range of applications. Dataset details are provided in Appendix.

Baselines. We compare LeADER with recent SOTA fully supervised and self-supervised pretrained models. Particularly, we consider models pretrained on large-scale labeled medical datasets, including RadImageNet and Google CXR-FM. Moreover, we compare LeADER with a representative set of SOTA publicly-available SSL baselines tailored for medical imaging tasks, encompassing PCRL [94], Adam [31], DiRA [26], DINO [8], Medical-MAE [83], and LVM-Med [60]. Among the baselines, DiRA and PCRL are multi-task learning methods, while DINO and LVM-Med are knowledge distillation methods. Notably, LVM-Med is pretrained on a large corpus of 1.3 million medical images.

Fine-tuning settings. Following the standard transfer learning protocol [32], we fine-tune the student network of LeADER (S_θ) for diverse classification and segmentation tasks. For transfer learning to classification tasks, we attach a classification head to the pretrained backbone, and for segmentation tasks, we use a UperNet network, where the encoder is initialized with the pretrained weights. We perform end-to-end training by fine-tuning all the parameters of the downstream models. Details of the fine-tuning hyperparameters are provided in the Appendix.

5. Results

5.1. LeADER elevates anatomy understanding

To assess LeADER's proficiency in understanding anatomy, we investigate its ability to discriminate various anatomical structures in a zero-shot setting (with no fine-tuning). To do so, we (1) leverage a dataset of 1,000 images (from ChestX-ray14 dataset [80]) with ten distinct anatomical landmarks manually annotated by human experts in each image, (2) extract 224^2 patches around each landmark across images, (3) generate embeddings of landmark instances using each model under study, and (4) visualize the embeddings with t-SNE plot. We compare LeADER with Google CXR-FM [65] and RadImageNet [55], two recently

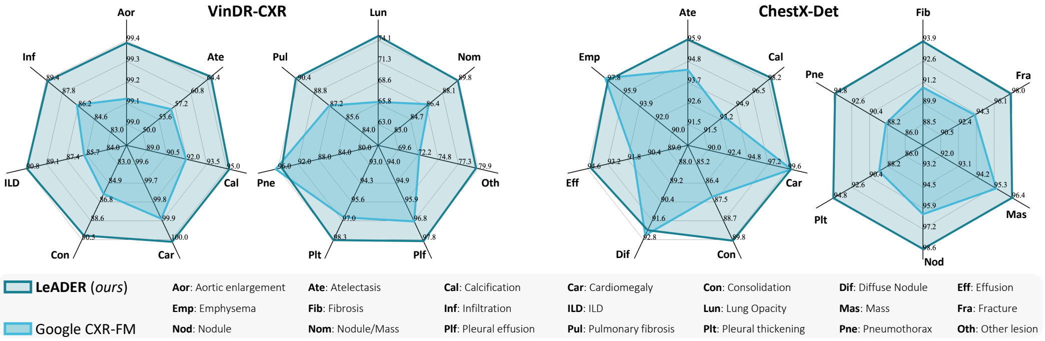


Figure 5. **Disease classification at lesion level:** LeADER provides significantly better performance ($p < 0.05$) compared with Google CXR-FM in disease classification at lesion level across datasets and diverse diseases, showcasing its effectiveness in enhancing representation learning by capturing entangled anatomy and disease information.

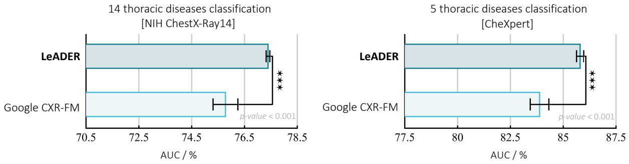


Figure 6. **Disease classification at image level:** LeADER significantly ($p < 0.05$) outperforms Google CXR-FM in disease classification at image level across tasks, highlighting the significance of our framework in integrating anatomy and disease learning.

developed medical models pretrained on large-scale medical datasets with disease labels. As seen in Fig. 3, RadImageNet and CXR-FM fail to distinguish between different anatomical structures. By contrast, LeADER effectively discriminates between various anatomical landmarks, resulting in well-separated clusters in its embedding space. This highlights LeADER's ability to learn semantics-rich anatomical embeddings, a desired property that is absent in existing disease expert foundation models, thus facilitating the more effective capture of disease-related information.

5.2. LeADER excels in zero-shot disease retrieval

To showcase the significance of LeADER's learned entangled representations, we examine LeADER's capability in identifying diseased patterns across images compared with Google CXR-FM and RadImageNet model by conducting zero-shot disease retrieval experiments (no training or tuning). To do so, we focus on Pneumonia disease in the RSNA Pneumonia dataset [70], where we: (1) select a random diseased image and then extract a diseased patch (disease query) based on the provided bounding box label for the image; (2) extract embeddings of the disease query as well as images in the RSNA Pneumonia dataset using each model under the study; (3) compute cosine similarity between the embedding of disease query and embeddings of each image; (4) rank the top-K most similar images to the disease query and calculate the precision score for the retrieved images. We set K to 100 and replicated the experiment using multiple random query images, wherein LeADER exhibits 15.3% and 3.7% higher disease retrieval performance compared to RadImageNet and Google CXR-FM, respectively (RadImageNet: 71.67±6.66, CXR-FM: 83.33±7.02 vs. **LeADER: 87.00±7.21**). In addition to our reported quantitative results, we also provide qualitative findings in Fig. 4, which presents a sample disease query and its top eight most similar images retrieved by both LeADER and CXR-FM. As seen, contrary to CXR-FM, which exhibits errors in its top similar samples (red boxes), LeADER accurately identifies diseased images. These findings underscore the effectiveness of our framework in capturing both complementary disease- and anatomy-related visual information from medical images, in contrast to CXR-FM, which solely focuses on capturing disease-related information.

5.3. LeADER provides superior representations for disease identification

To further scrutinize the effectiveness of our learned entangled representations, we conduct a comparative analysis against CXR-FM in thoracic diseases classification using linear probing across four downstream tasks. These tasks encompass image-level classification on ChestX-ray14 and CheXpert [34] with 10% labeled data, as well as lesion-level classification on ChestX-Det [46] and VinDR-CXR [61]. As seen in Fig. 6, in both the ChestX-ray14 and CheXpert datasets, each comprising labels for 14 and 5 thoracic diseases at the image level, LeADER demonstrates superior performance compared to the CXR-FM, achieving an average performance improvement of 1.6% and 1.9%, respectively. To delve deeper into the effectiveness of LeADER's representations compared to CXR-FM, we employ linear probing with embeddings extracted from the CXR-FM API

Method	Classification				Segmentation	
	14 Thoracic diseases	Tuberculosis	COVID-19	Lung nodule	Pneumothorax	13 Thoracic diseases
	(ChestX-ray14)	(Shenzhen)	(COVIDx)	(JSRT)	(SIIM-ACR)	(ChestX-Det)
ImageNet	81.74±0.13	93.35±0.77	95.90±0.65	51.73±0.81	70.22±0.57	74.89±0.16
RadImageNet	79.96±0.11	93.77±0.58	95.90±1.19	54.74±3.27	70.31±0.52	73.75±0.79
PCRL	80.56±0.08	93.18±1.38	95.50±1.95	60.69±4.10	68.01±1.28	71.45±0.59
DiRA	81.12±0.17	92.94±0.98	94.95±3.76	61.05±2.91	69.24±0.41	71.90±0.96
DINO	79.01±0.08	92.11±0.22	93.20±0.48	56.32±0.42	68.18±2.28	73.97±0.21
Medical-MAE	82.24±0.03	95.77±0.33	95.10±0.38	52.08±1.80	70.13±0.62	75.49±0.40
Adam	81.72±0.44	94.60±0.90	95.20±0.65	57.43±5.45	69.59±0.55	74.17±0.57
LVM-Med	81.76±0.09	95.11±1.02	96.45±0.48	50.79±5.43	69.06±1.26	74.79±0.03
LeADER (*Ours*)	**82.52±0.06**	**98.28±0.21**	**96.60±0.29**	**61.65±4.84**	**70.93±0.54**	**76.72±0.27**

Table 2. **Full transfer learning results:** LeADER demonstrates superior transfer performance compared to both fully- and self-supervised baselines across all tasks, showcasing the significance of our framework in capturing transferable features. ± denotes standard deviation. For each task, we conducted the independent two-sample t-test between the best (bolded) vs. others. Highlighted boxes in blue indicate statistical significance at the $p = 0.05$ level.

and the pretrained LeADER model on the ChestX-Det and VinDR-CXR datasets, which include the evaluation of diagnosing 13 and 14 common thoracic diseases, respectively, at the lesion level. As seen in Fig. 5, LeADER outperforms CXR-FM in 13 and 10 diseases within the VinDR-CXR and ChestX-Det datasets, leading to an average performance enhancement of 3.5% and 2.6%, respectively, demonstrating LeADER's capability in providing superior representations for the identification of a broad range of thoracic diseases. Our attribution of LeADER's superior representations over Google CXR-FM, evidenced in both zero-shot disease retrieval and disease identification, is grounded in the significance of our anatomy learning in conjunction with disease learning, which is neglected in existing disease learning approaches, including Google CXR-FM.

5.4. LeADER provides transferable representations for a variety of tasks

To highlight the importance of leveraging both disease and anatomical cues as supervisory signals for representation learning, we evaluate the transferability of LeADER's representations across six target tasks, including classification on the ChestX-ray14, NIH Shenzhen, COVIDx, and JSRT datasets, as well as segmentation on the SIIM-ACR and ChestX-Set datasets. We compare LeADER against a representative set of eight publicly available supervised and self-supervised baselines, including fully-supervised ImageNet and RadImageNet, as well as self-supervised models PCRL [94], Adam [31], DiRA [26], DINO [8], Medical-MAE [83], and LVM-Med [60]. As shown in Tab. 2, LeADER consistently demonstrates significantly superior transfer performance ($p < 0.05$) compared to both fully- and self-supervised baselines across all downstream tasks. Notably, our LeADER shows a significant performance improvement over Adam, its anatomical expert teacher, underscoring LeADER's efficacy in capturing more discriminative features for disease detection through its integrated disease learning branch. Additionally, LeADER outperforms multi-task learning (i.e. PCRL & DiRA) and knowledge distillation (DINO & LVM-Med) baselines across all tasks. These results suggest that LeADER serves as a comprehensive representation learning framework adept at capturing intricate pathological and anatomical information from images, thereby enriching visual representations to generalize more effectively across various medical tasks.

Method	SIIM-ACR		ChestX-Det	
	5%	10%	5%	10%
RadImageNet	54.56	61.48	64.22	67.10
PCRL	47.12	54.48	60.36	63.49
DiRA	42.44	48.27	61.63	64.86
DINO	47.85	52.08	46.84	52.64
Medical-MAE	60.54	61.05	66.86	67.31
Adam	52.47	65.82	64.30	65.80
LVM-Med	54.13	62.31	65.11	67.14
LeADER (*Ours*)	**62.23** (↑1.69)	**68.80** (↑2.98)	**68.81** (↑1.95)	**70.24** (↑2.93)

Table 3. **Transfer learning in limited data regimes:** LeADER excels in limited labeled data regimes, highlighting its significance for medical applications with scarce labeled data.

5.5. LeADER enhances robustness in limited data regimes

To demonstrate the robustness of representations learned via LeADER in small data regimes, we examine transfer learning using partially labeled data. We randomly sample different fractions (5% and 10%) of training data from the SIIM-ACR and ChestX-Det datasets and fine-tune the pretrained models under study on these training data subsets. As shown in Tab. 3, LeADER provides superior performance across all label fractions and downstream tasks. Specifically, LeADER yields performance boosts of 1.69 and 2.98 in SIIM-ACR, and 1.95 and 2.93 in ChestX-Det when using 5% and 10% of the training data, respectively. These results underscore the superiority of our framework in capturing more robust and generalizable representations,

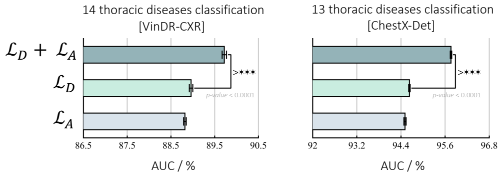
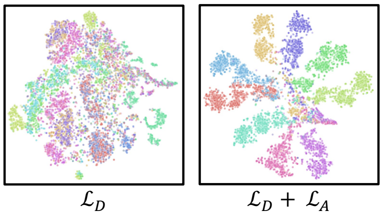


Figure 7. **Ablation study on LeADER's learning objectives:** integrating disease and anatomy branches significantly enhances performance compared to using each individual branch.



$\mathcal{L}_D$ $\mathcal{L}_D + \mathcal{L}_A$

Figure 8. **Ablation study on impact of anatomy branch on anatomy understanding:** anatomy branch plays a crucial role in learning more discriminative anatomical representations.

paving the way for the development of more accurate models for medical tasks with a dearth of labeled data.

6. Ablation studies

Impact of LeADER's learning branches. We investigate the impact of each learning branch in LeADER by comparing the performance of individual branches with the unified model that incorporates both branches. We evaluate the models via linear probing on the ChestX-Det and VinDR-CXR datasets. As illustrated in Fig. 7, integrating both branches significantly enhances performance across tasks compared to the individual branches. Moreover, Fig. 8 demonstrates the impact of anatomy learning on enhancing LeADER's anatomy understanding capabilities. These results demonstrate the significance of anatomy learning in conjunction with disease learning, which not only boosts disease identification but also equips the model with discriminative anatomical representation.

Comparison with ensemble methods. To demonstrate the efficacy of our training strategy for learning entangled anatomy-disease representation, we compare LeADER with two baselines that ensemble the anatomy and disease expert models: (1) concatenation, which concatenates the representations of expert teachers and uses them as input for downstream tasks, and (2) averaging ensemble, which uses each teacher individually for downstream tasks, followed by averaging their outputs for a final prediction. Fig. 9 shows superiority of LeADER, highlighting the limitations of these ensemble methods in generating entangled representations essential for diagnosing various disorders.

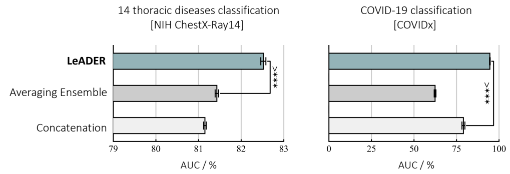


Figure 9. **Comparison with ensemble methods:** LeADER outperforms simple ensemble of disease and anatomy teachers.

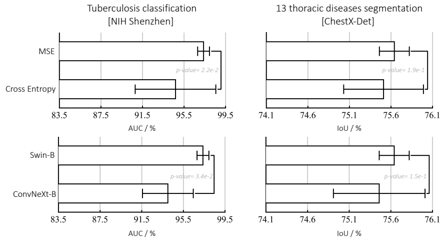


Figure 10. Ablation studies on (a) loss function (top-row) and (b) architecture choice for LeADER's backbone (bottom-row).

Impact of training loss function. We examine the impact of the knowledge distillation loss function (ℓ_s) on downstream performance by comparing different distance metrics, specifically MSE and Cross Entropy. For computational efficiency, we train the models with $\mathcal{L}_D$ on the ChestX-ray14 dataset for 100 epochs. We evaluate the models via fine-tuning on two downstream tasks. As illustrated in Fig. 10 (top-row), MSE loss yields superior performance than Cross Entropy across downstream tasks.

Impact of architecture. We investigate the impact of architecture choices on downstream performance by evaluating SOTA ConvNet and vision transformer backbones for the student model, specifically Swin-B [52] and ConvNeXt-B [53]. We use the same settings as the loss function ablation study. Fig. 10 (bottom-row) shows the superiority of Swin-B backbone over ConvNeXt-B.

7. Conclusion

We present LeADER, a framework that aims to offer powerful representation with higher generalizability, robustness, and interpretability by training deep models to comprehend medical images akin to human perception. The major novelty of LeADER is learning entangled disease-anatomy representation by distilling anatomical and pathological information from two expert models via disease and anatomy learning branches, enabling the collaborative and simultaneous learning of anatomical and disease knowledge, yet overlooked in existing supervised/self-supervised learning methods. Our experiments demonstrate the efficacy of LeADER in zero-shot, linear probing, and full fine-tuning settings.

References

[1] Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. 4

[2] Shekoofeh Azizi, Laura Culp, Jan Freyberg, and et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7:756–779, 2023. 2

[3] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. *arXiv:2101.05224*, 2021. 2

[4] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24256–24265, June 2023. 2

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. 2

[6] Eric Baum and Frank Wilczek. Supervised learning of probability distributions by neural networks. In D. Anderson, editor, *Neural Information Processing Systems*, volume 0. American Institute of Physics, 1987. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. 2, 3, 5, 7

[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. 2

[10] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3430–3437, Apr. 2020. 3

[11] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7028–7036, May 2021. 3

[12] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22742–22751, June 2023. 2

[13] 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. 2

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2

[15] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020. 2, 3

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 2

[17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021. 2

[18] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3015–3024. PMLR, 18–24 Jul 2021. 2

[19] Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10386–10395, June 2023. 2

[20] Yuting Gao, Jia-Xin Zhuang, Shaohui Lin, Hao Cheng, Xing Sun, Ke Li, and Chunhua Shen. Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 237–253, Cham, 2022. Springer Nature Switzerland. 3

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 2

[22] Shuangchun Gui, Zhenkun Wang, Jixiang Chen, Xun Zhou, Chen Zhang, and Yi Cao. Mt4mtl-kd: A multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging*, 43(4):1628–1639, 2024. 3

[23] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[24] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? *Medical Image Analysis*, 94:103086, 2024. 2

[25] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 137–147, Cham, 2020. Springer International Publishing. 2

[26] 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 (CVPR)*, pages 20824–20834, June 2022. 2, 5, 7

[27] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 40(10):2857–2868, 2021. 2

[28] 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 (CVPR)*, pages 16000–16009, June 2022. 2

[29] 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 (CVPR)*, June 2020. 2

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[31] Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. In *Domain Adaptation and Representation Transfer*, pages 94–104, Cham, 2024. Springer Nature Switzerland. 2, 3, 4, 5, 7

[32] 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, Cham, 2021. Springer International Publishing. 5

[33] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B. Gotway, and Jianming Liang. Caid: Context-aware instance discrimination for self-supervised learning in medical imaging. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 535–551. PMLR, 06–08 Jul 2022. 2

[34] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, and et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv:1901.07031*, 2019. 1, 2, 4, 5, 6

[35] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6), 2014. 4, 5

[36] Sajid Javed, Arif Mahmood, Talha Qaiser, and Naoufel Werghi. Knowledge distillation in histology landscape by multi-layer features supervision. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2037–2046, 2023. 3

[37] Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15859–15869, October 2023. 2

[38] Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, and Zhangyang Wang. Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[39] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. 2, 4

[40] Aakash Kaku, Sahana Upadhya, and Narges Razavian. Intermediate layers matter in momentum contrastive self supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 24063–24074. Curran Associates, Inc., 2021. 2

[41] Mintong Kang, Bowen Li, Zengle Zhu, Yongyi Lu, Elliot K. Fishman, Alan Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. 2

[42] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 2

[43] Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7918–7928, June 2023. 2

[44] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021. 2

[45] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *The Twelfth International Conference on Learning Representations*, 2024. 2

[46] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 4, 5, 6

[47] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6252–6261, June 2023. 2

[48] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[49] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2655, 2020. 4

[50] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. 3

[51] 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 (ICCV)*, pages 10012–10022, October 2021. 4

[52] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 8

[53] 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 (CVPR)*, pages 11976–11986, June 2022. 8

[54] DongAo Ma, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Foundation ark: Accruing and reusing knowledge for superior and robust performance. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 651–662, 2023. 2

[55] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, and et al. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022. 2, 5

[56] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198, Apr. 2020. 3

[57] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations, 2022. 2

[58] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, Shruthi Prabhakara, Umesh Telang, Alan Karthikesalingam, Neil Houlsby, and Vivek Natarajan. Supervised transfer learning at scale for medical imaging. *CoRR*, abs/2101.05913, 2021. 2

[59] Ivars Namatevs, Arturs Nikulins, Anda Slaidina, Laura Neimane, Oskars Radziņš, and Kaspars Sudars. Towards explainability of the latent space by disentangled representation learning. *Information Technology and Management Science*, 26:41–48, 11 2023. 1

[60] Duy M. H. Nguyen, Hoang Nguyen, Nghiem T. Diep, Tan N. Pham, Tri Cao, Binh T. Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, Daniel Sonntag, and Mathias Niepert. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching, 2023. 2, 5, 7

[61] Ha Q. Nguyen and et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020. 4, 5, 6

[62] Shishi Qiao, Ruiping Wang, Shiguang Shan, and Xilin Chen. Hierarchical disentangling network for object representation learning. *Pattern Recognition*, 140:109539, 2023. 1

[63] Dian Qin, Jia-Jun Bu, Zhe Liu, Xin Shen, Sheng Zhou, Jing-Jun Gu, Zhi-Hua Wang, Lei Wu, and Hui-Fen Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40(12):3820–3831, 2021. 3

[64] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, and et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. 1

[65] Andrew B. Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, and et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022. 1, 2, 3, 4, 5

[66] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174(1):71–74, 2000. 5

[67] Ecem Sogancioglu, Bram van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, and et al. Nodule detection and generation on chest x-rays: Node21 challenge, 2024. 4

[68] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using mul-

tiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9395–9404, October 2021. 3

[69] Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multimode online knowledge distillation for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11848–11857, June 2023. 3

[70] Anouk Stein, Carol Wu, Chris Carr, and et al. Rsna pneumonia detection challenge, 2018. 4, 5, 6

[71] Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Representing part-whole hierarchies in foundation models by learning localizability composability and decomposability from anatomy via self supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11281, June 2024. 2, 4

[72] 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 (CVPR)*, pages 20730–20740, June 2022. 2

[73] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, June 2023. 2

[74] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. 3

[75] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning?, 2020. 2

[76] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10375–10385, June 2023. 2

[77] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2122–2131, June 2023. 2

[78] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covidnet: 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. 5

[79] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning, 2023. 1

[80] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestxray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017. 4, 5

[81] Yongwei Wang, Yuheng Wang, Jiayue Cai, Tim K. Lee, Chunyan Miao, and Z. Jane Wang. Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Medical Image Analysis*, 84:102693, 2023. 3

[82] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[83] 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 (WACV)*, pages 3588–3600, 2023. 2, 5, 7

[84] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[85] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *CoRR*, abs/2105.04553, 2021. 2

[86] 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. 2

[87] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10365–10374, June 2023. 2

[88] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[89] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. 5

[90] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230*, 2021. 2

[91] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022. 2

[92] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502, 2022. 3

[93] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[94] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3499–3509, 2021. 2, 5, 7

[95] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 2

[96] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, June 2021. 3