Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification

Junfei Xiao Yutong Bai Alan Yuille Zongwei Zhou*
Johns Hopkins University

Code: https://github.com/lambert-x/Medical_MAE

Abstract

Vision Transformer (ViT) has become one of the most popular neural architectures due to its great scalability, computational efficiency, and compelling performance in many vision tasks. However, ViT has shown inferior performance to Convolutional Neural Network (CNN) on medical tasks due to its data-hungry nature and the lack of annotated medical data. In this paper, we pre-train ViTs on 266,340 chest X-rays using Masked Autoencoders (MAE) which reconstruct missing pixels from a small part of each image. For comparison, CNNs are also pre-trained on the same 266,340 X-rays using advanced self-supervised methods (e.g. *MoCo v2). The results show that our pre-trained ViT performs comparably (sometimes better) to the state-of-the-art CNN (DenseNet-121) for multi-label thorax disease classification. This performance is attributed to the strong recipes extracted from our empirical studies for pre-training and fine-tuning ViT. The pre-training recipe signifies that medical reconstruction requires a much smaller proportion of an image (10% vs. 25%) and a more moderate random resized crop range (0.5∼1.0 vs. 0.2∼1.0) compared with natural imaging. Furthermore, we remark that in-domain transfer learning is preferred whenever possible. The fine-tuning recipe discloses that layer-wise LR decay, RandAug magnitude, and DropPath rate are significant factors to consider. We hope that this study can direct future research on the application of Transformers to a larger variety of medical imaging tasks.*

1. Introduction

There has been great progress in the Vision Transformer (ViT) architecture [26] and its variants [62, 43, 37, 91], showing that Transformers surpass and supersede Convolutional Neural Networks (CNN) in various natural imaging tasks. In comparison with CNNs, Transformers can better leverage the rapidly increasing image data, long-range spatial context of an image [21, 25], and share properties of the human visual system [71, 77, 30, 92]. Training Transformers requires considerably more data than CNNs [83, 89], but medical data are small and labels are hard to obtain. As a result, directly applying Transformers to the medical domain is found to be problematic and challenging. There are several early attempts [70, 99, 67], but their performance is often inferior to state-of-the-art CNNs (elaborated in §2). Recent surveys suggest that a range of successful cases are using a hybrid architecture of Transformers and CNNs [58, 81]. In contrast, the *stand-alone* and *vanilla* ViT architecture remains the concentration of this study to strive for simplicity. We ask: *What is the full potential of ViT architecture in medical imaging tasks?* The answer, based on our study, is that vanilla ViT can achieve a similar or even better performance than state-of-the-art CNNs if equipped with (I) a large-scale pre-training on unlabeled medical data and (II) strong pre-training and fine-tuning recipes, customized by unique characteristics of medical images.

The pre-training of CNNs has been widely investigated in the medical domain [49], resulting in several publicly available Foundation models [113, 17, 99]. Numerous pre-training methods can enable CNNs to learn representation from unlabeled images, including contrastive learning [84], predictive learning [115], restorative learning [15], and their combination [38, 39]. At the time this paper is written, however, neither contrastive nor predictive pre-training is mature for vanilla ViT architectures yet. The most popular pre-training scheme for ViTs is called Masked Autoencoders (MAE) [45]. Its task is to mask random patches of the input image and reconstruct the missing pixels. We adopt MAE in this paper because of its great scalability, computational efficiency, and compelling performance in many vision tasks.

This paper customizes the recipe of pre-training and fine-tuning MAE for the medical domain and verifies its effectiveness on three chest X-ray datasets. We have also made the pre-training and fine-tuning code publicly available and released ViT-Small and ViT-Base that are pre-trained on 510K X-ray images as well as the pre-trained CNNs. The pre-trained ViT encoder can be fine-tuned to improve classification tasks (validated in §5) and detection tasks (see Github). In summary, four contributions are made.

*Corresponding author: Zongwei Zhou (zzhou82@jh.edu)

1. The usefulness of ViT pre-trained on ImageNet (14M data & labels) and chest X-rays (0.3M data) is evaluated, underlining the opportunity of in-domain transfer learning and self-supervised learning (Table 1).

2. A strong pre-training recipe, consisting of more unlabeled data (266,340), a higher masking ratio (90%), and a modest random cropping scale (0.5∼1.0), is developed for MAE to learn image representation from chest X-rays efficiently (§4.2–4.4).

3. Three of the most important hyper-parameters are determined to fine-tune ViT in multi-label thorax disease classification: layer-wise LR decay, RandAug magnitude, and DropPath rate (Table 3).

4. This is among the first efforts to approach vanilla ViT's performance to the state-of-the-art CNNs on three predominant chest X-ray benchmarks, yielding mAUC of 82.3%, 89.2%, 99.3% on NIH ChestX-ray14, Stanford CheXpert, and COVIDx, respectively (§5.1–5.3).

By intention or non-intention, the empirical comparisons between old and new techniques (*e.g.* CNN vs. ViT) are often biased to the newer one [63, 4]. In this paper, we try our best not to over-sell or under-analyze the ViTs' potential in the medical domain. To provide a fair and comprehensive benchmark, the performance of CNNs is truly state-of-the-art in each dataset based on our extensive literature review.

2. Related Works

Preliminary. Radiography images possess unique characteristics compared with photographic images, resulting in considerable difficulties when switching computer vision advancements to medical imaging [111, 112, 58, 81]. Photographic images, particularly those on ImageNet [23], contain large, apparent objects in the center of the images, residing in varying backgrounds. Learning discriminative features (*e.g.* color, texture, and shape) primarily from the foreground objects is important in computer vision. In contrast, radiography images are generated from pre-defined imaging protocols, so the background exhibits anatomical consistency across images (see examples of the chest anatomy in Figure 4). Clinically relevant information is dispersed throughout the image, whereas the diseased region (as the foreground) often encloses much more local, subtle, and fine-grained variations than photographic images. As a result, the model must be able to extract both global and local features to identify various diseases from the normal anatomy. In the following sections, we describe the difference between computer vision and medical imaging in the choice of *model architectures* and *self-supervised methods*, followed by a review of current state-of-the-art solutions for multi-label thorax disease classification.

ViTs or CNNs for medical imaging? Transformers have gained prevalence in numerous AI applications (*e.g.* AlphaFold2 [53], Google Translate [8]). In computer vision, there is a heated debate between the adoption of ViTs and CNNs, in terms of performance [63, 110, 6, 97, 91, 25], robustness [4, 69, 105, 109], data requirement [26, 83, 89], computational efficiency [74]. This discussion has finally been converging to an agreement that ViTs could serve as alternatives to CNNs in a variety of tasks [56, 42]. ViT has substantial potential for radiography imaging tasks, but currently, the superior performance of ViT has not been translated to radiography imaging, where CNN is still the dominant architecture. (1) ViTs' performance has lagged behind that of SOTA CNNs [72, 88], in which we believe the poorly configured training recipe[1] is one of the major causes; (2) most existing studies report ViTs' performance on medical tasks without comparing with CNN under an *similar* experimental setting [70, 58]; and (3) multiple works focus on designing *hybrid* architectures by integrating benefits of ViTs and CNNs to claim the superior performance to CNNs [14, 107, 98, 86]. Conducting a fair comparison between ViTs and CNNs should take into account the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes. So far, there is no broad benchmark to fairly compare ViTs and CNNs in medical tasks, leaving us wondering whether we could trivially switch to ViTs in medical tasks. Unlike the aforementioned studies, our objective is to faithfully benchmark between ViTs and SOTA CNNs in radiography imaging tasks; to improve the recipe of existing ViTs with respect to data, model, optimization aspects; and to visualize how ViTs and CNNs interpret radiography images (§6).

Self-supervised methods in medical imaging. Self-supervised learning has shown enormous potential in medical imaging due to the sparsity of high-quality annotation [111]. Two major trends are based on contrastive and restorative pre-training. In computer vision, contrastive pre-training [19, 18, 34, 11] holds state-of-the-art performance, surpassing supervised ImageNet pre-training in some tasks; while in medical imaging, restorative pre-training [113, 86, 29] presently reaches a new height in performance. We attribute this popularity asymmetry to the marked difference between photographic and radiography images. Since radiography imaging protocols assess patients in a fairly consistent orientation, the generated images have great similarity across various patients [96, 39]. The inherent consistency eases the analysis of numerous critical problems but also causes a significant problem for contrastive pre-training. Contrastive pre-training (*e.g.* MoCo [46, 19]) treats each image as a distinct class and minimizes the simi-

[1] Isensee *et al.* [52] remark that most of the performance improvement comes for medical imaging is choosing the perfect data process, model training, and optimization strategy of the network (U-Net in their case).

larity of representations derived from different images. This concept, in theory, might not work properly for radiography imaging because the negative pairs appear too similar (empirically evidenced in our Table 1). In contrast, restorative pre-training is good at conserving fine-grained textures embedded in image context, so it has been widely adopted in medical pre-training. Restorative pre-training is formulated as the task of pixel-wise image reconstruction [2, 15, 114, 115, 16, 100]. Following this spirit, we take masked autoencoders (MAE) [45] as pre-training task for its simplicity, efficiency, scalability, and compelling performance. We are among the first to configure a strong recipe for both ViT pre-training and fine-tuning on enormous chest X-rays. Besides, we extend MAE to pre-train CNNs on the same scale medical data, establish the first direct benchmark between ViTs and SOTA CNNs on public radiography imaging datasets, and extract reusable insights to the medical vision community.

3. Method

Data. Data from three public X-ray datasets are used to pre-train ViTs (and CNNs as comparison): NIH ChestX-ray14 (75,312 X-rays), Stanford CheXpert (191,028 X-rays), and MIMIC-CXR (243,334 X-rays). All data are in the posteroanterior (PA) or anteroposterior (AP) view, and resized to 256×256 as input. All the X-rays are standardized by mean and standard deviation computed from ImageNet. We perform random resized cropping with a scale range of (0.5∼1.0) and random horizontal flipping. No other data augmentation is applied unless noted. The pre-training does not require any annotations shipped with the datasets.

Task. The ViT pre-training[2] is analogous to the image reconstruction task proposed in MAE [45]: to reconstruct the masked image patches from visible ones. Mean squared error is computed between the reconstructed and original images in the pixel space, averaged over masked patches [24]. An image is divided into regular non-overlapping patches as a sequence of embeddings. We randomly sample patches to be masked. The optimal masking ratio we observe is 90%, which substantially accelerates the pre-training by 2.5× compared with the original MAE [45][3] and enables us to scale up ViTs with greater model capability (Figure 1a).

Model. The vanilla ViT [26] is used as encoder and applied *only* on the visible image patches. This design reduces time and memory complexity [45]: a masking ratio of 90% (used in our paper) can reduce the encoder complexity to <1/10. The decoder is another ViT and only used during pre-training to reconstruct the masked patches. Therefore, the decoder is made to be more lightweight than the encoder (depth=2, width=512). As a result, although the decoder processes both visible and masked image patches, its complexity is much smaller than the encoder. Positional embeddings are added to visible and masked patches in this full set to preserve information about their original location in the image. We use ViT-S/16 and ViT-B/16 to denote ViT-Small and ViT-Base with a patch size of 16×16 for simplicity.

[2]The MAE-style pre-training for CNNs (for comparison in Table 1) is similar to the image in-painting task proposed in Models Genesis [113].

[3]It should have taken ∼16.7 GPU days for the original MAE to pre-train ViT-S/16 (the smallest ViT) on 510K X-rays.

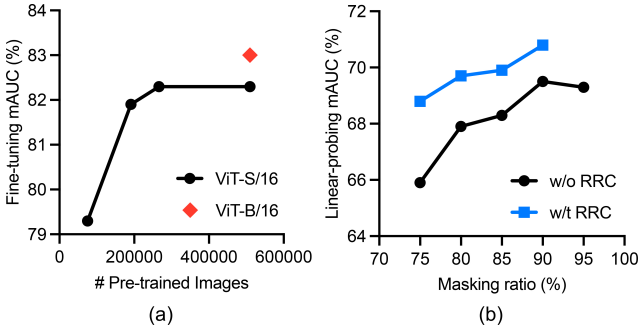

Figure 1: **The pre-training recipe.** (a) Using more images for pre-training can enhance the transferability of ViTs to some extent (§4.2). While ViT-S/16 (Params=22M) seems to be saturated at 266K images, ViT-B/16 (Param=86M) has the potential to scale up to more data. (b) MAE shows the optimal performance at a 90% masking ratio (§4.3). Besides, random resized crop (RRC) brings consistent performance gain to MAE pre-training (§4.4).

4. Pre-training: Recipe and Results

Implementation details. We adopt AdamW [65] optimizer with $\beta_1 = 0.9, \beta_1 = 0.95$ and set the weight decay to 0.05. Transformer blocks are initialized with xavier_uniform [32]. We set learning rate (lr) and batch size to $1.5e$-4 and 2048. lr is warmed up for 20 epochs [33] and scheduled with cosine annealing strategy [64]. The pre-training stage takes 800 epochs in total. Random resized crop and horizontal flip are used as data augmentation.

4.1. On the importance of in-domain transfer

Table 1 provides a comprehensive comparison on three sets of model initialization: random, ImageNet pre-training, and X-ray pre-training. Unlike CNNs, ViTs trained from scratch show very poor performance even with a strong training recipe and 2.7× larger number of training epochs than fine-tuning. On the contrary, ViTs achieve comparable and sometimes superior performance with the help of pre-training on large-scale datasets (*e.g.* ImageNet and X-rays). Specifically, after supervised pre-trained on ImageNet (following [91]), ViT-S/16 shows acceptable performance on the three datasets but is still distanced to the state-of-the-art CNN-based methods. In-domain transfer

Architecture	Pre-training Dataset	Method	Annotation	ChestX-ray14	CheXpert	COVIDx
CNN (DenseNet-121)	N/A	Random	0	80.4	87.8	93.0
	ImageNet (14M)	Categorization	14M	82.2	**89.4**	94.4
		MoCo v2† [19]	0	80.9	87.9	95.5
		BYOL† [34]	0	81.0	87.8	95.0
		SwAV† [11]	0	81.5	88.0	95.8
	X-rays (0.3M)	MoCo v2 [19]	0	80.6	88.7	94.0
		MAE††	0	81.2	88.7	**96.5**
ViT (ViT-S/16)	N/A	Random	0	67.9	77.9	87.3
	ImageNet (14M)	Categorization	14M	79.6	88.1	94.3
		MAE	0	78.6	88.3	88.8
	X-rays (0.3M)	MAE	0	**82.3**	89.2	95.2

†The pre-trained weights of ResNet-50 were taken from Ericsson *et al*. [27] (*DenseNet is not available for advanced self-supervised ImageNet pre-training*).
††MAE was developed for ViT (not directly applicable to CNN), so we implement it based on image in-painting [73, 113, 40, 41].

Table 1: **Pre-training on ImageNet vs. X-rays.** A direct comparison is performed between ViT and three groups of CNNs on three public datasets, considering the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes. The results suggest that ViT (I) consistently exceeds the CNNs that are pre-trained by state-of-the-art pre-training schemes on ImageNet, underlining the importance of in-domain transfer learning (§4.1); (II) surpasses the CNNs that are pre-trained by MAE and MoCo v2 on the same number of medical data (0.3M X-rays); (III) performs comparably (or even better) than state-of-the-art CNNs reported in the literature (detailed in Tables 4–6). Additionally, several important observations are obtained: (*i*) training from scratch takes longer epochs to converge than fine-tuning pre-trained weights (200 vs. 75 epochs); (*ii*) ViT shows inferior performance to CNN when training from scratch on X-ray images or fine-tuning from ImageNet; (*iii*) restorative pre-training (MAE) outperforms contrastive pre-training (MoCo v2) in radiography imaging.

	RandomResizedCrop	Crop Scale	mAUC
MAE	✗	N/A	65.9
	✓	(0.2, 1.0)	69.8
	✓	(0.5, 1.0)	**70.8**

Table 2: **A modest random cropping scale** is preferred for medical pre-training because the diseased regions (as foreground) is more local than photographic images, and pathological disorder could disperse over the entire X-rays [39] (rather than the center of the image).

seeks to reduce domain disparities between photographic and medical images [49]. In doing so, we bridge the domain gap and satisfy the ViTs/CNNs appetite for data by pre-training on 0.3M unlabeled chest X-rays. ViTs benefit more on the in-domain transfer (improved mAUC from 78.6 to 82.3 on ChestX-ray14), whereas ImageNet pre-trained CNNs remain high performance compared with in-domain pre-training (82.2 vs. 81.2 on ChestX-ray14).

4.2. Learning from 266,340 unlabeled X-rays

Training ViTs from scratch is harder than CNNs because ViTs lack inductive bias in modeling local visual representation and generally require more data to figure out the image content on their own [83, 63, 70]. As shown in Table 1, supervised pre-training on ImageNet brings performance gain from 67.9% to 79.6% for ViTs, and from 80.4% to 82.1% for CNNs on ChestX-ray14. We ask: *How many X-rays are needed for ViT pre-training?* Figure 1a shows that ViT-S/16 pre-trained on 75K, 191K, 266K, and 510K X-rays achieve a mAUC of 79.3%, 81.9%, 82.3%, and 82.3% on ChestX-ray14. The improvement from 75K to 266K is statistically significant (p-value=1.2e-127), but the performance gain is negligible from 266K to 510K—a bottleneck for ViT-S/16 (with 22M parameters). Although larger ViTs (*e.g*. ViT-B/16 with 86M parameters) can produce higher performance, considering computational cost and the fairness of the ViT vs. CNN, we end up pre-training ViT-S/16 using 266,340 unlabeled X-rays for benchmarking.

4.3. Masking out 90% X-ray content

The optimal masking ratio is related to the information redundancy in the data: BERT [55] uses a masking ratio of 15% for language and MAE [45] uses a ratio of 75% for images. Most recent study suggests that videos, due to its greater redundancy in the temporal dimension, can apply only 90% masking ratio for pre-training [28]. Given the great similarity in the chest anatomy, naturally, we hypothesize that *chest X-rays require even larger masking ratio for pre-training*. This is in line with the assumption that chest X-rays are more information-redundant than photographic images. We experimented with masking ratios ranging from 75% to 95%, incremented by intervals of 5%. Figure 1b indicates that 90% is the optimal masking ratio for MAE pre-training on chest X-rays. The larger masking ratio results in a more efficient pre-training, which is 2.5× faster than the original MAE. The efficient pre-training, in turn, enables us to scale up to larger ViT architectures and more diverse datasets.

Layer-wise LR decay	mAUC (%)
45	82.1
55	**82.3**
65	82

(a) **Layer-wise LR decay**. Learning rate decay in layer-wise needs to be tuned closely.

RandAug magnitude	mAUC (%)
4	82.0
6	**82.2**
8	82.1

(b) **RandAug magnitude**. A modest level of augmentation is preferred for fine-tuning.

DropPath rate	mAUC (%)
0.1	81.5
0.2	**82.3**
0.3	82.1

(c) **DropPath rate**. Fine-tuning on Chest X-ray images needs stronger regularization than natural images.

Table 3: **The fine-tuning recipe.** The ablation studies are conducted using ViT-S/16 on NIH Chest X-ray14. We report the 14-class average AUC (%). Except for (b) using Layer-wise LR decay 0.65, all the experiments adopt the optimal value for the hyper-parameters (Layer-wise LR decay 0.55, RandAug magnitude 6, and DropPath rate 0.2).

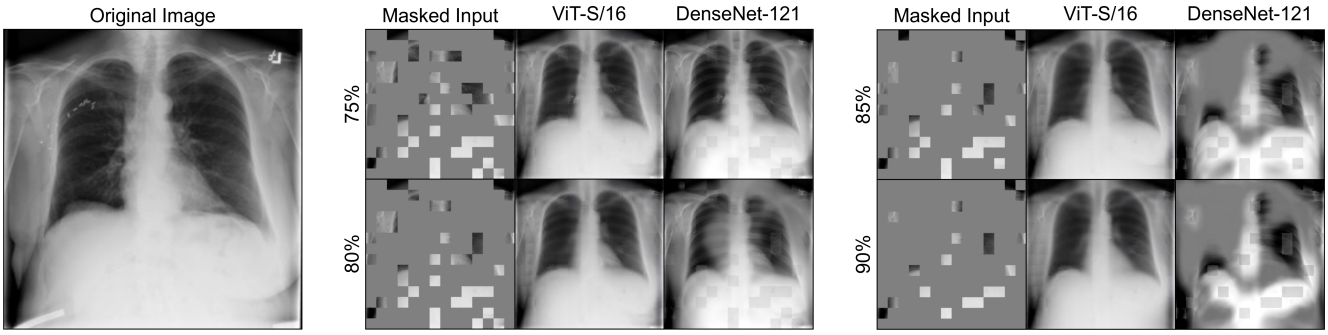

Figure 2: **Reconstruction of ChestX-ray14 *validation* images.** Pre-trained with a masking ratio of 75%, ViTs generalizes better than CNNs to the input images that are applied with higher masking ratios.

4.4. Cropping patches scaled of (0.5∼1.0)

As the spatial consistency of medical images is much higher than photographic images, there is a need to analyze the effectiveness of spatial data augmentation (*e.g.* random resized crop). Figure 1b suggests that RandomResizedCrop operation has consistent and noticeable benefit to MAE pre-training on chest X-ray imaging under different masking ratios. It enables ViTs to learn multiscale features from X-rays and to avoid the over-fitting problem due to the lack of training examples. More importantly, a relatively smaller cropping ratio than natural imaging is preferred (Table 2). Cropping patches scaled of (0.5∼1.0) yields 1.0% higher mAUC than those of (0.2∼1.0) (as suggested in [45]). It is intuitive that strong spatial augmentation is harmful since the informative lesions or organs could be cropped and biased and models will be learned with noisy annotations.

4.5. Quality assessment of image reconstruction

We assess reconstruction quality using validation images in Figure 2 for both ViT and CNN. The models are pre-trained on ChestX-ray14 and evaluated on the inputs with varying masking ratios, spanning from 75% to 90%, incremented by intervals of 5%. Both ViTs and CNNs can predict the overall anatomical structures in X-rays, but fail to reconstruct detailed texture such as shoulder bones. This is expected because the ViTs/CNNs only see 10% of the input image and attempt to reconstruct the rest 90% during the training—it is difficult even for expert radiologists. There is no clear evidence showing that the reconstruction capability is positively correlated to the transfer learning performance. On the contrary, the original autoencoders [48] (with a masking ratio of 0%) can certainly reconstruct images better than *masked* autoencoders, but their resulting representation is not as effective as the masked counterparts [113]. Moreover, studies in both CNNs [87, 113] and ViTs [45, 101] indicate that alternative loss functions (*e.g.* $l1$, smooth-$l1$, SSIM, and adversarial losses) for reconstruction would not contribute to the transfer learning performance. Therefore, we used $l2$ loss as default. Finally, we should remark that our ultimate goal is not the task of image reconstruction *per se*. While reconstructing patches is advocated and investigated as a pre-training scheme for ViTs/CNNs, the usefulness of the learned representation must be assessed objectively based on its generalizability and transferability to various downstream tasks (presented in §5).

5. Fine-tuning: Recipe and Results

Fine-tuning. The optimizer and lr scheduler are the same as pre-training. The choices of layer-wise LR decay, RandAug [22] magnitude[4], and DropPath [50] rate are crucial

[4]No improvement is obtained by more aggressive augmentation strategies (*i.e.* mixup [106] and cutmix [104]) since they could produce noisy

Method		Architecture	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	mAUC (%)
Allaouzi *et al*. [3]			72.0	**88.0**	77.0	87.0	90.0	82.8
Irvin *et al*. [51]			81.8	82.8	93.8	93.4	92.8	88.9
Seyyedkalantari *et al*. [80]			81.2	83.0	90.0	88.3	**93.8**	87.3
Pham *et al*. [75]		DN121	82.5	85.5	93.7	93.0	92.3	**89.4**
Hosseinzadeh *et al*. [49]			-	-	-	-	-	87.1
Haghighi *et al*. [39]			-	-	-	-	-	87.6
Kang *et al*. [54]			82.1	85.9	**94.4**	89.2	93.6	89.0
Ours	MoCo v2	DN121	78.5	77.9	92.5	92.8	92.7	88.7
	MAE	DN121	81.5	77.6	89.4	92.3	92.0	88.7
	MAE	ViT-S/16	**83.5**	81.8	93.5	**94.0**	93.2	89.2
	MAE	ViT-B/16	82.7	83.5	92.5	93.8	94.1	89.3

Table 4: **CheXpert benchmark.** ViT achieves comparable performance to the state-of-the-art CNNs on CheXpert (*official val*) over all five thorax diseases and the best o "Atelectasis" and "Edema" diseases.

Method			Input Resolution	# Params (M)	MACs (G)	Accuracy	COVID-19 Sensitivity
COVIDNet-CXR-3			480×480	29	29.1	98.3	97.5
COVIDNet-CXR-2				9	5.6	96.3	95.5
COVIDNet-CXR4-A				40	23.6	94.3	95.0
COVIDNet-CXR4-B				12	7.5	93.7	93.0
COVIDNet-CXR4-C				9	5.6	93.3	96.0
COVIDNet-CXR3-A				40	23.6	93.3	94.0
COVIDNet-CXR3-B				12	7.5	93.3	91.0
COVIDNet-CXR3-C				9	5.6	92.3	95.0
Ours	MoCo v2	DN121	448×448	7	11.6	96.0	96.5
	MAE	DN121		7	11.6	96.3	98.0
	MAE	ViT-S/16		22	16.9	95.3	95.0
	MAE	ViT-B/16	448×448	86	67.2	97.3	98.0
COVIDNet-CXR Small			224×224	117	2.3	92.6	87.1†
COVIDNet-CXR Large				127	3.6	94.4	96.8†
Ours	MoCo v2	DN121	224×224	7	2.9	94.0	94.5
	MAE	DN121		7	2.9	**96.5**	97.0
	MAE	ViT-S/16		22	4.2	95.2	94.5
	MAE	ViT-B/16	224×224	86	16.9	95.3	95.5

†The results are evaluated on 31 images; otherwise, the results are evaluated on the latest official testing set (400 images).

Table 5: **COVIDx benchmark.** ViTs show comparable performance to state-of-the-art CNNs on COVIDx (*official val*).

to fine-tune the pre-trained ViTs. The optimal settings are given by extensive studies in Table 3 and we reuse them for all three radiography imaging tasks in §5.3–5.2. Models are fine-tuned with 75 epochs on all three datasets.

Linear-probing. LARS [103] optimizer is used with *momentum*=0.9. We set learning rate (lr) and batch size to 0.1 and 16,384. lr is warmed up [33] for 10 epochs and scheduled with cosine annealing strategy [64]. The ViT is trained with 100 epochs. Linear-probing is used in Figure 1b.

5.1. Stanford CheXpert

Experimental setup. CheXpert is a large scale dataset containing 191,028 frontal-view chest X-rays. 14 diseases in radiology reports exist in the dataset and five common diseases are for benchmarking. We resized the images into 224×224 and the test is done on the official validation set. Mean Area Under the Curve (AUC) on five classes is reported for comparison.

labels by removing or overlapping thorax diseases in X-rays.

Results and analysis. As shown in Table 4, vanilla ViT-S achieves 89.2% mAUC which is very competitive to the best performance of 89.4%. Moreover, ViT-S yields the best performance on diseases of Atel (83.5%) and Edem (94.0%).

5.2. COVIDx

Experimental setup. COVIDx (version 9A) provides over 30,000 images containing 16,490 positive COVID-19 images. The dataset is annotated with 4 different classes for the training set of 30,130 images while the testing set only has 400 images of 3 classes. To ensure a fair comparison with previous methods, Accuracy and COVID-19 sensitivity on the testing set (3 classes) are reported.

Results and analysis. We compare our vanilla ViT-S model with the state-of-the-art models provided on the official github repository[5]. Our method beats the two other models when input resolution is 224×224 while achieving a very high accuracy of 95.2% and COVID-19 sensitivity of

[5] github.com/lindawangg/COVID-Net/blob/master/docs/models.md

Method		Architecture	Pre-training	mAUC
Wang *et al*. [95]		RN50	ImageNet (14M)	74.5
Yao *et al*. [102]		RN&DN		76.1
Li *et al*. [59]		RN50		75.5
Tang *et al*. [85]		RN50		80.3
Guendel *et al*. [36]		DN121		80.7
Guan *et al*. [35]		DN121		81.6
Wang *et al*. [93]		R152		78.8
Ma *et al*. [68]		R101		79.4
Baltruschat *et al*. [5]		RN50		80.6
Seyyed *et al*. [80]		DN121		81.2
Ma *et al*. [66]		DN121(×2)		81.7
Hermoza *et al*. [47]		DN121		82.1
Kim *et al*. [57]		DN121		82.2
Haghighi *et al*. [39]		DN121		81.7
Liu *et al*. [61]		DN121		81.8
Taslimi *et al*. [88]		SwinT		81.0
Ours	MoCo v2	DN121	X-rays (0.3M)	80.6
	MAE	DN121		81.2
	MAE	ViT-S/16		**82.3**
	MAE	ViT-B/16	X-rays (0.5M)	83.0

Table 6: **ChestX-ray14 benchmark.** ViT-S/16 achieves comparable performance to previous state-of-the-art CNN-based and Transformer-based methods on ChestX-ray14 (*official* split) reported in the literature. With the same pre-training scheme (MAE) on 0.3M X-rays, ViT significantly outperforms its CNN counterparts. In addition, ViT-B/16, pre-trained on 0.5M X-rays, hits a new record of 83.0 mAUC. RN, DN, and SwinT denote ResNet, DenseNet, and Swin Transformer.

94.5%. ViT-S shows a great balance between the model size, computation cost, and the performance.

5.3. NIH ChestX-ray14

Experimental setup. ChestX-ray14 has 112,120 frontal-view X-rays of 30,805 unique patients with the text-mined fourteen disease labels (where each image can have multiple labels). We follow the official data split which assigns 75,312 images for training and 25,596 images for testing. We resize the original images from size 1024×1024 into 224×224. Mean AUC on 14 classes is reported and 17 most popular and compelling baseline methods are compared.

Results and analysis. Table 6 provides a systematic comparison with state-of-the-art CNN and Transformer models on NIH ChestX-ray14 over years. The previous best CNN performance was obtained by DenseNet-121 [94] with a mean AUC of 82.6%. The previous best performance of Transformers was 81.0% [88], which was distanced to CNN's performance. Our vanilla ViT-S shows a very competitive result of 82.3% mean AUC over 14 diseases with the best classification performance on 6 out of 14 thorax diseases. It is worth noting that the research community took four years to improve the AUC score from 74.5 to 82.2 for CNN-type architectures, largely due to the difficulty of the training recipe.

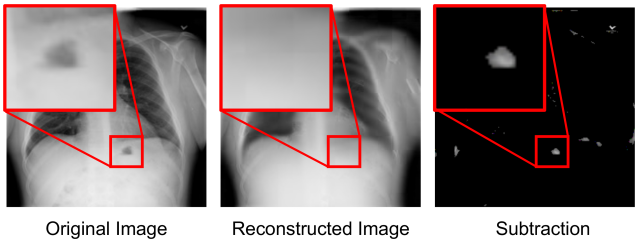

Figure 3: **MAE could reveal anomalies.** We input a chest X-ray (with anomalies) to the trained MAE and plot the difference map of the reconstructed output and original X-ray. Interestingly, we observe that the MAE happens to "heal" those anomalies by replacing them with normal patterns.

6. Discussion

Can MAE detect anomalies from an image? Anomalies are something that appear differently from the normal X-rays—can be diseases, medical devices, and clinical notations (*e.g*. arrows, numbers, letters). Since MAE is trained using the original X-rays as ground truth and the majority pixels in an X-ray are normal, the MAE should be able to overfit the normal anatomical patterns. Now, *if an anomaly is masked out, can MAE reconstruct a normal pattern?* If so, by subtracting the reconstructed output and the original X-ray, the anomaly can be detected and localized. A similar point has been discussed in Zhou *et al*. [113], but was illustrated using CNNs. Specifically, we input an original X-ray to the trained MAE and plot the difference map between reconstructed output and the original image. As shown in Figure 3, MAE happens to "heal" those anomalies and reconstruct with normal patterns. This behavior can be thought of as an attempt to detect and localize anomalies. More importantly, unlike weakly-supervised detection strategies [108, 7, 9, 82, 96], neither image-level nor pixel-level annotation is required for this approach, making it an attractive and challenging direction to explore [76, 90].

Weakly-supervised disease localization by ViT and CNN. With the help of Grad-CAM[6] [31], we are able to check which part of the X-ray image is responsible for the model prediction (the diseased region). We use the last dense-block (4th) of DenseNet-121 and the LayerNorm layer in the last transformer block (12th) of ViT-S/16 as the "target layers" for Grad-CAM. The experiments are done in a small subset of ChestX-ray14, which offers 787 cases with bounding-box of a total of eight thorax diseases. The final predicted bounding-box of the diseased region is generated with the thresholded Grad-CAM heatmap, largest connected component, and box regression. The results are evaluated by IoU between ground truth bounding box and the bounding box of the largest connected component in

[6] github.com/jacobgil/pytorch-grad-cam

Disease	Size (# of px)	DenseNet-121 AP_{25}	DenseNet-121 AP_{50}	ViT-S/16 AP_{25}	ViT-S/16 AP_{50}
Nodule	224	0.0	0.0	9.2	3.9
Mass	756	25.4	1.6	27.0	11.1
Atelectasis	924	10.1	2.0	31.5	8.1
Pneumothorax	1899	11.6	2.3	4.7	0.0
Infiltrate	2754	32.9	12.7	11.4	1.3
Effusion	2925	24.5	2.9	8.8	1.0
Pneumonia	2944	32.0	6.2	27.8	9.3
Cardiomegaly	8670	89.6	53.3	16.3	3.0
All	2300	31.0	12.3	18.0	4.7

Table 7: **Weakly-supervised disease localization.** We report the average precision (AP) on 25% and 50% IoUs. The IoU is calculated between the ground truth bounding box and bounding box of the largest connected component in the Grad-CAM heatmap. We also present the statistics of disease sizes, measured by the number of pixels within the bounding box, showing that CNN can detect large diseases (*e.g.* Cardiomegaly, Pneumonia) better than ViT, while ViT can capture smaller diseases (*e.g.* nodule).

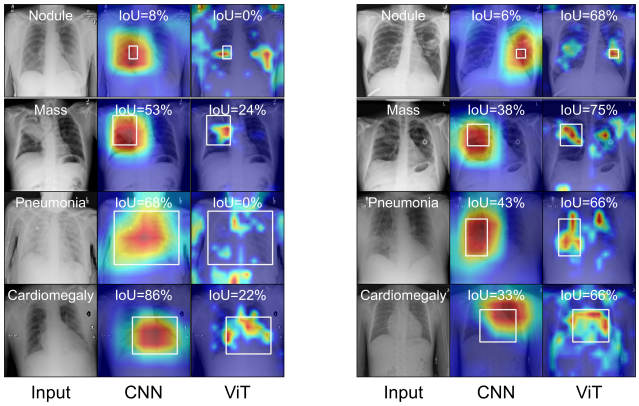

Figure 4: **Grad-CAM of CNN and ViT.** [Better viewed on-line, in color, and zoomed in for details] ChestX-ray14 provides bounding boxes for some of the thorax diseases, shown in white boxes. Left and right panels display successful cases predicted by CNN and ViT, respectively.

the attention response. We then compute Average Precision (AP) as the detection metric [60]. Precision is defined as $tp/(tp+fp)$, where tp and fp denote the number of true positives and false positives, respectively. AP_{25} considers cases with IoU>25% as true positives and AP_{50} with IoU>50%. Table 7 shows the detection results (including disease-wise results and all diseases). We observe that the CNN provides better localization of diseases in a larger size (*e.g.* Cardiomegaly and Pneumonia) while ViT is robust to diseases in a smaller size (*e.g.* Nodule). Although the classification performance of CNN and ViT is comparable (82.1% vs. 82.3% AUC), CNN significantly exceeds the localization ability to ViT, and their attention maps generated by Grad-CAM behave differently. Figure 4 provides examples of GradCAM of CNN and ViT. Attentions in CNN are relatively larger and more concentrated than those in ViT. This observation is consistent with those in Chefer *et al.* [13]. This study suggests that class activation maps are more suitable for visualizing the explainability of CNN-type models. In the future, other than class activation maps, we will seek to explore the explainability for Vision Transformers in multi-label classification tasks, with the help of self-attention derived from the Transformer architectures [12, 79, 1, 13].

7. Conclusion and Future Work

This paper has unleashed the potential of stand-alone, vanilla ViT by devising strong pre-training and fine-tuning recipes. We overcome several technical barriers and bring reusable insights for the medical vision community. Specifically, we (*i*) improve computational efficiency; (*ii*) customize data augmentation; (*iii*) explore larger data scale; and (*iv*) optimize learning parameters. As a result, the vanilla ViT achieves a comparable (sometimes better) performance to state-of-the-art CNNs. Code and pre-trained models are available.

This paper has also presented an up-to-date benchmark on three predominant chest X-ray datasets. Taking into account the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes, we have performed a fair and comprehensive comparison between vanilla ViT and (*i*) state-of-the-art CNNs reported in the literature, (*ii*) CNNs that are pre-trained by advanced pre-training schemes on ImageNet, (*iii*) CNNs that are pre-trained on the same number of medical data. We hope this study can direct future research on the application of Transformers to a larger variety of medical imaging tasks.

As future work, we will consider three extensions to our current study. First, assembling more publicly available X-ray datasets for pre-training (which account for a total of ∼1M images [10]). Scaling up the data is perhaps the most straightforward way to enhance larger ViTs (*e.g.* ViT-Large, ViT/Huge) in terms of performance and generalizability based on Figure 1a. Second, extending ViT to its 3D form for higher dimensional medical modalities (*e.g.* CT, MRI), which is expected to take a considerable computational resource [44] and larger data for pre-training [86], therefore requiring a more efficient method. Third, exploiting paired information of radiology reports and image data for pre-training. We acknowledge the unique ability of Transformers in processing multi-modality data [78, 20].

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Y. Zhang for providing data loader of the COVIDx dataset; A. Delaney for improving the writing of this paper.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 8

[2] 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. 3

[3] 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. 6

[4] 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. 2

[5] 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. 7

[6] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[7] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018. 7

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[9] Jinzheng Cai, Le Lu, Adam P Harrison, Xiaoshuang Shi, Pingjun Chen, and Lin Yang. Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–598. Springer, 2018. 7

[10] 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. 8

[11] 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. 2, 4

[12] 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 International Conference on Computer Vision (ICCV)*, 2021. 8

[13] 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. 8

[14] 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*, 2021. 2

[15] 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. 1, 3

[16] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels. *Advances in Neural Information Processing Systems*, 2020. 3

[17] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 1

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 4

[20] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 8

[21] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019. 1

[22] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. In *NeurIPS*, 2020. 5

[23] 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. 2

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3

[25] 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. 1, 2

[26] 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. 1, 2, 3

[27] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021. 4

[28] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 4

[29] 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. 2

[30] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021. 1

[31] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 7

[32] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010. 3

[33] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 3, 6

[34] 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. 2, 4

[35] Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018. 7

[36] 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. 7

[37] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[38] Zuwei Guo, Nahid UI Islam, Michael B Gotway, and Jianming Liang. Discriminative, restorative, and adversarial learning: Stepwise incremental pretraining. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 66–76. Springer, 2022. 1

[39] 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. 1, 2, 4, 6, 7

[40] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–147. Springer, 2020. 4

[41] 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*, 2021. 4

[42] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[43] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021. 1

[44] 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. 8

[45] 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. 1, 3, 4, 5

[46] 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. 2

[47] 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. 7

[48] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 5

[49] 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. 1, 4, 6

[50] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In

European conference on computer vision, pages 646–661. Springer, 2016. 5

[51] 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. 6

[52] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 2

[53] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 2

[54] 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. 6

[55] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 4

[56] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 2

[57] 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. 7

[58] 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. 1, 2

[59] 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. 7

[60] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 8

[61] 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, 2022. 7

[62] 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. 1

[63] 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, 2022. 2, 4

[64] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 3, 6

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 3

[66] Congbo Ma, Hu Wang, and Steven C. H. Hoi. Multi-label thoracic disease image classification with cross-attention networks, 2020. 7

[67] 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. 1

[68] 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. 7

[69] 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. 2

[70] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021. 1, 2, 4

[71] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 1

[72] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. *arXiv preprint arXiv:2103.07055*, 2021. 2

[73] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 4

[74] 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. 2

[75] 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. 6

[76] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 7

[77] Eva Portelance, Michael C Frank, Dan Jurafsky, Alessandro Sordoni, and Romain Laroche. The emergence of the shape bias results from communicative efficiency. *arXiv preprint arXiv:2109.06232*, 2021. 1

[78] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 8

[79] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 8

[80] 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. 6, 7

[81] 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. 1, 2

[82] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019. 7

[83] 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. 1, 2, 4

[84] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Caid: Context-aware instance discrimination for self-supervised learning in medical imaging. *arXiv preprint arXiv:2204.07344*, 2022. 1

[85] 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. 7

[86] 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. 2, 8

[87] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 238–248. Springer, 2020. 5

[88] 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. 2, 7

[89] 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. 1, 2

[90] Yu Tian, Guansong Pang, Yuyuan Liu, Chong Wang, Yuanhong Chen, Fengbei Liu, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. *arXiv preprint arXiv:2203.11725*, 2022. 7

[91] 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. 1, 2, 3

[92] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 1

[93] Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. Thoracnet: 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. 7

[94] Hongyu Wang, Shanshan Wang, Zibo Qin, Yanning Zhang, Ruijiang Li, and Yong Xia. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, 67:101846, 2021. 7

[95] 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. 7

[96] 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. 2, 7

[97] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Transforming radiograph imaging with transformers: Com-

paring vision transformers with convolutional neural networks in multi-label thorax disease classification. In *Radiological Society of North America (RSNA)*, 2022. 2

[98] 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. 2

[99] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unified 2d and 3d pre-training for medical image classification and segmentation. *arXiv preprint arXiv:2112.09356*, 2021. 1

[100] 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. 3

[101] 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*, pages 9653–9663, 2022. 5

[102] 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. 7

[103] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 6

[104] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 5

[105] 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. 2

[106] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 5

[107] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021. 2

[108] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 7

[109] 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, 2022. 2

[110] 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*, 2022. 2

[111] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021. 2

[112] Zongwei Zhou, Michael Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health: The Role of AI*. Springer, 2022. 2

[113] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 1, 2, 3, 4, 5, 7

[114] 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, 2019. 3

[115] 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. 1, 3