ARTICLES
https://doi.org/10.1038/s42256-021-00425-9
nature machine intelligence

Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports

Hong-Yu Zhou[1,3], Xiaoyu Chen[2,3], Yinghao Zhang[2,3], Ruibang Luo[1], Liansheng Wang[2] ✉ and Yizhou Yu[1] ✉

Pre-training lays the foundation for recent successes in radiograph analysis supported by deep learning. It learns transferable image representations by conducting large-scale fully- or self-supervised learning on a source domain; however, supervised pre-training requires a complex and labour-intensive two-stage human-assisted annotation process, whereas self-supervised learning cannot compete with the supervised paradigm. To tackle these issues, we propose a cross-supervised methodology called reviewing free-text reports for supervision (REFERS), which acquires free supervision signals from the original radiology reports accompanying the radiographs. The proposed approach employs a vision transformer and is designed to learn joint representations from multiple views within every patient study. REFERS outperforms its transfer learning and self-supervised learning counterparts on four well-known X-ray datasets under extremely limited supervision. Moreover, REFERS even surpasses methods based on a source domain of radiographs with human-assisted structured labels; it therefore has the potential to replace canonical pre-training methodologies.

Medical image analysis has achieved tremendous progress in recent years thanks to the development of deep convolutional neural networks (DCNNs)[1–5]. At the core of DCNNs is visual representation learning[6], where pre-training has been widely adopted and become the most dominant approach to obtain transferable representations. Typically, a large-scale dataset—also called the source domain—is first used for model pre-training. Transferable representations from the pre-trained model are further fine-tuned on other smaller downstream datasets, called target domains.

As one of the most general forms of medical images, radiographs have a great potential to be used in widespread applications[7–9]. To achieve (or at least approximate) radiologist-level diagnosis performance in these applications, it is common to transfer learned representations from natural images to radiographs[10,11], and ImageNet-based[12] pre-training is most widely adopted in this context. On the other hand, self-supervised learning[13–16] has attracted much attention in the community as it is capable of learning transferable radiograph representations without any human annotations. Both methodologies have been proven to be effective in solving medical image analysis tasks, especially when the amount of labelled data in the target domain is quite limited. However, in the first approach, there is an inevitable problem, which is the existence of domain shifts between medical and natural images. For instance, it is possible to introduce harmful noises from natural images as radiographs have a different pixel intensity distribution. As for self-supervised learning, to the best of our knowledge, clear performance gaps still exist between radiograph representations learned through self-supervised and label-supervised pre-training. To avoid these problems, building large-scale annotated radiograph datasets for label-supervised pre-training becomes an essential and urgent issue in radiograph analysis.

Radiologists and computer scientists have recently managed to build medical datasets for label-supervised pre-training at the size of hundreds of thousands of images, such as ChestX-ray[11], MIMIC[17] and CheXpert[18]. To acquire accurate labels for radiographs, these datasets often rely on a two-stage human intervention process. A radiology report is first prepared by radiologists for every patient study as part of the clinical routine. In the second stage, human annotators extract and confirm structured labels from these reports using artificial rules and existing natural language processing (NLP) tools; however, there are two major limitations of this label extraction workflow. First, it is still complex and labour intensive; for example, human annotators have to define a list of alternate spellings, synonyms and abbreviations for every target label. Consequently, the final accuracy of extracted labels heavily depends on the quality of human assistance and various NLP tools. A small mistake in a single step or a single tool may give rise to disastrous annotation results. Second, those human-defined rules are often severely restricted to application-oriented tasks instead of general-purpose tasks. It is difficult for DCNNs to learn universal representations from such application-oriented tasks.

In this paper we propose reviewing free-text reports for supervision (REFERS) to directly learn radiograph representations from accompanying free-text radiology reports. We believe abstract and complex logic reasoning sentences in radiology reports provide sufficient information for learning well-transferable visual features. As shown in Fig. 1a, REFERS is realized using a set of transformers, in which the most important part is a radiograph transformer serving as the backbone. The main reason why we choose the transformer as the backbone in REFERS is that it not only exhibits the advantages of DCNNs, but also has been shown to be more effective[19] due to the self-attention mechanism[20]. Moreover, we have found that,

[1]Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong. [2]Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China. [3]These authors contributed equally: Hong-Yu Zhou, Xiaoyu Chen and Yinghao Zhang. ✉e-mail: lswang@xmu.edu.cn; yizhouy@acm.org

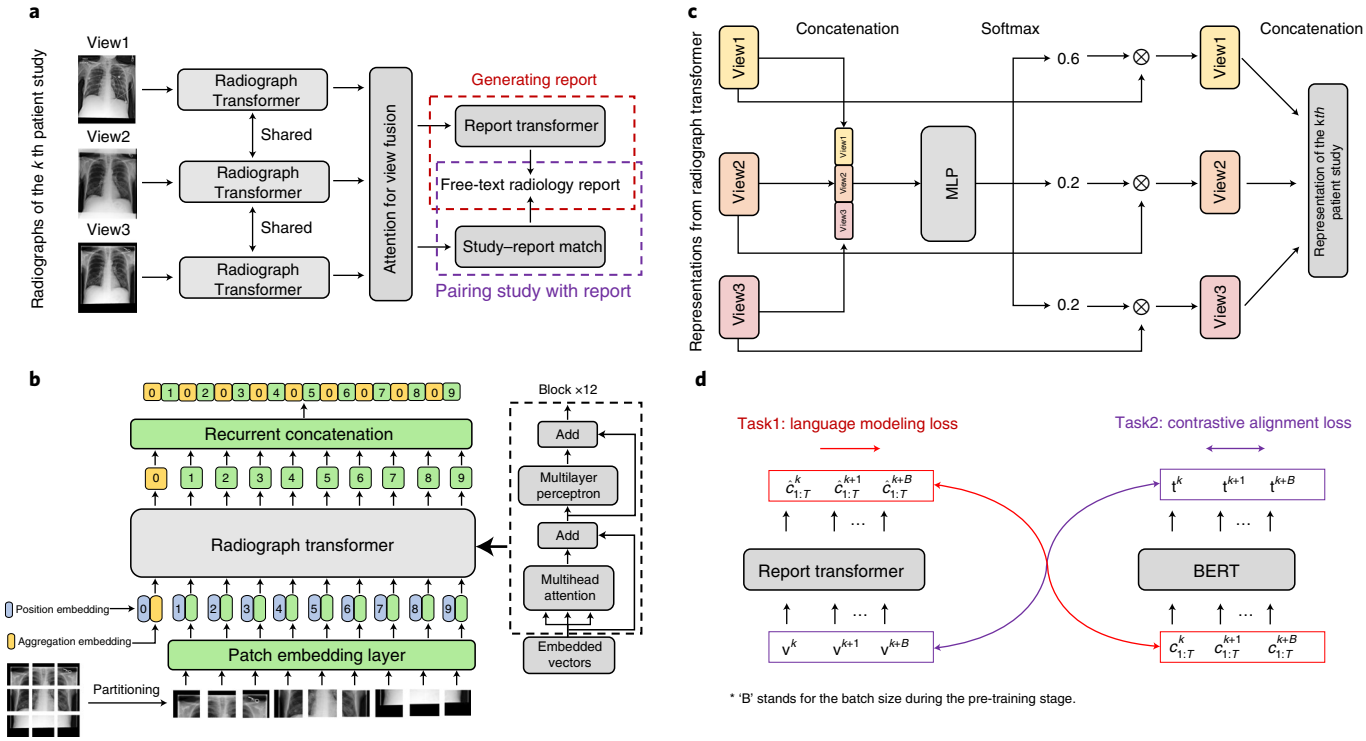

Fig. 1 | REFERS workflow. We forward radiographs of the *k*th patient study through the radiograph transformer, fuse representations of different views using an attention mechanism, and use report generation and study–report representation consistency reinforcement to exploit the information in radiology reports. **a**, An overview of the whole pipeline. **b**, The architecture of the radiograph transformer. **c**, Attention for view fusion is elaborated. MLP stands for a multi-layer perceptron. **d**, Two supervision tasks are shown, report generation and study–report representation consistency reinforcement; $\mathbf{v}^k$ and $\mathbf{t}^k$ denote the visual and textual features of the *k*th patient study, respectively, whereas $\hat{c}^k_{1:T}$ and $c^k_{1:T}$ are the token-level prediction and ground truth of the *k*th radiology report whose length is *T*, respectively.

in comparison to features generated from DCNNs, features from transformers are more compatible with textual tasks.

Unlike aforementioned representation learning methodologies, REFERS performs cross-supervised learning and does not need structured labels during the pre-training stage. Supervision signals are instead defined by automatically cross-checking the two different data modalities, radiographs and free-text reports. Considering that—in daily clinical routine—there is typically a free-text report associated with every patient study, which usually involves more than one radiograph. To fully utilize the study-level information in each report, we design a view fusion module based on an attention mechanism to process all radiographs in a patient study simultaneously and fuse the resulting multiple features. In this way, the learned representations are able to preserve both study- and image-level information. By contrast, only image-level information is addressed in traditional representation learning paradigms[11,13–16], which use a single image as input. On top of the view fusion module, we conduct two tasks (report generation and study–report representation consistency reinforcement) to extract study-level supervision signals from free-text reports. To carry out the first task, we apply a decoder called report transformer to the fused feature, with the goal to reproduce the radiology report associated with the study. For the second task, we apply our radiograph transformer and an NLP transformer to a study–report pair. These transformers produce a pair of feature representations for the patient study and radiology report, respectively. The consistency between such a pair of feature representations within every study–report pair is reinforced via a contrastive loss function. Some previous works[21,22] tried to learn joint text–image representations for single-domain medical image analysis tasks. Compared with them, REFERS focuses on learning well-transferable image features from study-level free-text reports on a large-scale source domain and fine-tuning them on one or more target domains.

On four well-known X-ray datasets, REFERS outperforms self-supervised learning and transfer learning on natural source images in producing more transferable representations, often bringing impressive improvements (greater than 5%) under limited supervision from target domains. This capability can be extremely important in real-world applications as medical data are scarce and their annotations are usually hard to acquire. More surprisingly, we found that REFERS clearly surpasses those methods that employ a source domain with a large collection of medical images with structured labels. In terms of specific abnormalities and diseases, REFERS is quite effective under extremely limited supervision (<1,000 annotated radiographs during fine-tuning). For instance, REFERS brings about 9% improvements on pneumothorax. Meanwhile, over 7% improvements are achieved on two common lung diseases (atelectasis and emphysema).

Results

All self-supervised learning and label-supervised pre-training (LSP) baselines, as well as REFERS, are first pre-trained on a source domain of medical images (that is, MIMIC-CXR-JPG[23]). Pre-trained models are then fine-tuned on each of four well-established datasets (target domains with labels), including NIH ChestX-ray[11], VinBigData Chest X-ray Abnormalities Detection[24], Shenzhen Tuberculosis[25] and COVID-19 Image Data Collection[26]. During the fine-tuning stage, we always perform fully supervised learning on the target domain, which only consists of radiographs with structured labels. Furthermore, we verify model performance by varying the

Table 1 | Comparison with self-supervised learning and transfer learning baselines

	NIH	NIH	NIH	VBD	VBD	VBD	SZ	C-T1	C-T2
Method	800 (1%)	8,000 (10%)	80,000 (100%)	100 (1%)	1,000 (10%)	10,000 (100%)	All	All	All
REFERS	**76.7**	**80.9**	**84.7**	**83.0**	**88.2**	**90.1**	**98.0**	**82.1**	**80.4**
Model Genesis	70.3	75.7	81.0	70.7	82.7	85.8	94.9	76.0	71.8
C2L	71.0	76.6	82.2	75.3	83.3	85.9	95.5	77.8	73.0
Context Restoration	67.8	73.9	78.7	67.9	82.4	83.8	92.7	74.6	69.8
TransVW	71.2	74.3	81.7	73.6	83.8	86.2	94.2	76.1	71.5
ImageNet-based pre-training	69.8	74.4	80.0	69.7	82.9	84.5	94.5	74.1	70.3
P-value	8.35×10^{-4}	8.72×10^{-4}	1.94×10^{-3}	8.72×10^{-5}	4.34×10^{-4}	9.33×10^{-4}	1.73×10^{-3}	5.88×10^{-4}	3.59×10^{-4}

NIH, VBD and SZ represent NIH ChestX-ray, VinBigData Chest X-ray Abnormalities Detection and Shenzhen Tuberculosis datasets, respectively. C-T1 and C-T2 denote the two tasks in COVID-19 Image Data Collection, where one task is to distinguish COVID-19 from non-COVID-19 cases (C-T1) and the other task is to separate viral pneumonia cases from bacterial ones (C-T2). Note that for fairness, all baselines use the same transformer-based backbone as the radiograph transformer of REFERS (that is, a vision transformer (ViT)-like architecture plus the recurrent concatenation operator). Each *P*-value is calculated between the results from REFERS and the best-performing baseline. The evaluation metric is the AUC. The best results are bold.

Table 2 | Comparison with methods using human-assisted structured labels

	NIH	NIH	NIH	VBD	VBD	VBD	SZ	C-T1	C-T2
Method	800 (1%)	8,000 (10%)	80,000 (100%)	100 (1%)	1,000 (10%)	10,000 (100%)	All	All	All
REFERS	**76.7**	**80.9**	**84.7**	**83.0**	**88.2**	**90.1**	**98.0**	**82.1**	**80.4**
LSP (Transformer)	74.2	78.2	82.1	78.5	85.8	87.6	96.4	80.2	76.6
LSP (ConvNet)	65.8	74.5	81.9	76.0	85.2	87.2	96.7	80.1	76.2
P-value	3.25×10^{-3}	2.89×10^{-3}	5.23×10^{-3}	3.56×10^{-4}	8.69×10^{-4}	1.05×10^{-3}	9.65×10^{-3}	7.61×10^{-3}	1.47×10^{-3}

Note that for fairness, both LSP (Transformer) and REFERS share the same transformer-based backbone (that is, the ViT-like architecture plus the recurrent concatenation operator). Each *P*-value is calculated between the results from REFERS and LSP (Transformer). AUC is the evaluation metric. The best results are in bold.

percentage of actually used training images (sampled from the predefined whole training set) in the target domain; this percentage is called the label ratio. When the label ratio is 100%, we use the whole training set in the target domain for fine-tuning.

NIH ChestX-ray. Table 1 and Extended Data Figs. 1 and 2 present experimental results from REFERS and other approaches under different label ratios. As shown in Table 1 and Extended Data Fig. 1, our approach outperforms self-supervised baselines and transfer learning on natural source images substantially. To be specific, REFERS achieves the highest area under the receiver operating characteristic curve (AUC) on all 14 classes using different amounts of training data during the fine-tuning stage. Moreover, REFERS exhibits the greatest performance improvements with respect to these baselines when only 800 training images (1% label ratio) in the target domain are utilized. For example, REFERS surpasses the widely adopted ImageNet-based pre-training[11] by about 7% on average. REFERS even gives quite competitive results when compared with LSP. Table 2 shows that the average performance of REFERS actually surpasses LSP, consistently maintaining an advantage of at least 2%. Compared with self-supervised baselines[13–16] and ImageNet-based pre-training[11], REFERS achieves the largest improvements on emphysema (7%) and cardiomegaly (>10%), especially under limited supervision. When compared with LSP, our method achieves consistent improvements on mass (>4%).

VinBigData Chest X-ray Abnormalities Detection. REFERS exhibits a greater advantage on this target domain dataset than it does on NIH ChestX-ray, as VinBigData comprises a much smaller number of annotated radiographs (about one-eighth of the NIH dataset). This again demonstrates REFERS's ability to manage with limited supervision. REFERS consistently maintains large advantages over other methods under different conditions (see Tables 1 and 2, and Extended Data Figs. 3 and 4). For instance, when we only have 105 annotated radiographs (1% label ratio) as fine-tuning data, REFERS surpasses C2L[16]—the best-performing self-supervised method—by over 7% in AUC. The performance of REFERS once again surpasses LSP with human-assisted structured labels even when all-annotated training data (100% label ratio) in the target domain are used. When we check specific abnormalities and diseases, we found REFERS consistently improves the diagnosis of atelectasis, lung opacity and pneumothorax in comparison to LSP.

COVID-19 and Shenzhen Tuberculosis image collections. Both datasets serve as target domains and comprise a small number of labelled images (fewer than 1,000 X-rays), which are employed to test the transferability of the representation learned on the source domain. We adopted these two datasets as target domains because the few training images in such small target domains are not capable of training powerful models themselves; thus, the performance of the trained models is more dependent on the quality of the learned representation. In Table 1, although separating tuberculosis from normal cases is not a hard task, our method still achieves 2.5% improvements over C2L[16] in AUC. When looking at the COVID-19 Image Data Collection dataset, which includes two harder tasks, we can find that the relative performance improvements over self-supervised baselines[13–16] and transfer learning on natural source images[11] become quite clear. For instance, on the viral versus bacterial task, REFERS outperforms C2L[16] by 7% in AUC, demonstrating the effectiveness of REFERS in helping achieve better performance over small-scale target datasets. Even if we compare REFERS against LSP, the performance advantage is still maintained at more than 1%.

Discussion

REFERS outperforms self-supervised learning and transfer learning on natural source images by substantial margins. This is the most prominent observation obtained from our experimental results, which holds on different datasets and with different amounts of annotated training data during fine-tuning. Among self-supervised baselines[13–16], C2L[16] and TransVW[15] are the two best-performing methods. REFERS outperforms C2L and TransVW by at least 4% when very limited annotated training data (at most 10% label ratio) from the NIH ChestX-ray and VinBigData datasets are used. Somewhat interestingly, as the label ratio increases, ImageNet-based pre-training[11] gradually narrows the gap on self-supervised learning. Nonetheless, REFERS still surpasses it by a large margin (at least 4%). Similar results can also be observed on the Shenzhen Tuberculosis and COVID-19 Image Data Collection datasets. As REFERS employs a cross-supervised learning manner, it does not require structured labels as conventional fully supervised learning approaches. As radiographs and radiology reports are readily available medical data, we believe our approach is as practical as self-supervised learning methodologies in real-world scenarios.

REFERS consistently surpasses label-supervised pre-training with human-assisted structured labels. This is another clear observation obtained from our experimental results. Although our approach does not use any structured labels in the source domain, our pre-trained model exhibits clear advantages over all four target domain datasets. Specifically, REFERS outperforms the most competitive LSP method, LSP (Transformer), which is based on transformer- and human-assisted structured labels in the source domain. In particular, our method shows more advantages at small label ratios. For instance, when NIH ChestX-ray and VinBigData are used as target domain datasets, REFERS achieves about 2.5% improvements when the number of training images is smaller than 10,000. Similarly, REFERS consistently surpasses LSP by significant margins (P-values < 0.01) on the Shenzhen Tuberculosis and COVID-19 Image Data Collection datasets. It is worth mentioning that when a classification problem is difficult to solve and has limited supervision, REFERS becomes more advantageous and achieves impressive improvements. For example, on the viral versus bacterial task (the last column in Table 2), REFERS surpasses label-supervised pre-training methods based on two-stage human intervention by approximately 4%. These improvements demonstrate that raw radiology reports contain more useful information than human-assisted structured labels. In other words, the advantages exhibited by our approach on small-scale target domain training data can be attributed to the rich information carried by radiology reports in the source domain. Such information provides additional supervision to help learn transferable representations for radiographs, whereas the supervision signals from structured labels have less information. We believe that this is an important step towards directly using natural language descriptions as supervision signals for image representation learning. As an example, REFERS can be used to learn natural image representations from text descriptions at corresponding websites.

REFERS reduces the need of annotated data in target domains. Figure 2a,b presents the performance of our approach under various label ratios. On the NIH ChestX-ray dataset, REFERS needs 90% fewer annotated target domain data (10% label ratio) to deliver performances comparable with those of Model Genesis[14] and ImageNet-based pre-training[11]. Similarly, on VinBigData, our method only needs 10% annotated training data to achieve much better results than those of Model Genesis and ImageNet-based pre-training under 100% label ratio. This phenomenon shows the potential of REFERS in providing high-quality pre-trained representations for downstream fine-tuning tasks with limited annotations.

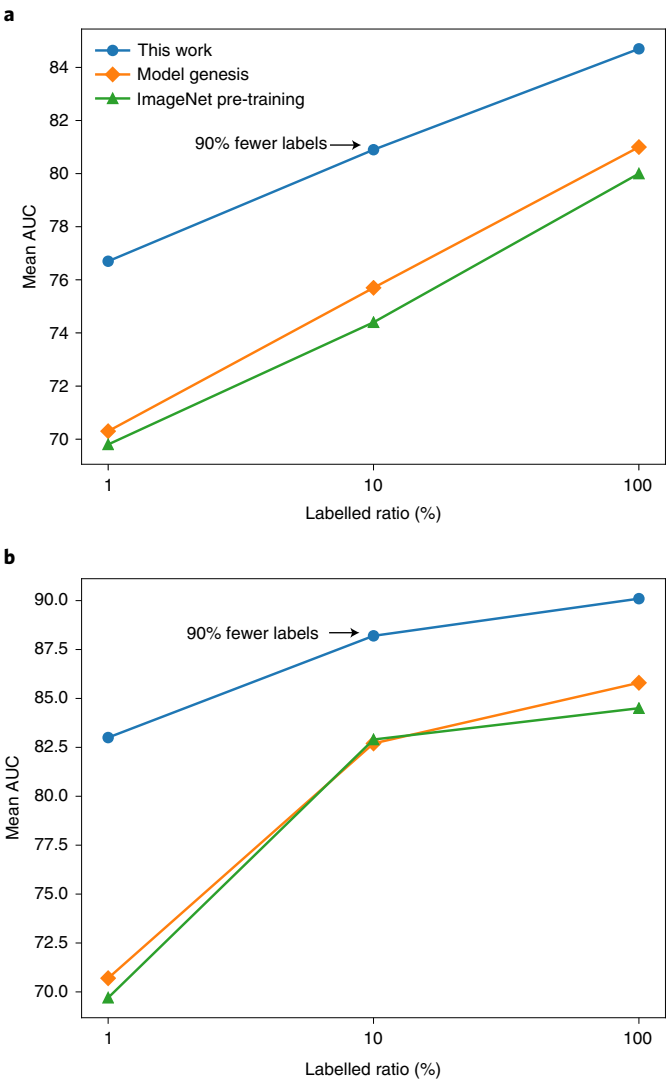

Fig. 2 | Performance of our approach under various label ratios. **a**,**b**, Performance obtained with different amounts of annotated training data in the target domain: NIH ChestX-ray (**a**) and VinBigData Chest X-ray Abnormalities Detection (**b**). We also denote the percentage of annotated training data in the target domain that REFERS needs to achieve comparable results with those of Model Genesis and ImageNet pre-training. Note that all three methods share the same transformer-based backbone.

Due to the difficulty to acquire reliable annotations for medical image analysis, the ability to achieve good performance with limited annotations means much to the community.

Improvements on specific abnormalities and diseases. In Extended Data Fig. 2, REFERS brings a 5% performance gains on emphysema and mass even when compared with LSP with limited supervision in the target domain (<10,000 training images). As both abnormalities have a dispersed spatial distribution in the lung area, the considerable improvements demonstrate that REFERS is able to handle elusive chest abnormalities in radiographs well. REFERS becomes more advantageous when the amount of supervision in the target domain becomes extremely limited (for example, when using 105 training images from VinBigData). For instance, REFERS outperforms LSP on atelectasis and pneumothorax by over 7% and 9%, respectively (Extended Data Fig. 4). Unlike emphysema, mass and atelectasis, pneumothorax maintains a concentrated spatial

Table 3 | An ablation study of REFERS by removing or replacing individual modules

Row	ViT	RecConcate	View Fusion	Task1	Task2	Viral versus bacterial
0	✓	✓	✓	✓	✓	**80.4**
1			✓	✓	✓	73.3
2	✓		✓	✓	✓	77.1
3	✓	✓		✓	✓	78.6
4	✓	✓				76.6
5	✓	✓	✓	✓		79.1
6	✓	✓	✓		✓	79.3

RecConcate stands for the recurrent concatenation operation in the radiograph transformer. Task1 and Task2 refer to the two tasks in cross-supervised learning. Row 1 corresponds to the result of a convolutional neural network while row four corresponds to LSP (Transformer)

distribution and is often located around the pleura. These successes imply that REFERS can deal with the diagnosis of both elusive and regular abnormalities and diseases well using a small number of training radiographs in the target domain. A similar phenomenon can be observed when REFERS is used for distinguishing viral pneumonia cases from bacterial cases in Tables 1 and 2.

Transformer is more effective under limited supervision. We observe a trend of CNNs (that is, ResNet series[4]) in Tables 1 and 2: LSP (ConvNet) shows mediocre performance when a relatively small number of training images in the target domain are used; however, when all training data (100% label ratio) are used, ConvNet shows competitive results. It seems that LSP (ConvNet) cannot manage small amounts of supervision well. By contrast, LSP (Transformer) exhibits much better performance at small label ratios. This comparison demonstrates that pre-trained transformers generate more transferable representations than pre-trained CNNs. The underlying reason might be that the self-attention mechanism in transformers makes the learned representations more transferable due to captured long-distance dependencies.

REFERS provides reliable evidences for clinical decisions. Figure 3 presents randomly chosen radiographs and their corresponding class activation maps[27]. We can find that REFERS generates reliable attention regions, on top of which we can apply a fixed confidence threshold to further identify the location of different types of lesions (green boxes in Fig. 3). The overall intersection over unions between green and red boxes (drawn by radiologists) are mostly higher than 0.5, indicating that the generated attention regions can well match radiologists's diagnoses. When lesions have a large size (for example, the fifth image from NIH ChestX-ray, i.e., Fig. 3e), our method captures well-aligned lesion areas. Even when lesions are quite small and therefore hard to detect (such as the last image from NIH ChestX-ray and the first image from VinBigData), REFERS can still identify the right locations.

Replication of experimental results and their statistical significance. There are a number of factors that influence pre-training results exhibit a certain level of randomness. These factors include—but are not limited to—network initialization, training strategy (for example, how to randomly crop images and perform mini-batch gradient descent) and even non-deterministic characteristics in computational tools (for example, cuDNN [28] would choose different algorithms in different runs due to benchmarking noise and hardware configuration). A good pre-training methodology should be able to produce relatively stable pre-trained representations when randomness in these factors is controlled within an acceptable limit. To take into account the influence of such randomness on experimental results, when REFERS and baseline pre-trained models are fine-tuned, we independently repeat each experiment three times and report their average results in Tables 1 and 2. We then calculate *P*-values between mean class AUCs of REFERS and the best-performing baseline model according to their fine-tuned performance using independent two-sample *t*-test. According to Tables 1 and 2, nearly all *P*-values are much smaller than 0.01, indicating that REFERS is significantly better than its counterparts when various amounts of labelled training data in the target domain is used. By contrast, making the number of times (repeating each experiment) smaller than three would give rise to less stable mean AUCs while simply repeating more times would produce meaninglessly smaller *P*-values.

Last but not the least, we provide a thorough ablation study of REFERS in Table 3. More details can be found in the Methods.

Methods

Dataset for pre-training (source domain). MIMIC-CXR-JPG[23] contains over 370,000 radiographs organized into patient studies, each of which may have one or more radiographs taken from different views, or at different times for the same patient. Each patient study has one free-text radiology report and each radiograph is associated with a set of abnormality/disease labels obtained from two-stage human-assisted intervention, as mentioned above. There are two major sections in each report: findings and impressions. The former includes detailed descriptions of important aspects in the radiographs, whereas the latter summarizes most immediately relevant findings.

To acquire human-assisted structured labels for radiographs (that is, two-stage human intervention), annotators need to first define a list of labels for abnormalities and diseases, including alternate spellings, synonyms and abbreviations. On the basis of local contexts and existing NLP tools, mentions of labels in reports are classified as positive, uncertain or negative. An aggregation procedure is further applied to aggregate multiple mentions of a single label. Uncertain labels need to be double-checked by radiologists.

As radiology reports were originally prepared by radiologists as part of the daily clinical routine, they can be regarded as freely available information that does not require extra human efforts, in contrast to structured labels. In practice, we only keep the findings and impressions sections in the reports. We also remove all study–report pairs—where the text section has less than three tokens (words and phrases)—from the dataset. This screening procedure produces 217,000 patient studies.

Datasets for fine-tuning (target domains). We do not require these datasets adopted for fine-tuning to have radiology reports. Instead, only human-assisted annotations are used during the fine-tuning stage. We follow the official split of NIH ChestX-ray, where the percentages of training, validation and testing sets are 70%, 10% and 20%, respectively. The same set of ratios are also employed for the VinBigData Chest X-ray, Shenzhen Tuberculosis and COVID-19 Image Data Collection datasets to build randomly split training, validation and testing sets.

- NIH ChestX-ray is a dataset for multilabel classification of 14 chest abnormalities (that is, atelectasis, cardiomegaly, consolidation, oedema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia and pneumothorax). There are over 100,000 frontal-view X-ray images of about 32,000 patients in NIH ChestX-ray, where labels of radiographs were extracted from associated reports following a similar procedure as that for MIMIC-CXR-JPG.
- VinBigData Chest X-ray provides labels of 14 chest diseases (that is, aortic enlargement, atelectasis, pneumothorax, lung opacity, pleural thickening, interstitial lung disease, pulmonary fibrosis, calcification, pleural effusion, consolidation, cardiomegaly, other lesion, nodule-mass and infiltration) and consists of 15,000 postero-anterior chest X-ray images. Here we did not use the test set in Kaggle, which does not provide any annotations. All images were labelled by a panel of experienced radiologists.
- Shenzhen Tuberculosis is a small dataset containing 662 frontal chest X-ray images primarily from a hospital clinical routine; 336 abnormal X-rays show various manifestations of tuberculosis, whereas the remaining 326 images are normal. We simply perform binary classification on this dataset.
- COVID-19 Image Data Collection is a dataset involving more than 900 pneumonia cases with chest X-rays, which was built to improve the identification of COVID-19. We conduct experiments on two tasks, which (1) distinguish COVID-19 from non-COVID-19 cases and (2) separate viral pneumonia cases from bacterial cases.

Baselines and label-supervised pre-training. As our method does not need the structured labels that are required by traditional fully supervised learning,

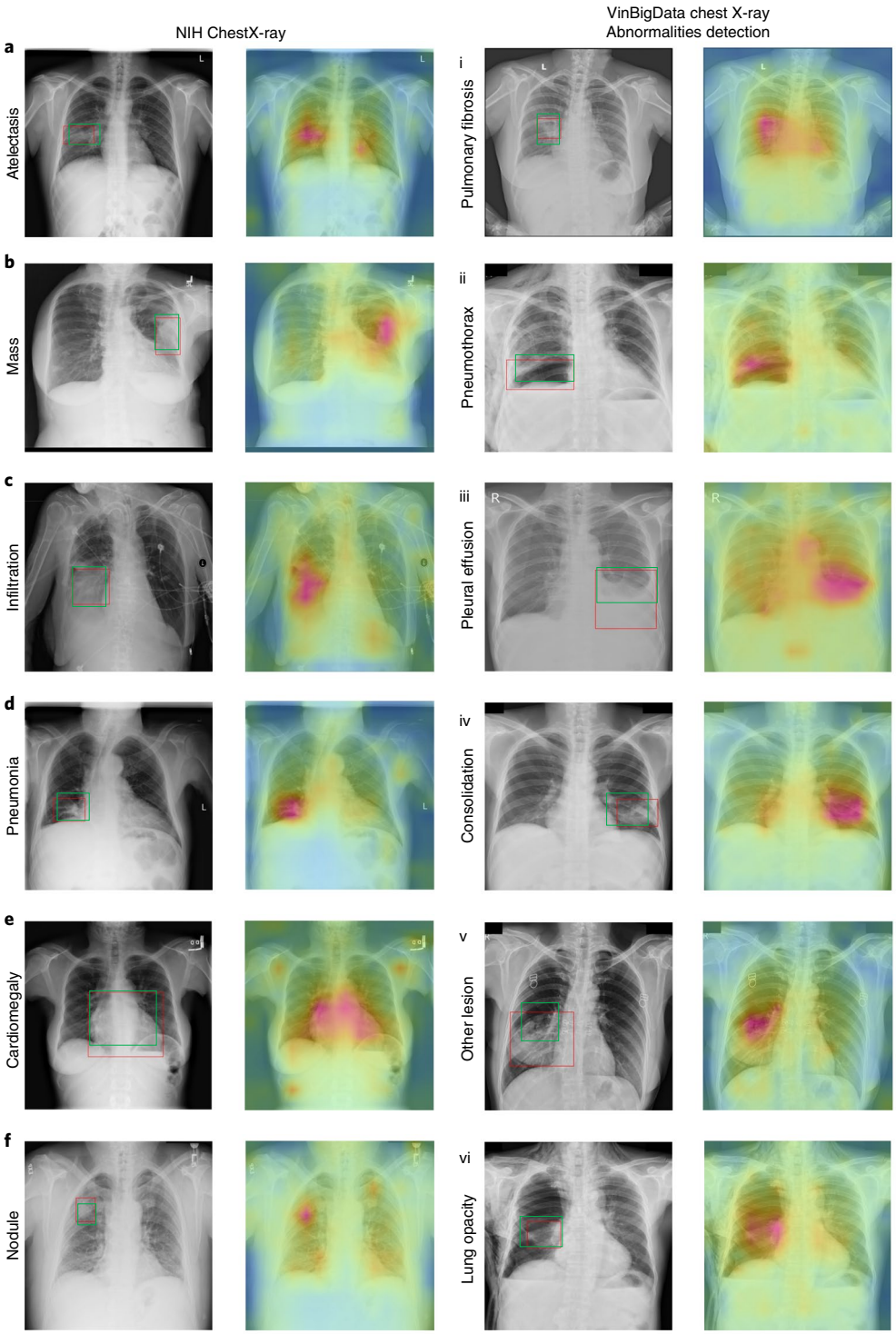

Fig. 3 | Sample visualization. a–f, Sample visualization of twelve randomly chosen samples from NIH ChestX-ray (**a**–**f**) and VinBigData (i–vi) (fine-tuned with all-annotated training data) or each sample, we present both the original image (left) and an attention map generated from REFERS. In each original image, red boxes denote lesion areas annotated by radiologists. In the attention maps, the fuchsia color represents attention values generated from REFERS. The darker the color, the higher the confidence of a specific disease. Green boxes in original images are our predicted lesion areas generated by applying a fixed confidence threshold to attention maps.

we compare it against four recent self-supervised learning methods[13–16] and ImageNet-based pre-training[11]:

- Context Restoration[13] repeats the operation of swapping two randomly chosen small X-ray patches a fixed number of times, and the neural network is asked to restore each altered image back to its original version.
- Model Genesis[14] applies multiple types of distortions to the input X-ray, including local shuffling, non-linear transformation, in- and out-painting. Similar to Context Restoration, Model Genesis asks the model to reconstruct the original image from the distorted one.
- TransVW[15] contrasts local X-ray patches to exploit the semantics of anatomical patterns while restoring distorted image contents.
- C2L[16] proposes to construct homogeneous and heterogeneous data pairs by mixing both images and features on top of MoCo[29]. C2L outperforms MoCo by observable margins on multiple X-ray benchmarks.

- ImageNet-based pre-training[11] is taken as a representative method that sets a large-scale dataset of annotated natural images as the source domain.
- Note that all above baselines are implemented using the same transformer-based network architecture as REFERS (that is, a ViT architecture plus the proposed recurrent concatenation module). Such an implementation arrangement is meant to rule out the influence of network architectures on final performance and maintain fairness in experimental comparisons.

Finally, our approach is compared with LSP, which directly sets a large collection of X-ray images with human-assisted structured labels as the source domain. For better comparison, we implement LSP on top of both CNN- and Transformer-based backbone networks. Specifically, LSP (Transformer) adopts the same Transformer-based network architecture as REFERS and the aforementioned self-supervised and ImageNet-based pre-training baselines. LSP (ConvNet) represents the best-performing residual network among ResNet-18, ResNet-50 and ResNet-101[4].

Data augmentation and image resizing. During the pre-training stage, we resize each radiograph in the source domain to 256×256 pixels and then apply random cropping to produce 224×224 images. Random horizontal flip, random rotation (−10 to 10 degrees) and random grayscale (brightness and contrast) are also applied to generate augmented images. When using random horizontal flip, we change the words left and right in the accompanying radiology report accordingly. During the fine-tuning stage, we apply the same set of data-augmentation strategies—random cropping, random rotation, random grayscale and random horizontal flip—to all four target domain datasets. As in the pre-training stage, we resize each radiograph in a target domain to 256×256, and then generate 224×224 cropped and augmented radiographs as input images.

Algorithm overview. REFERS performs cross-supervised learning on top of a transformer-based backbone, called radiograph transformer. Given a patient study, we first forward its views to the radiograph transformer for extracting view-dependent feature representations. We next perform cross-supervised learning that acquires study-level supervision signals from free-text radiology reports. To this aim, it is necessary and essential to use view fusion to obtain a unified visual representation for an entire patient study as each radiology report is associated with a patient study but not individual radiographs within the patient study. Such fused representations are then used in two tasks during the pre-training stage: report generation and study–report representation consistency reinforcement. The first task takes the free texts in original radiology reports to supervise the training process of the radiograph transformer. The second task reinforces the consistency between the visual representations of patient studies and the textual representations of their corresponding reports.

Radiograph transformer. The radiograph transformer accepts image patches as inputs. We divide each image into a grid of 14×14 cells, each of which has 16×16 pixels. We then flatten each image patch to form a one-dimensional vector of pixels and feed it into the transformer. At the beginning of the transformer, a patch-embedding layer linearly transforms each one-dimensional pixel vector into a feature vector. This vector is concatenated with a position feature produced from a learnable position embedding to help clarify the relative location of each patch in the whole input patch sequence. The concatenated feature is then passed through another linear transformation layer to make its dimensionality the same as that of the final radiograph feature. We stack twelve self-attention blocks at the core part of the radiograph transformer, which have the same architecture but independent parameters (Fig. 1b). We first follow the practice in ref. [20] to build a single self-attention block and then repeat its operations multiple times. In each block, we apply layer normalization[30] before the multi-head attention and perceptron layers, after which residual connections are added to stabilize the training process. In the perceptron layer, we employ a two-layer perceptron with the rectified linear unit[31] as the activation function. Moreover, we add an aggregation embedding, which is responsible for gathering the information from different input features. As shown in Fig. 1b, in the last layer, recurrent concatenation is performed to repeatedly concatenate the learned aggregation embedding with the learned representation of every patch. This is different from the operation in ViT[19], which only concatenates the aggregation embedding with patch features once.

Cross-supervised learning. There are two major components in cross-supervised learning: the view fusion module for producing study-level representations and two report-related tasks exploiting study-level information from associated free-text reports.

As aforementioned, we forward all radiographs in a patient study through the radiograph transformer simultaneously to obtain their individual representations. We further employ an attention mechanism to fuse these individual representations and obtain an overall representation of the given study. Supposing a study has three radiographs (that is, views), as shown in Fig. 1c. We first concatenate the features of all views and then feed the concatenated features to a multilayer perceptron to compute an attention value for each view. We next apply the softmax function to normalize these attention values, which are used as weights to produce a weighted version of the individual representations. Finally, these weighted representations are concatenated to form a unified visual feature for describing the whole study. Note that for studies that contain fewer than three radiographs, we randomly select one of the radiographs, and then repeat it once or twice to have a total of three views. For studies that contain more than three radiographs, we randomly select three of them from each study as input views.

We design two report-related tasks that acquire cross-supervision signals from free-text reports: report generation and study–report representation consistency reinforcement. In practice, these two tasks exploit study-level free-text information to better train study-level visual representations produced from the view fusion module. The first task applies a decoder called report transformer to the unified visual feature $\mathbf{v}^k$ of the kth patient study to reproduce its associated radiology report, denoted as $c^k_{1:T}$. Here, c^k_1 and c^k_T represent the start- and end-of-sequence tokens, respectively. As a result, the report transformer generates a sequence of token-level predictions, $\hat{c}^k_{1:T}$, for the kth patient study. The prediction of the tth token in this sequence depends on the predicted subsequence $\hat{c}^k_{1:t-1}$ and the visual feature $\mathbf{v}^k$. The network architecture of the report transformer follows the architecture of the decoder in ref. [20]. We wish the predicted token sequence ($\hat{c}^k_{1:T}$) resembles the sequence ($c^k_{1:T}$) representing the original report of the kth patient study; therefore, as shown in Fig. 1d, we apply a language modelling loss to both $\hat{c}^k_{1:T}$ and $c^k_{1:T}$ to maximize the following log-likelihood of the tokens in the original report.

$$\mathcal{L}^k_{\text{language}} = \sum_{t=2}^{T} \log P\left(c^k_t | \hat{c}^k_{1:t-1}, \mathbf{v}^k; \phi_v, \phi_t\right), \tag{1}$$

where P denotes conditional probability, $\hat{c}^k_1$ is a special symbol indicating the start of the predicted sequence, and ϕ_v and ϕ_t are the parameters of the radiograph and report transformers, respectively.

For the second task on study–report representation consistency reinforcement, we employ a contrastive loss[32] to align cross-modal representations. Here we use $\mathbf{t}^k$ to represent the textual feature vector of the k*th* radiology report. In practice, we obtain $\mathbf{t}_k$ by forwarding the sequence of tokens in the k*th* report (that is, $c^k_{1:T}$) to a bidirectional encoder representations from transformer (BERT) model[33]. BERT is built on top of the encoder in ref. [20] using large-scale pre-training on a great number of corpus resources; thus, BERT can help produce a generalized textual representation for the input report. Suppose we have B patient studies in each training mini-batch, as shown in Fig. 1d. The contrastive loss for the k*th* study can be formulated as

$$\mathcal{L}^k_{\text{contrast}} = -\log \frac{e^{\cos(\mathbf{v}^k, \mathbf{t}^k)/\tau}}{\sum_{i=1}^{B} e^{\cos(\mathbf{v}^k, \mathbf{t}^i)/\tau}}, \tag{2}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity, $\cos(\mathbf{v}^k, \mathbf{t}^k) = \frac{(\mathbf{v}^k)^\top \mathbf{t}^k}{\|\mathbf{v}^k\| \|\mathbf{t}^k\|}$, $\top$ denotes the transpose operation, $\|\cdot\|$ represents L2 normalization, and τ is the temperature factor. Finally, for each patient study, we simply sum up $\mathcal{L}^k_{\text{contrast}}$ and $\mathcal{L}^k_{\text{language}}$ as the overall loss. During the fine-tuning stage, we typically use the cross-entropy loss for model tuning.

Training and testing methodologies. We first pre-train the radiograph transformer on the source domain and then fine-tune it on downstream target domain datasets to verify the quality of pre-training. During the pre-training stage, we sample 4,600 studies to form a held-out validation set according to the official division of the MIMIC-CXR-JPG dataset[23]. We train the entire network using stochastic gradient descent (SGD) while setting the momentum value to 0.9 (ref. [34]) and the weight decay to 1×10^{-4}. Following ref. [33], we do not apply weight decay to layer normalization and the bias terms in all layers. We use a fixed batch size of 32 for 300,000 iterations (about 45 epochs). We calculate the validation loss after each epoch and save the checkpoint that achieves the lowest validation loss. We adopt the linear learning rate warm-up strategy[35] for the first 10,000 iterations, and then switch to cosine decay[36] until the end. Empirically, we found that training the radiograph transformer requires a large learning rate for fast convergence; thus, its learning rate is set to 3×10^{-3}, and set to 3×10^{-4} for the report transformer and BERT. We initialize the aggregation embedding to all zeros while randomly initializing all position embeddings. We use PyTorch[37] and NVIDIA Apex for mixed-precision training[38]. The complete pre-training process on the MIMIC-CXR dataset takes about two days on a single RTX 3090 GPU.

During the fine-tuning stage, we fine-tune all transformer-based models (including transformer-based baselines) using SGD with the momentum set to 0.9 and the initial learning rate set to 3×10^{-3} for all datasets. We fine-tune ResNet models using Adam[39] instead of SGD, and set the initial learning rate to 1×10^{-4}. All downstream models use the same learning rate decay strategy as that used in the pre-training stage, and are trained with a batch size of 128.

Ablation study. We conduct a thorough ablation study of REFERS by removing or replacing individual modules; the results are shown in Table 3.

First, we investigate the impact of replacing the radiograph transformer (rows 1 and 2 in Table 3). If we replace the radiograph transformer with ResNet-101[4] (row 1), the overall performance of REFERS on COVID-19 Image

Data Collection would drop by about 7% (compared with row 0). This comparison demonstrates that the radiograph transformer is more effective at dealing with limited annotations, which is also verified by the results in Tables 1 and 2. Next, when we replace the radiograph transformer with the original ViT architecture (row 2), which does not have the recurrent concatenation operator, the overall performance would drop by 3.3%. This result verifies the helpfulness of recurrently concatenating the learned aggregation embedding with patch representations. We also note that there exists a 3.8% performance difference between ResNet and ViT based architectures (rows 1 and 2), showing the advantage of a transformer-like architecture.

In addition to the radiograph transformer, we also investigate the impact of cross-supervised learning. First of all, we remove the view fusion module so that different radiographs within a patient study become associated with the same study-level radiology report (row 3). Such an operation is counter-intuitive as each individual radiograph alone cannot provide enough information to produce a study-level report. By comparing row 3 with row 0, we found that dropping the view fusion module would reduce the performance by nearly 2% on COVID-19 Image Data Collection. This result implies that learning study-level pre-trained representation is better than image-level pre-training as the former includes more patient-level information. Next we completely replace cross-supervised learning with label-supervised learning (row 4), and REFERS deteriorates into LSP (Transformer) in Table 2. We found that dropping the two report-related tasks would adversely affect the performance by 2%. Last but not least, we study the two report-related learning tasks individually. By comparing row 0 with rows 5 and 6, respectively, we observed that dropping either of them would not affect the overall performance too much (about 1%). This result implies that the effects of both tasks may partially overlap to some extent. Nonetheless, either of them along with the view fusion module can still outperform LSP (Transformer) (row 4). We also found that although both of them improve the overall performance, reinforcing the consistency between representations of each patient study and its associated report (that is, the second task) is more crucial than report generation (that is, the first task). We believe the reason behind this is that the representation learned in the second task can be regarded as a summary of each report, and thus provides more global information than token-level predictions in the first task. Such advantages make it more beneficial for the second task to include more study-level information for learning better study-level radiograph features.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

MIMIC-CXR-JPG data can be found at https://physionet.org/content/mimic-cxr-jpg/2.0.0/. NIH Chest X-ray data can be found at https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345. VinBigData Chest X-Ray Abnormalities Detection data can be found at https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection. Shenzhen Tuberculosis data can be found at https://www.kaggle.com/raddar/tuberculosis-chest-xrays-shenzhen. The COVID-19 Image Data Collection can be found at https://github.com/ieee8023/covid-chestxray-dataset.

Code availability

All codes are available at https://github.com/funnyzhou/REFERS (ref. [40]).

Received: 22 June 2021; Accepted: 11 November 2021;
Published online: 20 January 2022

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. 26th Advances in Neural Information Processing Systems* 1097–1105 (NeurIPS, 2012).
2. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *2nd International Conference on Learning Representations* (ICLR, 2014).
3. Szegedy, C. et al. Going deeper with convolutions. In *Proc. 28th IEEE Conference on Computer Vision and Pattern Recognition* 1–9 (IEEE, 2015).
4. He, K. M., Zhang, X. Y., Ren, S. Q. & Sun, J. Deep residual learning for image recognition. In *Proc. 29th IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
5. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
6. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (IEEE, 2013).
7. Phillips, N.A. et al. CheXphoto: 10,000+ photos and transformations of chest X-rays for benchmarking deep learning robustness. In *Proc. 5th Machine Learning for Health* 318–327 (PMLR, 2020).
8. Taylor, A. G., Mielke, C. & Mongan, J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med.* **15**, e1002697 (2018).
9. Carlile, M. et al. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. *JACEP Open* **1**, 1459–1464 (2018).
10. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Proc. 28th Advances in Neural Information Processing Systems* 3320–3328 (NeurIPS, 2014).
11. Wang, X.S. et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition* 2097–2106 (IEEE, 2017).
12. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 22nd IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
13. Chen, L. et al. Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **58**, 101539 (2019).
14. Zhou, Z. W., Sodha, V., Pang, J. X., Gotway, M. B. & Liang, J. M. Model genesis. *Med. Image Anal.* **67**, 101840 (2021).
15. Haghighi, F., Taher, M. R. H., Zhou, Z. W., Gotway, M. B. & Liang, J. M. Transferable visual words: exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Trans. Med. Imag.* **40**, 2857-2868 (IEEE, 2021).
16. Zhou, H.-Y. et al. Comparing to learn: surpassing ImageNet pretraining on radiographs by comparing image representations. In *Proc. 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention* 398–407 (Springer, 2020).
17. Johnson, A.E.W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 1–8 (2019).
18. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. 33rd AAAI Conference on Artificial Intelligence* 590–597 (AAAI, 2019).
19. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations* (ICLR, 2021).
20. Vaswani, A. et al. Attention is all you need. In *Proc. 31st Advances in Neural Information Processing Systems* 5998–6008 (NeurIPS, 2017).
21. Shin, H.-C. et al. Interleaved text/image deep mining on a very large-scale radiology database. In *Proc. 28th IEEE Conference on Computer Vision and Pattern Recognition* 1090–1099 (IEEE, 2015).
22. Wang, X. S., Peng, Y. F., Lu, L., Lu, Z. Y & Summers, R. M. Tienet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proc. 31st IEEE Conference on Computer Vision and Pattern Recognition* 9049–9058 (IEEE, 2018).
23. Johnson, A. E. W. et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at https://arxiv.org/abs/1901.07042 (2019).
24. Nguyen, H. Q. et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Preprint at https://arxiv.org/abs/2012.15029 (2021).
25. Jaeger, S. et al. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **4**, 475 (2014).
26. Joseph, P. C. et al. COVID-19 Image Data Collection: prospective predictions are the future. Preprint at https://arxiv.org/abs/2006.11988 (2020).
27. Zhou, B.L., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proc. 29th IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929 (IEEE, 2016).
28. Chetlur, S. et al. cuDNN: Efficient primitives for deep learning. Preprint at https://arxiv.org/abs/1410.0759 (2014).
29. He, K. M., Fan, H. Q., Wu, Y. X., Xie, S. N., & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. 33rd IEEE Conference on Computer Vision and Pattern Recognition* 9729–9738 (IEEE, 2020).
30. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. In *4th International Conference on Learning Representations* (ICLR, 2016).
31. Dahl, G.E., Sainath, T.N. & Hinton, G.E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Proc. 38th International Conference on Acoustics, Speech and Signal Processing* 8609–8613 (IEEE, 2013).
32. Gutmann, M. & Hyvärinen, A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In *Proc. 13th International Conference on Artificial Intelligence and Statistics* 297–304 (JMLR, 2010).
33. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4171–4186 (ACL, 2019).

34. Sutskever, I., Martens, J., Dahl, G. & Hinton, G.E. On the importance of initialization and momentum in deep learning. In *Proc. 38th International Conference on Machine Learning* 1139–1147 (PMLR, 2013).
35. Goyal, P., Mahajan, D., Gupta, A. & Misra, I. Scaling and benchmarking self-supervised visual representation learning. In *Proc. 17th International Conference on Computer Vision* 6391–6400 (IEEE, 2019).
36. Loshchilov, I. & Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations* (ICLR, 2017).
37. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. 33rd Advances in Neural Information Processing Systems* 8024–8035 (2019).
38. Micikevicius, P. et al. Mixed precision training. In *6th International Conference on Learning Representations* (ICLR, 2018).
39. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. In *2nd International Conference on Learning Representations* (ICLR, 2014).
40. Zhou, H.Y. et al. *Generalized Radiograph Representation Learning via Cross-supervision between Images and Free-text Radiology Reports* (Zenodo, 2021); https://doi.org/10.5281/zenodo.5624117

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities (grant nos. 20720190012 and 20720210121).

Author contributions

H.Z. and Y.Y. conceived the idea and designed the experiments. H.Z., X.C. and Y.Z. implemented and performed the experiments. H.Z. and Y.Y. wrote the manuscript. All authors analysed the data and experimental results, and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00425-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00425-9.

Correspondence and requests for materials should be addressed to Liansheng Wang or Yizhou Yu.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Training data	Method	Mean	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	Nodule	Pleural Thickening	Pneumonia	Pneumothorax	p-value
0.8k (1%)	Our REFERS	**76.7**	**77.5**	**85.6**	**78.6**	**84.9**	**85.4**	**79.5**	**72.3**	**77.1**	**67.5**	**76.2**	**66.5**	**71.6**	**69.3**	**81.7**	8.35e-4
	Model Genesis	70.3	72.1	67.1	75.8	76.1	80.6	72.6	64.8	73.5	65.7	65.2	62.2	67.6	64.8	76.2	
	C2L	71.0	75.1	67.1	77.6	75.1	83.4	71.5	66.8	70.0	63.8	70.1	66.2	68.1	65.7	74.4	
	Context Restoration	67.8	69.1	64.4	73.2	73.8	78.1	70.0	62.1	70.2	65.2	62.4	59.1	65.0	62.2	73.8	
	TransVW	71.2	74.5	68.9	76.7	79.8	81.1	67.9	68.7	68.2	66.8	66.5	66.2	68.5	68.8	75.0	
	ImageNet Pre-training	69.8	73.3	69.6	76.0	81.7	80.5	67.1	64.9	64.8	65.8	67.0	62.3	65.7	65.0	74.0	
8k (10%)	Our REFERS	**80.9**	**80.1**	**89.8**	**79.5**	**87.8**	**87.5**	**88.2**	**77.2**	**86.1**	**69.6**	**82.0**	**72.8**	**74.2**	**72.2**	**85.6**	8.72e-4
	Model Genesis	75.7	77.2	72.8	77.5	85.7	85.2	81.0	75.3	78.0	68.4	73.1	69.5	72.2	67.7	80.4	
	C2L	76.6	78.0	75.5	77.5	84.1	85.7	81.2	73.7	79.5	67.4	77.5	71.7	72.0	67.3	81.9	
	Context Restoration	73.9	75.5	70.6	77.1	84.5	84.2	79.4	73.1	67.5	68.1	70.9	66.9	71.7	65.2	79.1	
	TransVW	74.3	76.5	70.8	77.6	83.0	84.8	79.7	69.9	74.7	68.5	72.1	68.3	72.4	63.2	79.6	
	ImageNet Pre-training	74.4	74.2	79.8	75.9	85.7	83.2	80.4	72.1	74.0	64.1	71.7	65.6	69.6	66.2	79.7	
80k (100%)	Our REFERS	**84.7**	**83.0**	**92.3**	**82.1**	**90.2**	**88.7**	**91.4**	**83.9**	**93.3**	**74.1**	**85.5**	**76.7**	**78.5**	**77.0**	**89.1**	1.94e-3
	Model Genesis	81.0	78.8	84.5	79.2	87.8	86.6	89.7	81.0	85.2	71.1	81.9	73.2	75.8	73.0	85.6	
	C2L	82.2	81.1	90.2	81.0	88.1	88.0	88.3	80.8	86.8	72.0	82.7	74.1	76.2	75.3	85.9	
	Context Restoration	78.7	75.8	82.9	76.4	86.6	84.8	88.2	78.6	83.0	70.0	79.6	69.5	73.2	69.4	84.0	
	TransVW	81.7	79.8	85.0	80.0	88.2	87.1	90.1	81.8	85.9	72.3	82.6	74.4	76.6	74.0	86.1	
	ImageNet Pre-training	80.0	78.3	89.3	77.6	87.9	85.9	87.4	78.5	88.8	65.9	79.9	70.7	74.5	71.0	84.7	

Extended Data Fig. 1 | Comparison with self-supervised and transfer learning on NIH ChestX-ray. Comparison with self-supervised learning and transfer learning baselines on NIH ChestX-ray dataset. Note that for the sake of fairness, all baselines use the same transformer-based backbone as the radiograph transformer of REFERS (that is, a ViT-like architecture plus the recurrent concatenation operator). Each *P*-value is calculated between our REFERS and the best-performing baseline. The evaluation metric is Area under the ROC Curve (AUC). Best results are bolded.

Training data	Method	Mean	Aortic Enlargement	Atelectasis	Calcification	Cardiomegaly	Consolidation	ILD	Infiltration	Lung Opacity	Nodule-Mass	Other Lesion	Pleural Effusion	Pleural Thickening	Pneumothorax	Pulmonary Fibrosis	p-value
0.1k (1%)	Our REFERS	**83.0**	88.4	**85.1**	**71.2**	**91.4**	**88.6**	**83.7**	**81.9**	**86.3**	**74.7**	**77.7**	**86.0**	**81.8**	**85.7**	**79.4**	8.72e-5
	Model Genesis	70.7	87.8	65.7	57.6	87.5	72.9	67.0	64.3	63.6	67.7	68.7	77.5	77.7	63.1	68.4	
	C2L	75.3	**89.1**	71.3	64.5	88.7	77.4	71.2	70.3	70.2	72.5	75.0	81.4	80.7	69.5	72.5	
	Context Restoration	67.9	74.4	63.6	63.7	74.7	71.2	68.6	67.5	70.0	64.2	65.4	67.7	68.0	64.8	66.5	
	TransVW	73.6	82.8	68.4	68.5	83.4	74.7	73.6	72.6	75.8	69.0	70.3	74.1	75.1	68.7	73.5	
	ImageNet Pre-training	69.7	77.7	64.5	67.3	80.0	70.1	69.6	68.1	73.3	64.9	67.2	69.3	70.3	62.7	70.1	
1k (10%)	Our REFERS	**88.2**	**92.6**	**89.6**	78.4	**92.9**	**94.4**	**86.7**	**87.4**	**91.2**	**83.6**	**84.7**	**90.2**	**88.1**	**89.6**	**85.8**	4.34e-4
	Model Genesis	82.7	88.6	83.3	78.2	86.6	87.4	79.4	81.9	83.6	79.9	82.3	83.5	85.0	74.9	83.2	
	C2L	83.3	92.1	80.3	**78.6**	89.5	82.6	81.8	82.6	84.7	80.9	83.5	85.0	**88.1**	72.1	85.0	
	Context Restoration	82.4	91.4	75.1	72.4	89.6	81.6	80.7	79.6	85.5	80.7	82.6	86.0	87.6	76.6	84.7	
	TransVW	83.8	91.7	80.3	72.5	89.7	85.9	81.3	82.0	85.7	82.0	84.4	86.7	87.6	76.9	85.0	
	ImageNet Pre-training	82.9	91.6	81.1	73.3	89.6	87.8	79.6	82.1	85.7	82.4	84.0	86.6	87.5	75.1	85.5	
10k (100%)	Our REFERS	**90.1**	**93.6**	**90.2**	80.4	**93.6**	**95.1**	**91.2**	**90.3**	**93.2**	**84.6**	**86.9**	**92.3**	**90.5**	**89.7**	**89.5**	9.33e-4
	Model Genesis	85.8	90.8	84.8	79.0	89.4	89.0	81.6	82.6	85.6	81.4	85.0	86.3	86.5	80.3	84.4	
	C2L	85.9	92.5	82.5	**80.9**	90.6	86.3	85.0	85.2	87.1	82.1	84.7	87.2	88.5	82.6	86.9	
	Context Restoration	83.8	92.7	76.1	73.0	90.9	84.2	81.7	81.3	86.8	82.0	84.6	87.1	87.9	80.8	84.8	
	TransVW	86.2	92.0	81.7	76.0	90.1	86.2	86.6	88.0	87.1	83.5	85.3	88.3	88.8	86.0	86.6	
	ImageNet Pre-training	84.5	92.6	81.4	75.9	91.5	88.3	80.5	83.0	86.6	82.7	84.2	87.2	87.7	79.9	86.0	

Extended Data Fig. 2 | Comparison with self-supervised and transfer learning on VinBigData Chest X-ray Abnormalities Detection. Comparison with self-supervised learning and transfer learning baselines on VinBigData Chest X-ray Abnormalities Detection. Note that for the sake of fairness, all baselines use the same transformer-based backbone as the radiograph transformer of REFERS (that is, a ViT-like architecture plus the recurrent concatenation operator). Each *O*-value is calculated between our REFERS and the best-performing baseline. The evaluation metric is Area under the ROC Curve (AUC). Best results are bolded.

Training data	Method	Mean	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	Nodule	Pleural Thickening	Pneumonia	Pneumothorax	p-value
0.8k (1%)	LSP (Transformer)	74.2	75.5	85.0	77.2	**85.0**	85.3	71.5	70.6	64.5	66.8	72.4	66.0	69.4	68.8	80.4	
	LSP (ConvNet)	65.8	66.9	62.4	71.3	72.1	76.2	68.0	60.1	67.4	64.6	60.3	56.8	63.1	60.1	71.9	3.25e-3
	Our REFERS	**76.7**	**77.5**	**85.6**	**78.6**	84.9	**85.4**	**79.5**	**72.3**	**77.1**	**67.5**	**76.2**	**66.5**	**71.6**	**69.3**	**81.7**	
8k (10%)	LSP (Transformer)	78.2	77.9	86.3	77.7	87.2	85.5	83.8	76.0	80.2	67.3	76.0	69.7	73.0	71.4	82.4	
	LSP (ConvNet)	74.5	76.2	71.4	77.0	85.0	84.6	80.0	74.0	69.5	68.0	71.7	67.9	72.2	66.1	79.6	2.89e-3
	Our REFERS	**80.9**	**80.1**	**89.8**	**79.5**	**87.8**	**87.5**	**88.2**	**77.2**	**86.1**	**69.6**	**82.0**	**72.8**	**74.2**	**72.2**	**85.6**	
80k (100%)	LSP (Transformer)	82.1	80.1	90.0	80.2	89.2	87.3	88.7	81.1	89.4	70.0	81.3	74.5	76.8	75.5	85.4	
	LSP (ConvNet)	81.9	80.2	85.3	80.5	88.4	87.4	90.3	82.1	86.2	70.0	83.0	74.9	77.0	74.6	86.3	5.23e-3
	Our REFERS	**84.7**	**83.0**	**92.3**	**82.1**	**90.2**	**88.7**	**91.4**	**83.9**	**93.3**	**74.1**	**85.5**	**76.7**	**78.5**	**77.0**	**89.1**	

Extended Data Fig. 3 | Comparison with methods using human-assisted structured labels on NIH ChestX-ray. Comparison with label-supervised pre-training (LSP) on NIH ChestX-ray dataset. For fairness, both LSP (Transformer) and REFERS share the same transformer-based backbone (that is, the ViT architecture plus the recurrent concatenation operator). Each *P*-value is calculated between the results from our REFERS and LSP (Transformer). The evaluation metric is Area under the ROC Curve (AUC). Best results are bolded.

Training data	Method	Mean	Aortic Enlargement	Atelectasis	Calcification	Cardiomegaly	Consolidation	ILD	Infiltration	Lung Opacity	Nodule-Mass	Other Lesion	Pleural Effusion	Pleural Thickening	Pneumothorax	Pulmonary Fibrosis	p-value
0.1k (1%)	LSP (Transformer)	78.5	86.1	73.3	62.2	90.4	87.3	82.7	80.8	83.4	68.5	72.8	84.0	76.5	76.4	74.6	
	LSP (ConvNet)	76.0	**89.2**	71.1	64.2	88.9	78.4	72.4	70.7	70.9	72.8	75.4	81.8	80.9	74.4	72.8	3.56e-4
	Our REFERS	**83.0**	88.4	**85.1**	**71.2**	**91.4**	**88.6**	**83.7**	**81.9**	**86.3**	**74.7**	**77.7**	**86.0**	**81.8**	**85.7**	**79.4**	
1k (10%)	LSP (Transformer)	85.8	89.8	84.5	75.6	91.9	91.9	86.4	**87.8**	88.2	81.0	82.0	87.6	84.8	86.9	84.5	
	LSP (ConvNet)	85.2	91.7	86.1	76.9	89.7	84.8	85.4	83.6	85.8	82.6	84.5	86.1	88.5	81.3	**85.8**	8.69e-4
	Our REFERS	**88.1**	**92.6**	**89.6**	**78.4**	**92.9**	**94.4**	**86.7**	87.4	**91.2**	**83.6**	**84.7**	**90.2**	**88.1**	**89.6**	**85.8**	
10k (100%)	LSP (Transformer)	87.6	92.4	86.5	76.1	92.8	92.3	90.4	88.4	89.1	83.1	84.0	89.2	87.6	88.6	86.0	
	LSP (ConvNet)	87.2	92.8	87.7	79.3	91.3	88.7	87.9	86.0	88.1	83.2	85.7	88.1	88.7	86.0	86.8	1.05e-3
	Our REFERS	**90.1**	**93.6**	**90.2**	**80.4**	**93.6**	**95.1**	**91.2**	**90.3**	**93.2**	**84.6**	**86.9**	**92.3**	**90.5**	**89.7**	**89.5**	

Extended Data Fig. 4 | Comparison with methods using human-assisted structured labels on VinBigData Chest X-ray Abnormalities Detection. Comparison with label-supervised pre-training (LSP) on VinBigData Chest X-ray Abnormalities Detection. For fairness, both LSP (Transformer) and REFERS share the same transformer-based backbone (that is, the ViT architecture plus the recurrent concatenation operator). Each *P*-value is calculated between the results from our REFERS and LSP (Transformer). The evaluation metric is Area under the ROC Curve (AUC). Best results are bolded.

nature research

Corresponding author(s): Yizhou Yu

Last updated by author(s): Oct 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were obtained by direct download from public and online repositories via web browser.
Data analysis	Custom software was used to train machine learning models and analyze their outputs, and this code is available freely for download at https://github.com/funnyzhou/REFERS.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All radiographs are compiled from publicly-available data repositories and links for download are available at https://github.com/funnyzhou/REFERS.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The used radiograph datasets comprise images obtained from online repositories (as stated in the manuscript), and we used all images not excluded as described below. For our computational experiments, we chose the maximum sample sizes that would be feasible given our computing resources.
Data exclusions	In MIMIC-CXR-JPG, we exclude images without any radiology reports as the proposed pre-training method, REFERS, requires the information from radiology reports to learn radiograph representations. Also, we remove all study-report pairs, where the text section has less than 3 tokens (words and phrases), from MIMIC-CXR-JPG.
Replication	We include multiple replicates in each of our computational experiments. All attempts at replication of experimental findings were successful.
Randomization	In MIMIC-CXR-JPG, we sample 4.6k studies to form a held-out validation set, which follows the official division. In the rest 4 datasets (i.e., NIH ChestX-ray, VinBigData, Shenzhen Tuberculosis and COVID-19 Image Data Collection), radiographs were partitioned into training, validation, and testing folds at random.
Blinding	Since we always compare the performance of different computational algorithms by running these algorithms on the same dataset, there is no group allocation in our experiments. Thus blinding is not applicable in this paper.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging