MDPI

Article

Self-supervised Learning Application on COVID-19 Chest X-ray Image Classification Using Masked AutoEncoder

Xin Xing [1,2], Gongbo Liang [3], Chris Wang [4], Nathan Jacobs [5], Ai-Ling Lin [2,6,7]*

[1] Department of Computer Science, University of Kentucky, Lexington, KY, USA
[2] Department of Radiology, University of Missouri, Columbia, MO, USA
[3] Department of Computing and Cyber Security, Texas AM University-San Antonio, San Antonio, TX, USA
[4] Department of Computer Science, University of Missouri, Columbia, MO, USA
[5] Department of Computer Science & Engineering, Washington University in St. Louis, St. Louis, MO, USA
[6] Department of Biological Sciences, University of Missouri, Columbia, MO, USA
[7] Institute for Data Science and Informatics, University of Missouri, Columbia, MO, USA
* Correspondence: ai-ling.lin@health.missouri.edu;

Abstract: The COVID-19 pandemic has underscored the urgent need for rapid and accurate diagnosis facilitated by artificial intelligence (AI), particularly in computer-aided diagnosis using medical imaging. However, this context presents two notable challenges: high diagnostic accuracy demand and limited availability of medical data for training AI models. To address these issues, we proposed the implementation of Masked AutoEncoder (MAE), an innovative self-supervised learning approach, for classifying 2D Chest X-ray images. Our approach involved in performing imaging reconstruction using a Vision Transformer (ViT) model as the feature encoder, paired with a custom-defined decoder. Additionally, we fine-tuned the pre-trained ViT encoder using a labeled medical dataset, serving as the backbone. To evaluate our approach, we conducted a comparative analysis of three distinct training methods: training from scratch, transfer learning, and MAE-based training, all employing Covid-19 chest X-ray images. The results demonstrate that MAE-based training produces superior performance, achieving an accuracy of 0.985 and an AUC of 0.9957. We explored the mask ratio influence on MAE and found $ratio = 0.4$ shows the best performance. Furthermore, we illustrate that MAE exhibits remarkable efficiency when applied to labeled data, delivering comparable performance to utilizing only 30% of the original training dataset. Overall, our findings highlight the significant performance enhancement achieved by using MAE, particularly when working with limited datasets. This approach holds profound implications for future disease diagnosis, especially in scenarios where imaging information is scarce.

Keywords: Vision Transformer (ViT), Self-supervised learning, Chest X-ray Image, Image Classification

Citation: Xing, X.; Liang, G.; Wang, Chris; Jacobs, N.; Lin, AL MAE-Covid19. *Bioengineering* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

The COVID-19 pandemic has brought attention to the vital role of artificial intelligence (AI) in combating infectious diseases, specifically through the analysis of lung images, such as X-ray chest images. Computer-aided diagnosis (CAD) has emerged as a promising tool for accurate and rapid diagnosis in this context. Deep Learning (DL) models, known for their exceptional performance in computer vision tasks like image recognition [1–4], semantic segmentation [5,6], and object detection [7–9], have increasingly been adopted for CAD and other healthcare applications [10–13].

Despite the significant potential of DL models in medical data analysis, there are several practical challenges impeding their widespread adoption. First, medical datasets are often smaller compared to those used for natural image analysis, such as the widely used ImageNet dataset [14]. DL models have numerous parameters and require substantial data for effective training. Consequently, training such models with limited datasets can be

challenging and may lead to overfitting [15,16]. Second, labeling medical data is a resource-intensive and time-consuming process. CAD relies on labeled data for training, but labeling medical data necessitates specialized medical knowledge and expertise, making it more demanding than labeling natural images. Lastly, current DL models for medical image analysis primarily rely on convolutional neural networks (CNNs) [17,18]. While CNNs excel at capturing local features, they may not be optimally suited for capturing global information across an entire image. Therefore, further research is necessary to develop more effective DL network architectures capable of capturing both local and global features for medical image analysis, particularly when dealing with limited datasets.

To address these challenges, various methods have been proposed. For example, transfer learning [19] has gained widespread usage, wherein DL models are pre-trained on large-scale natural image datasets like ImageNet and subsequently fine-tuned on smaller medical datasets. This approach helps mitigate the overfitting issue caused by limited medical datasets, although it may not fully bridge the gap between natural and medical images. Regarding labeling issues, weakly supervised learning (WSL) [20,21] has become popular, where models are trained using only image-level labels for tasks like object detection [22]. However, WSL may not be suitable for classification tasks that still require image-level labels. Recently, novel DL models such as Vision Transformer (ViT) [23] and its variants [24,25] have demonstrated promising results in capturing global information from medical images. Nevertheless, these models often necessitate extensive amounts of data for effective training.

In our study, we explored a novel method for medical image analysis that addresses the challenges associated with training strategy and limited medical datasets. We propose the utilization of self-supervised learning (SSL) [26], a method that leverages the intrinsic attributes of the data as pre-training tasks, eliminating the reliance on labeled data. SSL implementation involves utilizing attributes such as image rotation prediction [27], patch localization [28], and image reconstruction [29], which can be accessed without manual labeling. Additionally, by substituting the conventional CNN backbone with a Vision Transformer (ViT) model, our method effectively captures both local and global features of medical images. To accomplish this, we employ a Masked Autoencoder (MAE) model [30], a self-supervised learning approach that utilizes the ViT model as its backbone. By combining SSL with ViT through the MAE model, we anticipate that this method can contribute to more accurate and efficient medical image analysis.

To our best knowledge, we are the first to apply MAE on Covid-19 X-ray imaging. During the exploration, we found the innovation application of MAE on the limited dataset, which is not studied by the previous work [30]. We demonstrated the superior performance of the MAE model compared to baseline models, explored the influence of mask ratios on the MAE model's performance, and evaluated the MAE model's performance using different proportions of limited training data. The contributions of our study are as follows:

- We conducted a comparative analysis of various training strategies using the same public Covid-19 dataset and observed that the MAE model outperformed other approaches, demonstrating superior performance.
- To further investigate the impact of different mask ratios on the MAE model's performance, we examined how varying mask ratios affected the effectiveness of the model. Our experiments revealed that the model achieved its best performance with a mask ratio of 0.4.
- Through extensive evaluations, we examined the applicability of the MAE model across different proportions of available training data. Remarkably, the MAE model achieved comparable performance even when trained with only 30% of the available data.

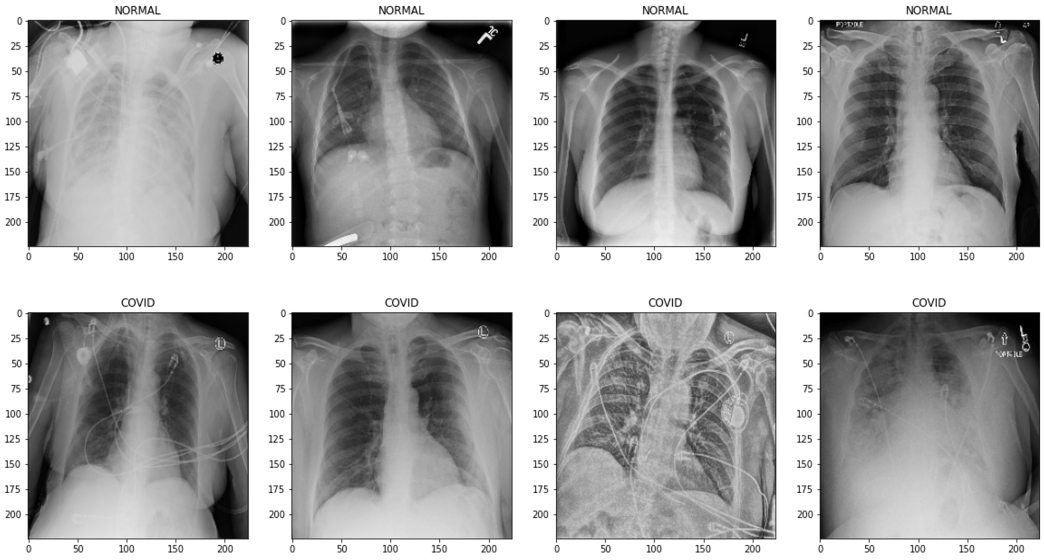

Figure 1. The visualization of the chest X-ray image of the: COVIDxCXR-3. The first row shows the negative subjects and the second row shows the positive subjects. The input image size is 224×224. We normalize the image pixel from 0 to 255.

Table 1. The images and patients distribution of the dataset COVIDxCXR-3.

Type	Negative	Positive	Total
Images Distribution			
Train	13992	15994	29986
Test	200	200	400
Patients Distribution			
Train	13850	2808	16648
Test	200	178	378

2. Materials and Methods

2.1. Data

Our study adopted the chest ray classification dataset: COVIDxCXR-3[31], which is a public dataset with more than 29,000 chest X-ray images, for positive/negative detection. COVIDxCXR-3 collects the data from different public data sources: covid-chest x-ray-dataset[32], COVID-19 Chest X-ray Dataset Initiative[33], Actualmed COVID-19 Chest X-ray Dataset Initiative[34], COVID-19 Radiography Database - Version 3[35], RSNA Pneumonia Detection Challenge[36], RSNA International COVID-19 Open Radiology Database (RICORD)[37], BIMCV-COVID19+: a large annotated dataset of RX and CT images of COVID19 patients[38], and Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU)[39]. Figure 1 visualizes the positive/negative image samples of the Covid-19 subject. Table 1 shows the details of the COVIDxCXR-3 dataset distribution. The dataset has a multinational cohort of over 16,600 patients. The whole dataset is split into training and testing sets by the dataset authors. The training dataset has 13992 negative and 16490 positive images. The testing dataset has 200 negative and positive images, respectively.

During the implementation of our experiment, we configured the image size to 224×224 pixels, while normalizing the image pixel values within the range of 0 to 255. Under normal circumstances, in order to provide an unbiased evaluation of a model, cross-validation is typically conducted during the training process. However, as demonstrated in Table 1, there exists an inequality in both image and patient distributions. There are 15994 images positive for Covid-19, derived from 2808 Covid-19 positive patients, indicating multiple X-ray images per patient in the dataset (approximately 6 images per

patient). Meanwhile, the dataset authors have not provided the specific subject information, precluding the possibility of conducting subject-level cross-validation. Concurrently, cross-validation at the image level would result in data leakage. Therefore, we chose to adopt the train/test split as defined by the dataset authors in these particular circumstances.

2.2. Vision Transformer

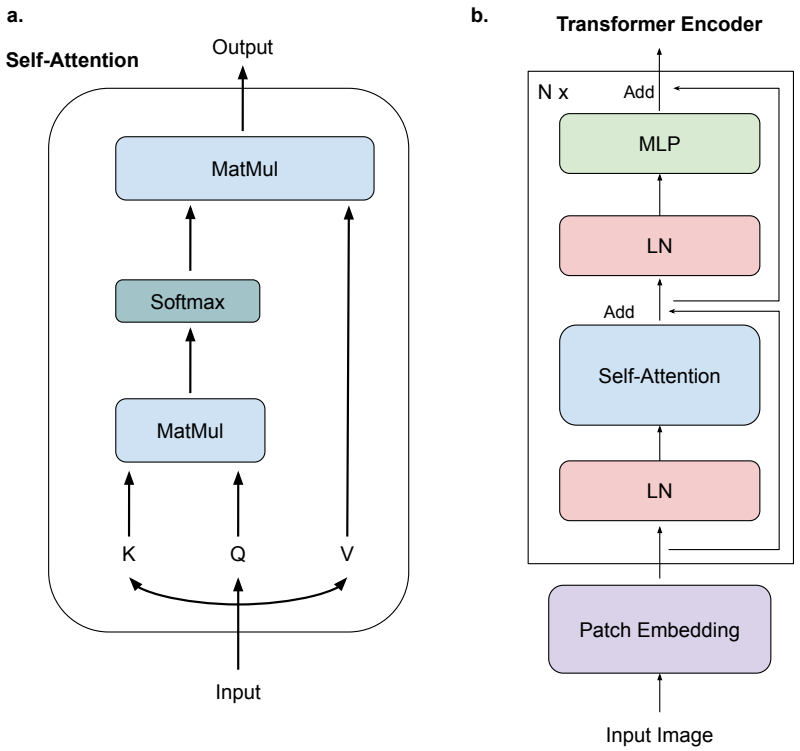

Figure 2. The structure of Vision Transformer. Sub-figure (a) illustrates the structure of the self-attention module. Sub-figure (b) shows the architecture of the vision transformer encoder.

Since ViT is built upon the self-attention mechanism and many works adopt multi-head attention in the implementation, we first introduce the basics of the attention mechanism, then describe the ViT architecture.

The attention mechanism includes three inputs: a query (Q), a key (K), and a value (V). The attention operation is defined as the equation 1:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Afterward, the multi-head attention is defined as equation 2:

$$\begin{aligned} MultiHeadAttn(Q, K, V) &= Concat(head_1, ..., head_n)W^o \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

where W_i^Q, W_i^K, and W_i^V are learnable projections matrices.

Figure 2 illustrates the ViT structure, which has a patch embedding module and $N\times$ stacked transformer encoder blocks. Each transformer encoder block contains a multi-head attention (MSA), two layer normalization (LN) [40], and a multi-layer perceptron (MLP).

In the implementation, the input image I is first transformed into a series of patch embeddings:

$$z_0 = PatchEmbedding(I) \tag{3}$$

The patch embeddings are forwarded through the transformer encoder under the following operations:

$$z^l = MLP(Norm(MSA(Norm(z^{l-1})))) \quad (4)$$

where z^{l-1} and z^l are the l^{th} transformer encoder block input and output, respectively.

2.3. MAE

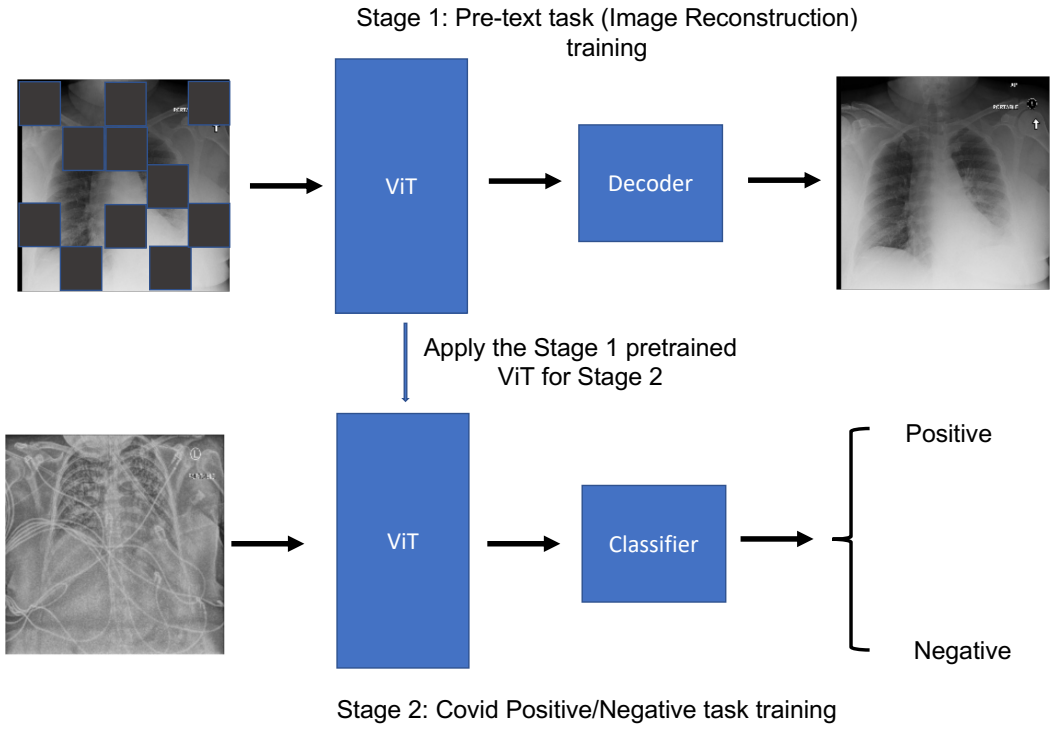

Figure 3. The workflow of the MAE method on the Covid-19 classification task. There are two stages for MAE training. The first stage is the image reconstruction pre-training stage, with the ViT backbone as the image encoder. The second stage is a fine-tuning stage, with the ViT backbone as the feature extractor for the labeled images.

Originally, MAE is a proficient self-supervised learning method employed in Natural Language Processing (NLP) tasks. Building upon its initial foundations, the MAE has significantly broadened its application scope beyond NLP field, marking its presence in the field of computer vision. This widening of its application is facilitated by the universality of its core principle - the strategic "masking" operation. This operation forms the core of the self-supervised learning methodology, by selectively omitting sections of input data to create a challenge for the model to reconstruct these masked elements. This process allows the model to develop a robust understanding of the intrinsic structure and properties of the target dataset, optimizing its ability to generate representative features during the pre-training phase. This initial phase is followed by a fine-tuning stage, which utilizes labeled input data to further refine the model's comprehension of the dataset, thereby improving its overall performance. This two-tiered approach equips the model with the essential tools to tackle novel data and perform reliably on the target dataset. The successful adaptation of the MAE methodology to the realm of computer vision was achieved by employing techniques parallel to its NLP counterpart. Images are decomposed into a multitude of patches, a subset of which are randomly masked. Subsequently, the model is trained to perform the reconstruction pre-training task, effectively learning to predict the obscured sections of the image. Following this, a fine-tuning phase is undertaken with labeled data, ensuring the model's efficient performance on the target dataset.

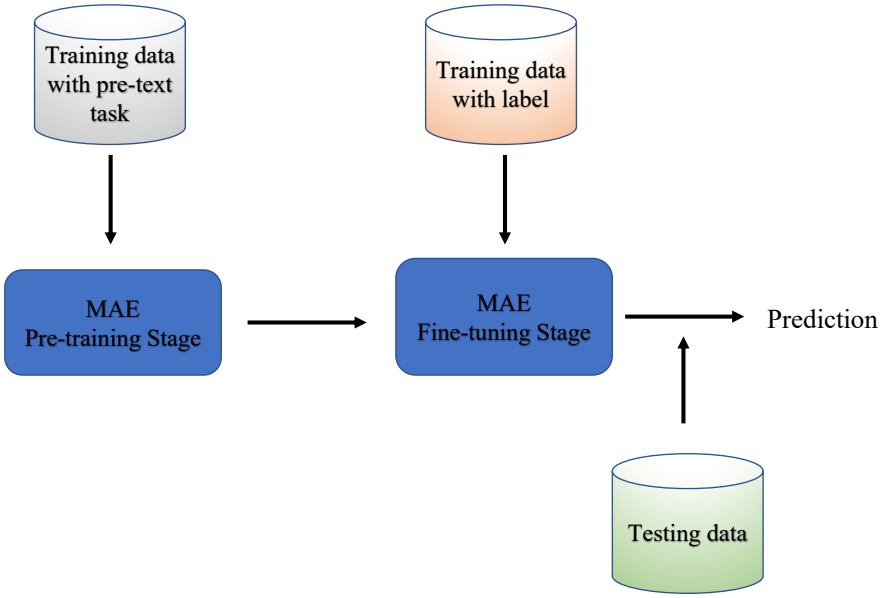

Figure 4. The block presentation of the MAE pipeline.

When considering the choice of backbone for the MAE model, ViT emerges as an optimal option when compared to the Convolutional Neural Network (CNN). As previously mentioned, the ViT architecture offers distinct advantages. One notable feature of the ViT model is its initial operation, where the input image is segmented into various patches. This patch-based approach can easily utilize the masking of random regions. Given these advantages, it is judicious to adopt a ViT architecture as the backbone for the MAE model.

Figure 3 illustrates the comprehensive workflow of a self-supervised learning system based on the MAE, encompassing two essential stages: the pretext task of image reconstruction and the subsequent fine-tuning stage. When it comes to the architecture of the model during the pre-training stage, it incorporates a Vision Transformer (ViT) as both the encoder and decoder. Serving as an encoder, the ViT is applied to mask certain segments of the input image patches. In its role as a decoder, it is tasked with the restoration of the masked patches. Upon transitioning to the fine-tuning phase, the pre-trained ViT encoder is trained further with samples and labels from the target dataset. In the context of our specific implementation, we elected to employ ViT-small as the encoder and the standard decoder within the framework of the MAE. The overall pipeline, founded on block representation, is presented in Figure 4.

2.4. Loss Function

Since our work concentrates on binary classification, the overall loss function is a binary cross-entropy. For a chest X-ray image V with label l and probability prediction $p(l|V)$, the loss function is:

$$loss(l, V) = l \log(p(l|V)) + (1 - l) \log(1 - p(l|V)) \quad (5)$$

where the label $l = 0$ indicates a negative sample and $l = 1$ indicates a positive sample, respectively.

2.5. Implementation and Metrics

We implemented the experiment models using PyTorch. We trained and tested the models based on the default setting of the dataset. For the pre-trained baseline, the model is pre-trained on ImageNet [14]. For the model training, we set the batch size to 16. Adam optimizer [41] with $beta1 = 0.9$, $beta2 = 0.999$, and a learning rate of 1×10^{-4} was used during the training. For the SSL model, we pre-trained the model with 100 epochs. In the fine-tuning stage, we trained all the models for 40 epochs.

To evaluate the performance of our model, we used accuracy (Acc), area under the curve of Receiver Operating Characteristics (AUC), F1 score (F1), Precision, Recall, and Average Precision (AP) as our evaluation metrics. We evaluated the training computation cost by the average epoch training time (e-Time). The accuracy is calculated with the following Eq. 6:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP is the True Positive, TN is the True Negative, FP is the False Positive, and FN is the False Negative.

The precision is calculated by the following Eq. 7:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

The recall is calculated by the following Eq. 8:

$$recall = \frac{TP}{TP + FN} \tag{8}$$

The F1-score is calculated by the following Eq. 9:

$$F1 = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{9}$$

AUC curves compare the true positive rate and the false positive rate at different decision thresholds. AP summarizes a precision-recall curve as the weighted mean of precision achieved at each threshold.

3. Results

3.1. Model performance increasing by MAE

Table 2. The performance of different training strategies over the ViT model. ViT-MAE model outperforms the other two training strategies.

Type	Acc	AUC	F1	Precision	Recall	AP
DenseNet121	0.9775	**0.9970**	0.9771	0.9948	0.96	0.9750
ResNet50	0.9650	0.9969	0.9641	0.9894	0.94	0.9601
ViT-scratch	0.7075	0.7808	0.7082	0.7065	0.7100	0.6466
ViT-pretrain	0.9350	0.9783	0.9340	0.9484	0.9200	0.9125
ViT-MAE	**0.9850**	0.9957	**0.9850**	**0.9950**	**0.9850**	**0.9859**

We conducted training experiments on the ViT model architecture using three different approaches: ViT-scratch, ViT-pretrain, and ViT-MAE. In the ViT-scratch approach, the ViT model was trained directly on the medical image data. The ViT-pretrain approach involved fine-tuning a pre-trained ViT model on ImageNet using the medical image data. ViT-MAE refers to training the ViT model using the MAE pipeline. Accuracy was chosen as the performance metric. As depicted in Table 2, ViT-MAE achieved a remarkable accuracy of 0.985 in Covid-19 positive/negative detection, surpassing the other approaches (ViT-scratch accuracy=0.7075 and ViT-pretrain accuracy=0.9350) on the same dataset. To further compare ViT-MAE with CNN models, namely ResNet50 and DenseNet121, we conducted additional experiments. It was observed that ViT-MAE outperformed both ResNet50 and DenseNet121 in terms of all metrics, except for AUC where the difference was minimal. We think this minimal difference is due to 1) the model experiment's randomness, a common characteristic of machine learning models, and 2) the size of the test dataset. The size of the test dataset would overestimate/underestimate the model. In our experiments, compared

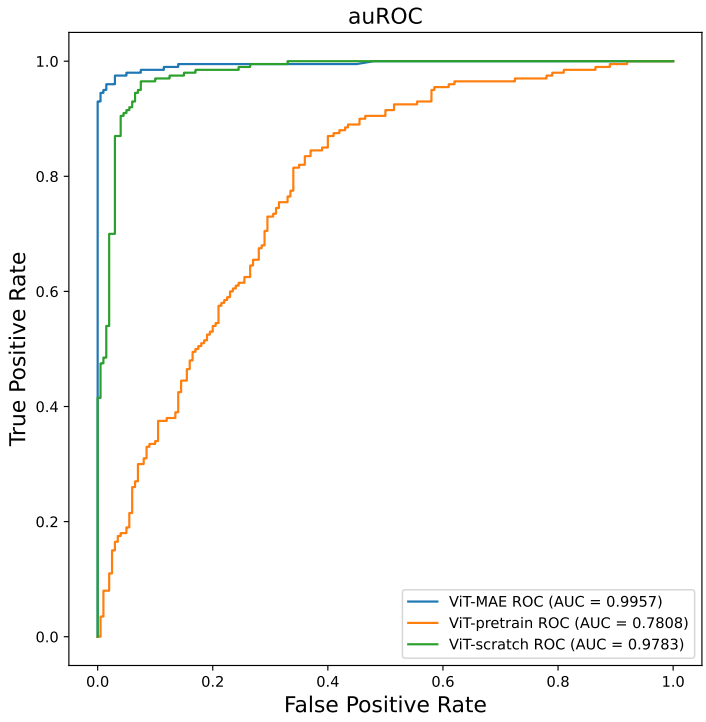

Figure 5. The AUC plot of the different training strategies for ViT model.

with the training dataset, the test dataset is relatively small with only 400 images, which may overestimate the models. However, in terms of the same ViT backbone, we think ViT-MAE models exhibit comparable performance in our experiments. Figure 6 illustrates the AUC curves for the three training approaches, clearly demonstrating that ViT-MAE outperforms the other strategies in terms of AUC performance.

We studied statistical tests that compare the ViT-MAE performance with ViT-scratch and ViT-pretrain. The metric chosen to evaluate their performance was accuracy. To ensure robustness, we conducted four independent experiments for each ViT model, employing different random seeds. The statistical summary of the three pretraining methods yielded the following mean and standard deviation values: For ViT-scratch, the mean was 0.7135 with a standard deviation of 0.0142. For ViT-pretrain, the mean was 0.9293 with a standard deviation of 0.0207. Lastly, for ViT-MAE, the mean was 0.9775 with a standard deviation of 0.006. In order to assess the significance of the differences in performance, we conducted two t-tests: ViT-scratch vs. ViT-MAE and ViT-pretrain vs. ViT-MAE. The resulting p-values for the two group t-tests were found to be less than 0.001 and 0.02, respectively. Our analysis revealed that ViT-MAE significantly outperformed ViT-scratch, indicating the critical influence of the training strategy on model performance. Additionally, we observed a relatively narrow performance gap between ViT-MAE and ViT-pretrain. These findings suggest that while ViT-MAE exhibits superior performance compared to ViT-scratch, the disparity in performance between ViT-MAE and ViT-pretrain is comparatively smaller.

3.2. Mask ratio influence on MAE performance

Since the pre-training of ViT-MAE is a reconstruction task, the mask ratio of the input image is a parameter that may affect the final performance. In this section, we study the mask ratio influence on ViT-MAE training. Table 3 shows the performance of different mask ratios over the MAE pre-training stage. Figure 7 illustrates the AUC curves of different mask ratios. The 40% percentage mask ratio outperforms the other mask ratio situations

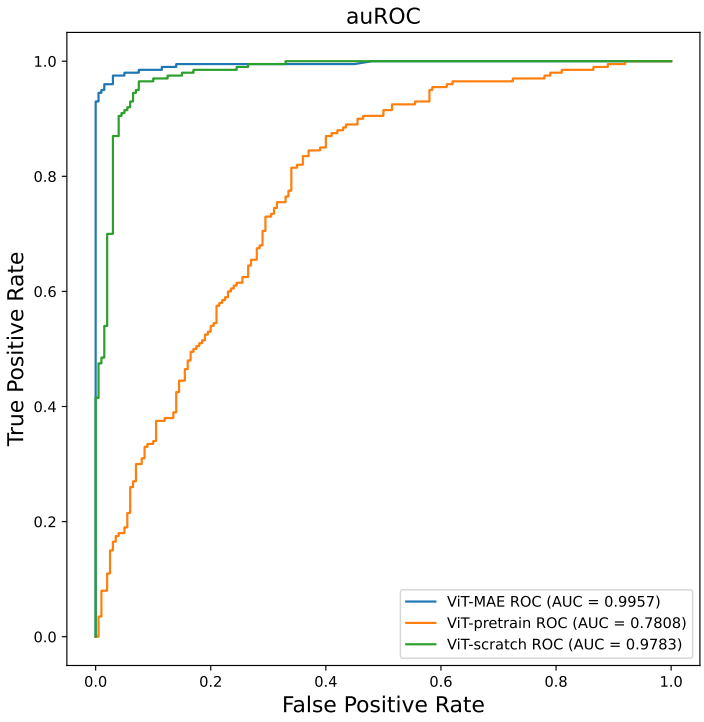

Figure 6. The AUC plot of the different training strategies for ViT model.

Table 3. The performance of different mask ratios over the MAE pre-training stage. The pre-training *maskratio* = 0.4 of MAE outperforms the other pre-training strategies.

Ratio	Acc	AUC	F1	Precision	Recall	AP
0.4	0.9850	0.9957	0.9850	0.9850	0.9850	0.9559
0.5	0.9100	0.9653	0.9086	0.9277	0.8950	0.8783
0.6	0.8875	0.9579	0.8819	0.9282	0.8400	0.8597
0.7	0.8900	0.9502	0.8894	0.8939	0.8850	0.8486
0.8	0.8925	0.9516	0.8900	0.9110	0.8700	0.8576

with Acc= 0.9850 and AUC=0.9957. The mask ratio result indicates that a large mask ratio may decrease the final performance for the medical image dataset, while the large mask ratio (mask ratio=0.75) shows good performance in the natural dataset. We think this may due to the difference between the medical and natural datasets, and the reconstruction results on the medical image may not show better performance than the natural image. We prove our thoughts in section 3.5.

3.3. MAE performance on the limited training dataset

One advantage of self-supervised learning is that people can use a small labeled dataset to train a large DL model well. To explore the potential of SSL, we conduct experiments on the limited training dataset. We randomly split the partial training dataset to train our model as the from 10% to 90% percentage. Table 4 shows the performance under different percentage splitting, and Figure 8 shows the AUC curves of different percentage situations. It appears that using only 30% of the training dataset is fair enough to achieve better performance than that of ViT-pretrain scenario with the whole training dataset (94.25 vs. 93.5). Meanwhile, it is straightforward that increasing the labeled training percentage will

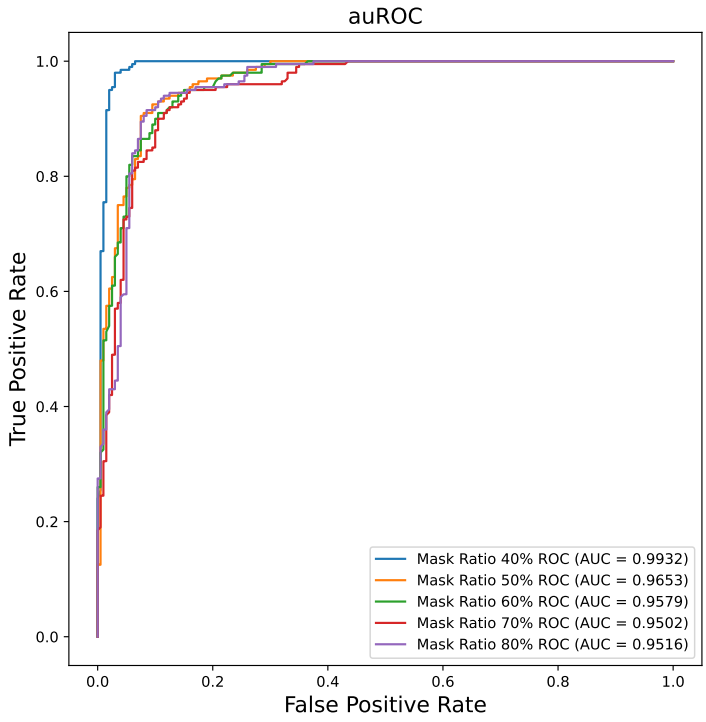

Figure 7. The AUC plot of the different mask ratios for ViT-MAE pre-training.

Table 4. The performance of different percentage of training dataset over the MAE pre-training stage. The pre-training mask ratio=0.4 of MAE outperforms the other pre-training strategies.

Percentage(%)	Acc	AUC	F1	Precision	Recall	AP
10	0.8800	0.9394	0.8776	0.8958	0.8600	0.8404
20	0.8925	0.9588	0.8877	0.9290	0.8500	0.8646
30	0.9425	0.9772	0.9415	0.9585	0.9250	0.9242
40	0.9600	0.9866	0.9602	0.9554	0.9650	0.9395
50	0.9675	0.9879	0.9673	0.9746	0.9600	0.9556
60	0.9650	0.9876	0.9645	0.9794	0.9500	0.9554
70	0.9675	0.9896	0.9669	0.9845	0.9500	0.9602
80	0.9775	0.9978	0.9771	0.9948	0.9600	0.9750
90	0.9825	0.9944	0.9823	0.9949	0.9700	0.9800

contribute to better performance of the model. The promising results provide the potential training procedure for small medical dataset training on DL models.

3.4. Visualization of MAE on image reconstruction

We present a visualization of the X-ray image reconstruction phase. As demonstrated in Figure 9, the depiction includes both Covid and Normal subjects, showcasing the original, masked, and reconstructed images. Upon visual comparison, it can be observed that the reconstructed images are relatively coarse. However, it should be emphasized that our primary objective is not to achieve pixel-perfect image reconstruction but to ensure that the deep learning model's parameters are properly initialized for the fine-tuning process on the specific dataset. Concurrently, the rough outcome of the reconstruction task implies that increasing the mask ratio will not contribute to enhanced model performance during the

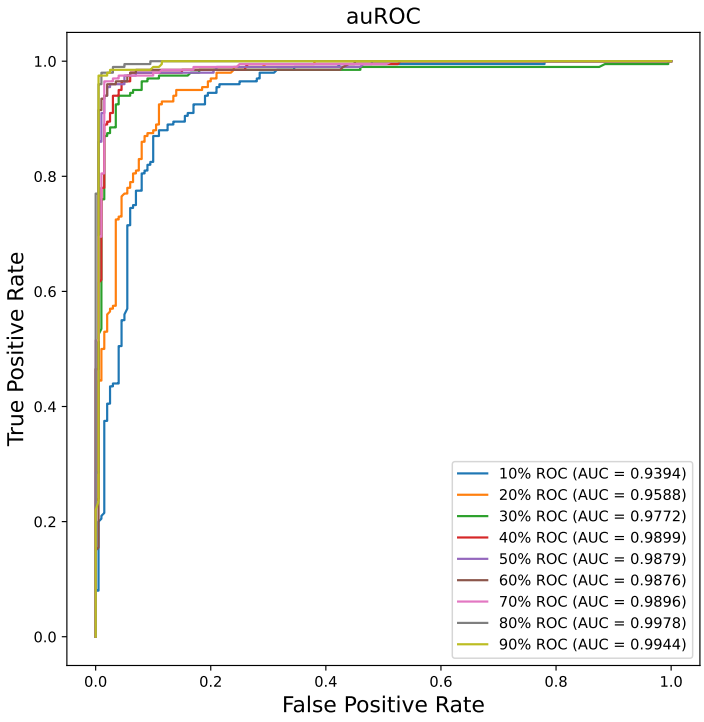

Figure 8. The AUC plot of the different percentages of training dataset for ViT-MAE model.

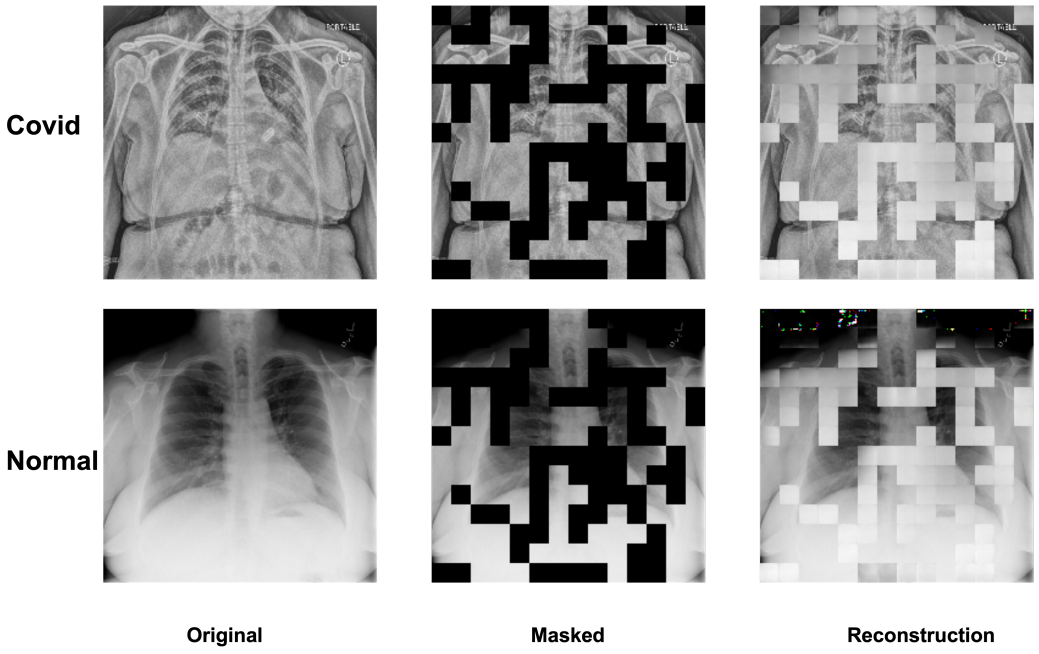

Figure 9. The visualization of the MAE for the image reconstruction pre-training. From the left column to the right are the original input image, the random masked image, and the reconstruction image. Even though the final reconstruction is not well-defined, the target of the pre-training stage is boosting the initial parameters of the ViT model.

fine-tuning stage. This is due to the model's inability to extract additional learning during the reconstruction stage.

4. Discussion

Our study involved an in-depth exploration of the MAE through a comprehensive series of experiments utilizing a publicly available Covid-19 Chest X-ray image dataset. Our study is innovative as we applied MAE to an X-ray imaging dataset for COVID-19 diagnosis, which has not been reported before. Further, we demonstrated that MAE exhibits remarkable efficiency when applied to labeled data, delivering comparable performance to utilizing only 30% of the original training dataset. The findings may have profound implications for various diseases diagnosis with the limited imaging dataset in the future, given that we showed that the accuracy can be maintained even with a reduced, smaller dataset. Our findings also yield several significant insights on MAE application in medical imaging as follows: first, enhanced model performance, by leveraging self-supervised learning with MAE, we observed notable improvements in model performance compared to alternative training methods. This underscores the efficacy of MAE in the context of medical image analysis. Second, the impact of masked ratio, The performance of the MAE model on medical images was found to be influenced by the masked ratio employed during training. Notably, we achieved optimal results with a masked ratio of 0.4 in our implementation. This indicates the importance of carefully selecting the appropriate ratio to achieve the best performance. Finally, labeled data efficiency: Our study demonstrates that MAE operates as a labeled data-efficient model, showcasing comparable performance even when trained on a partial dataset. This finding highlights the potential of MAE in situations where acquiring large quantities of labeled data may be challenging or resource-intensive.

In our implementation, we utilized Vision Transformer (ViT) as the DL model. Compared to traditional CNN models, ViT has shown promising performance across a range of tasks, but it is prone to being data-hungry during the training phase. We conducted experiments on the same DL model and training setting but with different training strategies: ViT-scratch, ViT-pretrain, and ViT-MAE. The results demonstrate the efficacy of self-supervised learning, yielding an accuracy of 0.985 and an AUC of 0.9957. Meanwhile, we compared the performance of the MAE model with the CNN-based models [42,43], which shows MAE outperforms the CNN models.

In our experiments, we explored the association between the mask ratio, a hyperparameter in the masked token reconstruction task, and the model's performance. We set the mask ratio from 0.4 to 0.8 and found that increasing the mask ratio led to a decrease in performance, which is different from the mask ratio result (0.75) of MAE on natural images [30]. To explain this trend, we visualized the original images, masked images, and reconstructed images. Comparing the original and reconstructed images, we observed that the reconstructed images were blurrier. The goal of reconstruction pretraining is to initialize the model parameters and enhance the model's understanding of the medical dataset. However, a high mask ratio may hinder the reconstruction process and weaken the model's understanding ability. Therefore, the mask ratio is a crucial factor in practical implementation.

To further highlight the advantages of self-supervised learning, we conducted limited dataset experiments. We randomly sampled the training dataset from 10% to 90% and applied the sample to conduct the reconstruction pretraining and fine-tuning of the model. The results strongly indicate the advantage of self-supervised learning on limited data. For example, using only a 30% sample of the training dataset, the ViT-MAE model still achieved 0.9425 accuracy, comparable to the performance of the ViT-pretrain model using the entire training dataset. This is particularly important in clinical applications, where datasets are often limited. Training large DL models on limited data can be challenging and easily overfitting due to the large number of parameters in the model. By this pretraining stage, the model can learn the good representation of the target dataset. Compared to the pretained model by natural image dataset, such as the ImageNet dataset, the pretraining stage MAE model has a narrow gap for the target dataset (i.e. small medical dataset) and is suitable for the later fine-tuning stage. Therefore, the MAE is suitable for the limited dataset. Additionally, using a smaller training dataset to train a large DL model reduces

the cost of labeling, as traditional supervised DL training requires a large labeled training dataset to ensure model convergence. However, data labeling can be another issue when the dataset size is large, as in the case of the ImageNet dataset with one million images. Furthermore, medical image labeling often requires professional domain knowledge, such as an X-ray radiologist, to ensure accurate labeling.

The limitations of our study include the focus on Covid-19 alone. For future work, we plan to extend our work in two directions: first, we will extend the MAE model to handle 3D medical images, such as 3D brain imaging for Alzheimer's Disease [44–46]; second, we will explore the potential of the MAE for other tasks, such as image segmentation and localization [6,47], beyond image classification.

In conclusion, we applied MAE to the X-ray imaging dataset for COVID-19 diagnosis and illustrated that MAE exhibits remarkable efficiency when applied to labeled data, delivering comparable performance to utilizing only 30% of the original training dataset. Overall, our findings highlight the significant performance enhancement achieved by using MAE, particularly when working with limited datasets. This approach holds profound implications for future disease diagnosis, especially in scenarios where imaging information is scarce.

Author Contributions: X.X., N. J., and A.-L.L.; Conceptualization. X.X., N. J.; Methodology. X.X.; Software. X.X and G.L.; Validation. X.X., N. J., and A.-L.L.; Formal analysis. A.-L.L.; Investigation. N. J. and A.-L.L.; Resources. X.X.; Data curation. X.X. and A.-L.L.; Writing-original draft. G. L., C.W, N. J., and A.-L.L.; Review Editing. X.X.; Visualization; N. J., and A.-L.L.; Supervision. A.-L.L.; Project administration and Funding acquisition. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by NIH grants R01AG054459 and RF1AG062480 to A.-L.L.

Data Availability Statement: Data used in this article is from the dataset: COVIDxCXR-3. The collected data are available here: https://www.kaggle.com/datasets/andyczhao/covidx-cxr2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114.
2. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
4. Xing, X.; Peng, C.; Zhang, Y.; Lin, A.L.; Jacobs, N. AssocFormer: Association Transformer for Multi-label Classification **2022**.
5. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
9. Ranjbarzadeh, R.; Jafarzadeh Ghoushchi, S.; Anari, S.; Safavi, S.; Tataei Sarshar, N.; Babaee Tirkolaee, E.; Bendechache, M. A deep learning approach for robust, multi-oriented, and curved text detection. *Cognitive computation* **2022**, pp. 1–13.
10. Anari, S.; Tataei Sarshar, N.; Mahjoori, N.; Dorosti, S.; Rezaie, A. Review of deep learning approaches for thyroid cancer diagnosis. *Mathematical Problems in Engineering* **2022**, *2022*, 1–8.
11. Xing, X.; Liang, G.; Zhang, Y.; Khanal, S.; Lin, A.L.; Jacobs, N. Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, 2022, pp. 1–4.
12. Xing, X.; Rafique, M.U.; Liang, G.; Blanton, H.; Zhang, Y.; Wang, C.; Jacobs, N.; Lin, A.L. Efficient Training on Alzheimer's Disease Diagnosis with Learnable Weighted Pooling for 3D PET Brain Image Classification. *Electronics* **2023**, *12*, 467.

13. Liang, G.; Xing, X.; Liu, L.; Zhang, Y.; Ying, Q.; Lin, A.L.; Jacobs, N. Alzheimer's disease classification using 2d convolutional neural networks. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021, pp. 3008–3012.
14. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the CVPR09, 2009.
15. Ying, X. An overview of overfitting and its solutions. In Proceedings of the Journal of physics: Conference series. IOP Publishing, 2019, Vol. 1168, p. 022022.
16. Wang, X.; Liang, G.; Zhang, Y.; Blanton, H.; Bessinger, Z.; Jacobs, N. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology* **2020**, *17*, 796–803.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60*, 84–90.
18. Xing, X.; Liang, G.; Blanton, H.; Rafique, M.U.; Wang, C.; Lin, A.L.; Jacobs, N. Dynamic image for 3d mri image alzheimer's disease classification. In Proceedings of the Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I. Springer, 2021, pp. 355–364.
19. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE* **2020**, *109*, 43–76.
20. Durand, T.; Mordan, T.; Thome, N.; Cord, M. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 642–651.
21. Zhou, Z.H. A brief introduction to weakly supervised learning. *National science review* **2018**, *5*, 44–53.
22. Liang, G.; Wang, X.; Zhang, Y.; Jacobs, N. Weakly-supervised self-training for breast cancer localization. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020, pp. 1124–1127.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, [arXiv:cs.CV/2010.11929].
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
25. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.
26. Goyal, P.; Caron, M.; Lefaudeux, B.; Xu, M.; Wang, P.; Pai, V.; Singh, M.; Liptchinsky, V.; Misra, I.; Joulin, A.; et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988* **2021**.
27. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* **2018**.
28. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.
29. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* **2020**, *33*, 21271–21284.
30. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
31. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports* **2020**, *10*, 19549. https://doi.org/10.1038/s41598-020-76550-z.
32. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv 2003.11597* **2020**.
33. COVID-19 Radiography Database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.
34. Actualmed COVID-19 Chest X-ray 71 Dataset Initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset.
35. Figure 1 COVID-19 Chest X-ray Dataset Initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset.
36. RSNA Pneumonia Detection Challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.
37. RSNA International COVID-19 Open Radiology Database. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70230281.
38. BIMCV-COVID19+. https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/.
39. COVID-19-NY-SBU. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=89096912.
40. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv preprint arXiv:1607.06450* **2016**.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
42. Joaquin, A. Using deep learning to detect pneumonia caused by ncov-19 from x-ray images. *Towards data science* **2020**.
43. Hasan, N.; Bao, Y.; Shawon, A.; Huang, Y. DenseNet convolutional neural networks application for predicting COVID-19 using CT image. *SN computer science* **2021**, *2*, 389.

44. Hammond, T.C.; Xing, X.; Wang, C.; Ma, D.; Nho, K.; Crane, P.K.; Elahi, F.; Ziegler, D.A.; Liang, G.; Cheng, Q.; et al. β-amyloid and tau drive early Alzheimer's disease decline while glucose hypometabolism drives late decline. *Communications biology* **2020**, *3*, 352.
45. Hammond, T.C.; Xing, X.; Yanckello, L.M.; Stromberg, A.; Chang, Y.H.; Nelson, P.T.; Lin, A.L. Human Gray and White Matter Metabolomics to Differentiate APOE and Stage Dependent Changes in Alzheimer's Disease. *Journal of cellular immunology* **2021**, *3*, 397.
46. Ying, Q.; Xing, X.; Liu, L.; Lin, A.L.; Jacobs, N.; Liang, G. Multi-modal data analysis for alzheimer's disease diagnosis: An ensemble model using imagery and genetic features. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021, pp. 3586–3591.
47. Zhao, Y.; Zeng, K.; Zhao, Y.; Bhatia, P.; Ranganath, M.; Kozhikkavil, M.L.; Li, C.; Hermosillo, G. Deep learning solution for medical image localization and orientation detection. *Medical Image Analysis* **2022**, *81*, 102529.