# MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models

**Hari Sowrirajan***, **Jingbo Yang***, **Andrew Y. Ng**, **Pranav Rajpurkar**
Department of Computer Science
Stanford University
{hsowrira,jingboy,pranavsr}@stanford.edu

## Abstract

Self-supervised approaches such as Momentum Contrast (MoCo) can leverage unlabeled data to produce pretrained models for subsequent fine-tuning on labeled data. While MoCo has demonstrated promising results on natural image classification tasks, its application to medical imaging tasks like chest X-ray interpretation has been limited. Chest X-ray interpretation is fundamentally different from natural image classification in ways that may limit the applicability of self-supervised approaches, including that (1) classification depends on differences in a small number of pixels, (2) X-rays are large and grayscale, (3) there are far fewer unlabeled chest X-ray images than natural images. In this work, we investigate whether MoCo-pretraining leads to better representations or initializations for chest X-ray interpretation. We conduct MoCo-pretraining on CheXpert, a large labeled dataset of X-rays, followed by supervised fine-tuning experiments on the pleural effusion task. Using 0.1% of labeled training data, we find that a linear model trained on MoCo-pretrained representations outperforms one trained on representations without MoCo-pretraining by an AUC of 0.096 (95% CI 0.061, 0.130), indicating that MoCo-pretrained representations are of higher quality. Furthermore, a model fine-tuned end-to-end with MoCo-pretraining outperforms its non-MoCo-pretrained counterpart by an AUC of 0.037 (95% CI 0.015, 0.062) with the 0.1% label fraction. These AUC improvements are observed for all label fractions for both the linear model and an end-to-end fine-tuned model with the greater improvements for smaller label fractions. Finally, we observe similar results on a small, target chest X-ray dataset (Shenzhen dataset for tuberculosis) with MoCo-pretraining done on the source dataset (CheXpert), which suggests that pretraining on unlabeled X-rays can provide transfer learning benefits for a target task. Our study demonstrates that MoCo-pretraining provides high-quality representations and transferable initializations for chest X-ray interpretation.

## 1 Introduction

Self-supervised approaches such as Momentum Contrast (MoCo) (He et al., 2019a; Chen et al., 2020d) can leverage unlabeled data to produce pretrained models for subsequent fine-tuning on labeled data. While MoCo and other self-supervised learning methods have demonstrated promising results on natural image classification tasks (Chen et al., 2020c; He et al., 2019a; Chen et al., 2020a;b), their application to medical imaging settings has been limited (Raghu et al., 2019; Cheplygina et al., 2019).

Chest X-ray is the most common imaging tool used in practice, and is critical for screening, diagnosis, and management of diseases. The recent introduction of large datasets (size 100k-500k) of chest X-rays (Irvin et al., 2019; Johnson et al., 2019; Bustos et al., 2020) has driven the development of deep learning models that can detect diseases at a level comparable to radiologists (Rajpurkar et al., 2020). Because there is an abundance of unlabeled chest X-ray data (Raoof et al., 2012), self-supervised approaches represent a promising avenue for improving chest X-ray interpretation models.

*Equal contribution

Chest X-ray interpretation is fundamentally different from natural image classification in that (1) disease classification may depend on abnormalities in a small number of pixels, (2) data attributes for chest X-rays differ from natural image classification in that X-rays are larger, grayscale and have similar spatial structures across images (always either anterior-posterior, posterior-anterior, or lateral), (3) there are far fewer unlabeled chest X-ray images than natural images. These differences may limit the applicability of self-supervised approaches, which were developed for natural image classification settings, to chest X-ray intepretation. For instance, MoCo data augmentations including random crops and blurring may eliminate disease-covering parts from an augmented image, while color jittering and random gray scale would not produce meaningful transformations for already grayscale images. Furthermore, given the availability of orders of magnitude fewer chest X-ray images than natural images, and larger image sizes, it remains to be understood whether MoCo-pretraining may improve on the traditional paradigm for automated chest X-ray interpretation, in which models are fine-tuned on labeled chest X-ray images from ImageNet-pretrained weights.

In this work, we investigate whether MoCo-pretraining leads to better representations or initializations than those acquired without MoCo-pretraining for chest X-ray interpretation. We conduct MoCo-pretraining on CheXpert (Irvin et al., 2019) followed by supervised fine-tuning experiments using different fractions of labeled data. In addition to fine-tuning on labeled dataset from the source dataset, we also experiment with transfer learning to another small chest X-ray dataset with a different task (Jaeger et al., 2014). We assess the quality of pretrained representations by fine-tuning a linear classifier on outputs of a frozen base model, and the transferability of pretrained initialization by fine-tuning the model end-to-end. Our main findings are as follows:

1. A linear model trained on MoCo-pretrained representations has higher performance on a chest X-ray interpretation task than one trained without MoCo-pretrained representations, with greater improvements in low labeled data regimes (Figure 2).
2. A model fine-tuned end-to-end with MoCo-pretraining has higher performance than one without MoCo-pretraining for small label fractions, and comparable performance for large label fractions (Figure 3).
3. On a target/external dataset (Shenzhen dataset for tuberculosis) with a small number of labeled samples, a linear model trained on MoCo-pretrained representations acquired from the source dataset (CheXpert) has higher performance than one without MoCo-pretrained representations. However, a model fine-tuned end-to-end on the external dataset with MoCo-pretraining on the source dataset does not have significantly higher performance than one without MoCo-pretraining (Figure 4).

Our study demonstrates that MoCo-pretraining provides high-quality representations and transferable initializations for chest X-ray interpretation.

## 2 Related Work

**Self-supervised Learning** Self-supervision is a form of unsupervised pretraining that uses a formulated pretext task on unlabeled data as the training goal. Handcrafted pretext tasks include solving jigsaw puzzles (Noroozi & Favaro, 2016), relative patch prediction (Doersch et al., 2015) and colorization (Zhang et al., 2016). However, many of these tasks rely on ad-hoc heuristics that could limit the generalization and transferability of learned representations for downstream tasks (Chen et al., 2020c). Contrastive learning of visual representations has emerged as the front-runner for self-supervision and has demonstrated superior performance on downstream tasks. These approaches include frameworks such as MoCo (He et al., 2019a; Chen et al., 2020d), SimCLR (Chen et al., 2020a;b), PIRL (Misra & Maaten, 2020) and FixMatch (Sohn et al., 2020). These approaches learn representations by contrasting positive image pairs against negative pairs and differ in the type of contrastive loss, generation of positive and negative pairs, and sampling method. The pretrained models can subsequently be fine-tuned on labeled data for a particular downstream task.

To evaluate learned representations of self-supervised models, the linear evaluation protocol is commonly used (Oord et al., 2018; Zhang et al., 2016; Kornblith et al., 2019; Bachman et al., 2019), where a linear classifier is trained on a frozen base model, and test performance is used as the metric for representation quality. Additionally, Kornblith et al. (2019) found that, for many natural image datasets, logistic regression models trained on fixed representations in sparse data regimes establish

strong baselines and in some cases outperform models fine-tuned end-to-end, where all layers of the network are allowed to be trained. Nevertheless, most deep learning models are still fine-tuned end-to-end following their pretrained initialization to maximize final performance, even though the underlying representations may shift during training. Kornblith et al. (2019) found that end-to-end models significantly outperform their logistic regression counterparts (when using the full training set) on 179 of 192 of the datasets tested, with generally larger gains for larger datasets. In this work, we perform both linear evaluation and end-to-end fine-tuning experiments to assess representation quality as well as initialization quality.

**Self-supervision for Chest X-rays** There has been limited work applying self-supervision to chest X-ray interpretation tasks. Liu et al. (2019) utilized contrastive learning for chest X-ray diagnosis, but in a supervised fashion where diseased X-rays are explicitly contrasted with healthy ones. Zhou et al. (2020) proposed C2L, which utilizes unsupervised contrastive learning for model pretraining. However, C2L utilizes a cross-entropy function to facilitate computation of contrastive loss in batches of multiple positive and negative pairs, whereas approaches like MoCo and SimCLR directly utilize contrastive loss functions such as InfoNCE, which contrasts a single positive image pair against a batch of negative samples. Furthermore, Zhou et al. (2020) did not assess whether their pretrained models generalize to datasets unseen during pretraining, or how the performance of the models scaled with different amounts of labeled training data.

**ImageNet Transfer For Chest X-ray Interpretation** The dominant computer vision approach of starting with an ImageNet-pretrained model has been proven to be highly effective at improving model performance in diverse settings such as object detection and image segmentation (Tan et al., 2018). Although high performance deep learning models for chest X-ray interpretation use ImageNet-pretrained weights (Kornblith et al., 2019) found that common regularization techniques limit ImageNet transfer learning benefits and that ImageNet features are less general than previously believed. Moreover, He et al. (2019b) showed that randomly-initialized models are competitive with their ImageNet-initialized counterparts on a vast array of tasks with sufficient labeled data, and that pretraining merely speeds up convergence. Raghu et al. (2019) further investigated the efficacy of ImageNet pretraining, observing that simple convolutional architectures are able to achieve comparable performance as larger ImageNet model architectures.

# 3 METHODS

## 3.1 CHEST X-RAY DATASETS AND DIAGNOSTIC TASKS

We use a large source chest X-ray dataset for pretraining and a smaller external chest X-ray dataset for the evaluation of model transferability. The source chest X-ray dataset we select is CheXpert, a large collection of chest X-ray images labeled for the presence or absence of several diseases (Irvin et al., 2019). CheXpert consists of 224k chest X-rays collected from 65k patients. We focus on identifying the presence of pleural effusion, a clinically important condition that has high prevalence in the dataset (with $45.63\%$ of all images labeled as positive or uncertain). In addition, we use the Shenzhen Hospital X-ray set (Jaeger et al., 2014) for evaluation of model transferability to an external target dataset. The Shenzhen dataset contains $662$ X-ray images, of which $336$ ($50.8\%$) are abnormal X-rays that have manifestations of tuberculosis.

## 3.2 MOCO PRETRAINING FOR CHEST X-RAY INTERPRETATION

We apply the MoCo-pretraining procedure to chest X-ray intepretation. MoCo is a form of self-supervision that utilizes contrastive learning, where the pretext task is to maximize agreement between different views of the same image (positive pairs) and to minimize agreement between different images (negative pairs).

Our choice to use MoCo is driven by two constraints in medical imaging AI: (1) large image sizes, and (2) the cost of large computational resources. Compared to other self-supervised frameworks such as SimCLR (Chen et al., 2020c), MoCo requires far smaller batch sizes during pretraining. Chen et al. (2020d) used a batch size of $256$ and achieved comparable performance on ImageNet as the SimCLR implementation, which used a batch size of $4096$; in contrast, SimCLR experienced

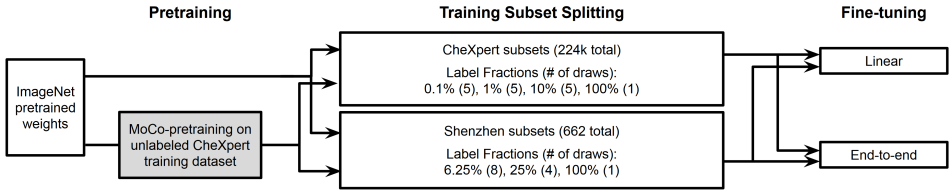


Figure 1: Semi-supervised CheXpert model training pipeline with MoCo as self-supervised training agent.

lower performance at a batch size of 256 (Chen et al., 2020a). MoCo's reduced dependency on mini-batch size is achieved by using a momentum updated queue of previously seen samples to generate contrastive pair encodings. Using MoCo, we are able to conduct experiments on a single NVIDIA GTX 1070 with a batch size of 16 using chest X-ray images of size $320 \times 320$. We perform MoCo-pretraining on the entire CheXpert training dataset. We choose to apply MoCo-pretraining on ImageNet-initialized models to leverage possible convergence benefits (Raghu et al., 2019). Due to the widespread availability of ImageNet-pretrained weights, there is no extra cost to initialize models with ImageNet weights before MoCo-pretraining.

We modify the data augmentation strategy used to generate views suitable for the chest X-ray interpretation task. Data augmentations used in self-supervised approaches for natural images may not port to chest X-rays. For example, random crop and Gaussian blur could change the disease label for an X-ray image or make it impossible to distinguish between diseases. Furthermore, color jittering and random gray scale do not represent meaningful augmentations for grayscale X-rays. Instead, we use random rotation (10 degrees) and horizontal flipping, a set of augmentations commonly used in training chest X-ray models (Irvin et al., 2019; Rajpurkar et al., 2017) driven by experimental findings in the supervised setting and clinical domain knowledge. Future work should investigate the impact of various individual augmentations and their combinations.

The overall training pipeline with MoCo-pretraining and the subsequent fine-tuning with CheXpert and Shenzhen datasets is illustrated in Figure 1. We maintain hyperparameters related to momentum, weight decay, and feature dimension from MoCo (Chen et al., 2020d). Checkpoints from top performing epochs are saved for subsequent checkpoint selection and model evaluation. We select hyperparameters based on performance of linear evaluations. We use two backbones, ResNet18 and DenseNet121, to evaluate the consistency of our findings across model architectures. We experiment with initial learning rates of $10^{-2}$, $10^{-3}$, $10^{-4}$ and $10^{-5}$, and investigate their effect on performance.

### 3.3 Fine-tuning and Evaluation

We fine-tune models on different fractions of labeled training data. We also conduct baseline fine-tuning experiments with ImageNet-pretrained models that were not subjected to MoCo-pretraining. As presented in Figure 1, the label fractions of training sets are $0.1\%$, $1\%$, $10\%$ and $100\%$ for the CheXpert dataset and $6.25\%$, $25\%$, $100\%$ for the external Shenzhen dataset. Fine-tuning experiments on small label fractions are repeated multiple times with different random samples and averaged to guard against anomalous, unrepresentative training splits. For the CheXpert dataset, each of the $0.1\%$, $1\%$ and $10\%$ label fraction sets is drawn $5$ times. For the Shenzhen dataset, the $6.25\%$ label fraction is drawn $8$ times and the $25\%$ label fraction is drawn $4$ times.

To evaluate the transfer of representations, we freeze the backbone model and train a linear classifier on top using the labeled data (MoCo/ImageNet-pretrained *Linear Models*). We also unfreeze all layers and fine-tune the entire model end-to-end using the labeled data to assess transferability on the overall performance (MoCo/ImageNet-pretrained *end-to-end Models*). Our models are fine-tuned using the same configurations as fully-supervised models designed for CheXpert (Irvin et al., 2019), which has determined an optimal batch size, learning rate and other hyper-parameters. To be specific, we use a learning rate of $3 \times 10^{-5}$, batch size of 16 and number of epochs that scale with the size of labeled data. For the CheXpert dataset, these are $220, 95, 41, 18$ epochs for the $4$ label fractions respectively.

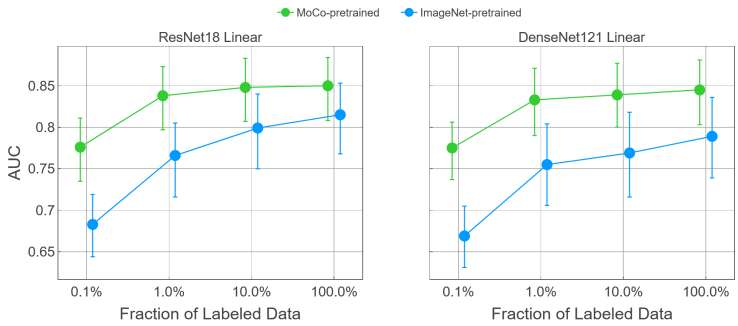


Figure 2: AUC for linear models with MoCo-pretraining is consistently higher than AUC of linear models with ImageNet-pretraining, showing that MoCo-pretraining produces higher quality representations than ImageNet-pretraining does.

### 3.4 Statistical analysis

We compare the performance of the models trained with and without MoCo-pretraining using the area under the receiver operating characteristic curve (AUC). To assess whether MoCo-pretraining significantly changed the performance, we compute the difference in AUC on the test set with and without MoCo-pretraining. The non-parametric bootstrap is used to estimate the variability around model performance. 500 bootstrap replicates from the test set are drawn, and the AUC and difference is calculated for the MoCo-pretrained model and non-Moco pretrained model on these same 500 bootstrap replicates. This produces a distribution for each estimate, and the 95% bootstrap percentile intervals are reported to assess significance at the $p = 0.05$ level.

## 4 Experiments

### 4.1 Transfer Performance of MoCo-pretrained Representations on CheXpert

We investigate whether representations acquired from MoCo-pretraining are of higher quality than those transferred from ImageNet. We visualize the performance of MoCo-pretrained and ImageNet-pretrained linear models at various label fractions in Figure 2 and present the AUC improvements in Table 1.

Trained on small label fractions, the ResNet18 MoCo-pretrained linear model demonstrates statistically significant performance gains over its ImageNet counterpart. With 0.1% label fraction, the improvement in performance is 0.096 (95% CI 0.061, 0.130) AUC; the MoCo-pretrained and ImageNet-pretrained linear models achieve performances of 0.776 and 0.683 AUC respectively. This improvement is also observed for DenseNet121 with 0.1% label fraction: the MoCo-pretrained linear model records an AUC of 0.775 whereas the ImageNet one records an AUC of 0.669, corresponding to an improvement of 0.107 (95% CI 0.075, 0.142) AUC. These findings support the hypothesis that MoCo-representations are of superior quality, and is most apparent when labeled data is scarce.

With larger label fractions, the MoCo-pretrained linear models demonstrate clear yet diminishing improvements over the ImageNet-pretrained linear models. Training with 100% of the labeled data, the ResNet18 MoCo-pretrained linear model achieves an AUC of 0.850 while the ImageNet one achieves an AUC of 0.815, yielding a performance gain of 0.034 (95% CI -0.009, 0.080). Similarly, the DenseNet121 model records an AUC improvement of 0.055 (95% CI 0.008, 0.102) with the 100% label fraction, which is statistically significant but lower than the difference observed with the 0.1% label fraction. Both backbones are observed to have monotonically decreasing performance gains with MoCo as we increase the amount of labeled training data. These results provide evidence that MoCo-pretrained representations retain their quality at all label fractions, but less significantly at larger label fractions.

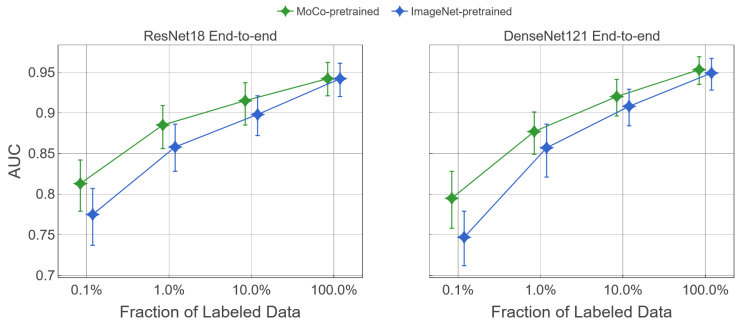


Figure 3: AUC for models fine-tuned end-to-end with MoCo-pretraining is consistently higher than AUC of models fine-tuned end-to-end without MoCo-pretraining, showing that MoCo-pretraining representations are more transferable than representations produced by ImageNet-pretraining only.

| Architecture | MoCo-pretrained | ImageNet-pretrained | 0.1% | 1.0% | 10.0% | 100% |
|---|---|---|---|---|---|---|
| ResNet18 | End-to-End | End-to-End | 0.037( 0.015, 0.062) | 0.027( 0.006, 0.047) | 0.017( 0.003, 0.031) | 0.000(-0.009, 0.009) |
| ResNet18 | Linear Model | Linear Model | 0.096( 0.061, 0.130) | 0.070( 0.029, 0.112) | 0.049( 0.005, 0.094) | 0.034(-0.009, 0.080) |
| ResNet18 | Linear Model | End-to-End | 0.001(-0.024, 0.025) | -0.022(-0.051, 0.009) | -0.050(-0.083, -0.018) | -0.094(-0.127, -0.062) |
| DenseNet121 | End-to-End | End-to-End | 0.048( 0.023, 0.074) | 0.019( 0.001, 0.037) | 0.012( 0.000, 0.023) | 0.003(-0.006, 0.013) |
| DenseNet121 | Linear Model | Linear Model | 0.107( 0.075, 0.142) | 0.078( 0.035, 0.121) | 0.067( 0.023, 0.111) | 0.055( 0.008, 0.102) |
| DenseNet121 | Linear Model | End-to-End | 0.029( 0.002, 0.055) | -0.024(-0.050, -0.003) | -0.070(-0.109, -0.036) | -0.107(-0.141, -0.073) |

Table 1: Table of AUC improvements achieved by MoCo-pretrained models against models without MoCo-pretraining on the CheXpert dataset

### 4.2 Transfer Performance of End-to-End MoCo-pretrained Models on CheXpert

We investigate whether MoCo-pretraining translates to higher performance for models fine-tuned end-to-end. We visualize the performance of the MoCo and ImageNet-pretrained end-to-end models at different label fractions in Figure 3.

We find that MoCo-pretrained end-to-end models outperform their ImageNet-pretrained counterparts more at small label fractions than at larger label fractions. With the $0.1\%$ label fraction, the ResNet18 MoCo-pretrained end-to-end model achieves an AUC of $0.813$ while the ImageNet-pretrained end-to-end model achieves an AUC of $0.775$, yielding a statistically significant AUC improvement of 0.037 (95% CI 0.015, 0.062). The AUC improvement with the $1.0\%$ label fraction is also statistically significant at 0.027 (95% CI 0.006, 0.047). Additionally, the DenseNet121 models also record statistically significant AUC improvements, with the $0.1\%$ label fraction seeing a difference of 0.048 (95% CI 0.023, 0.074) and the $1.0\%$ label fraction reaching a difference of 0.019 (95% CI 0.001, 0.037). However, both pretraining approaches converge to an AUC of $0.942$ with the $100\%$ label fraction.

These results demonstrate that MoCo-pretraining yields performance boosts for end-to-end training, and further substantiate the quality of the pretrained initialization, especially for smaller label fractions. This finding is consistent with Chen et al. (2020a), who also report larger performance gains for self-supervised models trained end-to-end on smaller label fractions of ImageNet.

### 4.3 Transfer Benefit of MoCo-pretraining on an External Dataset

We conduct experiments to test whether MoCo-pretrained chest X-ray representations transfer to a target dataset. Results of these experiments are presented in Figure 4.

We first examine whether MoCo-pretrained linear models improve AUC on the external Shenzhen dataset. With $6.25\%$ label fraction, which is approximately 25 images, the ResNet18 MoCo-pretrained model outperforms the ImageNet-pretrained one by 0.054 (95%, CI 0.024, 0.086). For this comparison, the MoCo-pretrained model reported an AUC of $0.902$ whereas the ImageNet-pretrained model reported an AUC of $0.852$. AUC improvement is lower with the $25\%$ label fraction, at 0.026 (95%, CI -0.001, 0.056). Here, AUC reported for the MoCo-pretrained model is $0.928$

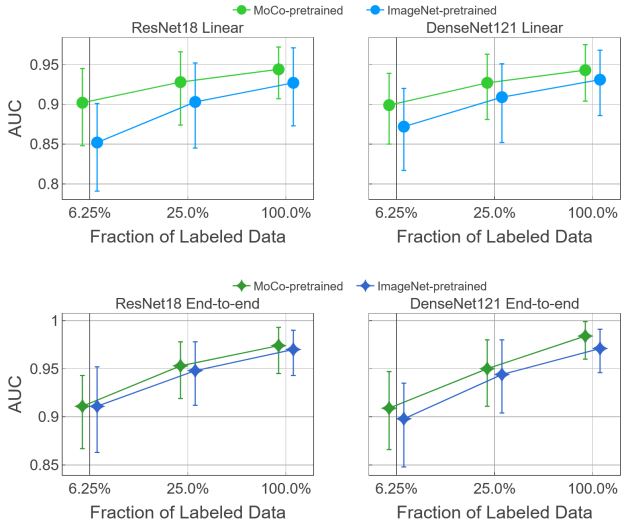


Figure 4: AUC performance on the Shenzhen tuberculosis task for models with and without MoCo-pretraining shows that MoCo pretraining still introduces significant performance improvement despite being fine-tuned on an external dataset.

| Learning Rate | AUC |
|---|---|
| $10^{-2}$ | 0.786 (0.699, 0.861) |
| $10^{-3}$ | 0.908 (0.853, 0.955) |
| $10^{-4}$ | 0.944 (0.907, 0.972) |
| $10^{-5}$ | 0.939 (0.891, 0.975) |

Table 2: Performance of MoCo pretrained ResNet18 on Shenzhen dataset at different pretraining learning rates with $100\%$ label fraction.

and the AUC for the ImageNet-pretrained model is $0.903$. AUC improvement with the $100\%$ label fraction is 0.018 (95% CI -0.011, 0.053). For this comparison, the AUC reported for the MoCo-pretrained model is $0.944$ and the AUC for the ImageNet-pretrained model is $0.927$. This is similar to the trend observed on CheXpert dataset discussed in Section 4.1. These observations suggest that representations learned from MoCo-pretraining are better suited for an external target chest X-ray dataset with a different task than representations learned from ImageNet-pretraining.

Next, we test whether MoCo-pretrained models with end-to-end training also perform well on the external Shenzhen dataset. With the $100\%$ label fraction, the ResNet18 MoCo-pretrained model is able to achieve an AUC of $0.974$, which is higher than the AUC of $0.970$ achieved by its ImageNet-pretrained counterpart. However, the corresponding AUC improvement of only 0.003 (95% CI -0.014, 0.020) is much less than the improvement observed for linear models. Since the Shenzhen dataset is limited in size, it is possible that training end-to-end quickly saturates learning potential at low label fractions. Regardless, the non-zero improvement still suggests that MoCo-pretrained initializations can transfer to an external dataset. This echoes Sun et al. (2019); Erhan et al. (2010), who found that unsupervised pretraining pushes the model towards solutions with better generalization to tasks that are in the same domain.

### 4.4 Sensitivity of MoCo-pretraining to Learning Rate

We investigate whether the quality of MoCo-pretrained representations is dependent on learning rates used for MoCo-pretraining. As shown in Table 2, the best AUC achieved by a linear model base on MoCo-pretraining is $0.944$ at a learning rate of $10^{-4}$. At a higher learning rate of $10^{-2}$, the AUC decreases to $0.786$. At a lower learning rate of $10^{-5}$, the AUC decreases slightly to $0.939$. Our optimal learning rate of $0.0001$ is much less than the $0.03$ used in the original MoCo (He et al., 2019a; Chen et al., 2020d) implementation. From this observation, we deduce that learning rate for self-supervised training may need to be adjusted differently for different data domains.

## 5 Conclusion

We find MoCo-pretraining provides high-quality representations and transferable initializations for chest X-ray interpretation. Despite many differences in the data and task properties between natural image classification and chest X-ray interpretation, we only make a small set of modifications to MoCo pretraining for successful application: we change the data augmentations to mimic those

used in supervised chest X-ray intepretation, and use a much smaller learning rate in the pretraining stage. This suggests the possibility for broad application of self-supervised approaches beyond natural image classification settings.

To the best of our knowledge, this work is the first to show the benefit of MoCo-pretraining across label fractions for chest X-ray interpretation, and also investigate representation transfer to a target dataset. All of our experiments are run on a single NVIDIA GTX 1070, demonstrating accessibility of this method. Our success in demonstrating improvements in model performance over the traditional supervised learning approach, especially on low label fractions, may be broadly extensible to other medical imaging tasks and modalities, where high-quality labeled data is expensive, but unlabeled data is increasingly easier to access.

## REFERENCES


Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

Aurelia Bustos, A. Pertusa, J. M. Salinas, and M. Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

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

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

Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, (11):625–660, 2 2010.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019a.

Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019b.

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

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. URL http://arxiv.org/abs/1901.07042.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10632–10641, 2019.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Phil Chen, Amirhossein Kiani, Jeremy Irvin, Andrew Y Ng, and Matthew P Lungren. Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379*, 2020.

Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pp. 270–279. Springer, 2018.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. *arXiv preprint arXiv:2007.07423*, 2020.

# A Appendix

| Pretraining | Architecture | Fine Tuning | 0.1% | 1.0% | 10.0% | 100.0% |
|---|---|---|---|---|---|---|
| MoCo | ResNet18 | Linear Model | 0.776(0.735, 0.811) | 0.838(0.797, 0.873) | 0.848(0.807, 0.883) | 0.850(0.808, 0.884) |
| ImageNet | ResNet18 | Linear Model | 0.683(0.644, 0.719) | 0.766(0.716, 0.805) | 0.799(0.750, 0.840) | 0.815(0.768, 0.853) |
| MoCo | DenseNet121 | Linear Model | 0.775(0.737, 0.806) | 0.833(0.790, 0.871) | 0.839(0.800, 0.877) | 0.845(0.803, 0.881) |
| ImageNet | DenseNet121 | Linear Model | 0.669(0.631, 0.705) | 0.755(0.706, 0.804) | 0.769(0.716, 0.818) | 0.789(0.739, 0.836) |
| MoCo | ResNet18 | End-to-end | 0.813(0.779, 0.842) | 0.885(0.856, 0.909) | 0.915(0.885, 0.937) | 0.942(0.921, 0.962) |
| ImageNet | ResNet18 | End-to-end | 0.775(0.737, 0.807) | 0.858(0.828, 0.886) | 0.898(0.872, 0.921) | 0.942(0.920, 0.961) |
| MoCo | DenseNet121 | End-to-end | 0.795(0.758, 0.828) | 0.877(0.849, 0.901) | 0.920(0.896, 0.941) | 0.953(0.935, 0.969) |
| ImageNet | DenseNet121 | End-to-end | 0.747(0.712, 0.779) | 0.857(0.821, 0.886) | 0.908(0.884, 0.929) | 0.949(0.928, 0.967) |

Table 1: Data table corresponding to Figure 2 and Figure 3. AUC of models trained to detect pleural effusion on the CheXpert dataset.

| Pretraining | Architecture | Fine Tuning | 6.25% | 25.0% | 100.0% |
|---|---|---|---|---|---|
| MoCo | ResNet18 | Linear Model | 0.902(0.848, 0.945) | 0.928(0.874, 0.966) | 0.944(0.907, 0.972) |
| ImageNet | ResNet18 | Linear Model | 0.852(0.791, 0.901) | 0.903(0.845, 0.952) | 0.927(0.873, 0.971) |
| MoCo | DenseNet121 | Linear Model | 0.899(0.850, 0.939) | 0.927(0.881, 0.963) | 0.943(0.904, 0.975) |
| ImageNet | DenseNet121 | Linear Model | 0.872(0.817, 0.920) | 0.909(0.852, 0.951) | 0.931(0.886, 0.968) |
| MoCo | ResNet18 | End-to-end | 0.911(0.867, 0.943) | 0.953(0.919, 0.978) | 0.974(0.945, 0.993) |
| ImageNet | ResNet18 | End-to-end | 0.911(0.863, 0.952) | 0.948(0.912, 0.978) | 0.970(0.943, 0.990) |
| MoCo | DenseNet121 | End-to-end | 0.909(0.866, 0.947) | 0.950(0.911, 0.980) | 0.984(0.960, 0.999) |
| ImageNet | DenseNet121 | End-to-end | 0.898(0.848, 0.935) | 0.944(0.904, 0.980) | 0.971(0.946, 0.991) |

Table 2: Data table corresponding to Figure 4. AUC of models trained to detect tuberculosis on the Shenzhen dataset.

| Architecture | MoCo-pretrained | ImageNet-pretrained | 6.25% | 25.0% | 100% |
|---|---|---|---|---|---|
| ResNet18 | End-to-End | End-to-End | 0.001(-0.022, 0.027) | 0.005(-0.012, 0.027) | 0.003(-0.014, 0.020) |
| ResNet18 | Linear Model | Linear Model | 0.054( 0.024, 0.086) | 0.026(-0.001, 0.056) | 0.018(-0.011, 0.053) |
| ResNet18 | Linear Model | End-to-End | -0.007(-0.029, 0.015) | -0.020(-0.040, -0.003) | -0.026(-0.052, -0.005) |
| DenseNet121 | End-to-End | End-to-End | 0.011(-0.006, 0.028) | 0.006(-0.010, 0.023) | 0.013(-0.003, 0.033) |
| DenseNet121 | Linear Model | Linear Model | 0.024(-0.001, 0.050) | 0.016(-0.011, 0.043) | 0.013(-0.014, 0.041) |
| DenseNet121 | Linear Model | End-to-End | -0.001(-0.023, 0.019) | -0.016(-0.035, 0.001) | -0.027(-0.053, -0.003) |

Table 3: AUC improvements achieved by MoCo-pretrained models against ImageNet-pretrained models on the Shenzhen tuberculosis task.