Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Domain Adapt Represent Transf Afford Healthc AI Resour Divers Glob Health (2021).
Author manuscript; available in PMC 2022 June 15.

Published in final edited form as:
Domain Adapt Represent Transf Afford Healthc AI Resour Divers Glob Health (2021). 2021 ; 12968: 3–13. doi:10.1007/978-3-030-87722-4_1.

A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis

Mohammad Reza Hosseinzadeh Taher[1],
Fatemeh Haghighi[1],
Ruibin Feng[2],
Michael B. Gotway[3],
Jianming Liang[1]

[1]Arizona State University, Tempe, AZ 85281, USA

[2]Stanford University, Stanford, California 94305, USA

[3]Mayo Clinic, Scottsdale, AZ 85259, USA

Abstract

Transfer learning from supervised ImageNet models has been frequently used in medical image analysis. Yet, no large-scale evaluation has been conducted to benchmark the efficacy of newly-developed pre-training techniques for medical image analysis, leaving several important questions unanswered. As the first step in this direction, we conduct a systematic study on the transferability of models pre-trained on iNat2021, the most recent large-scale fine-grained dataset, and 14 top self-supervised ImageNet models on 7 diverse medical tasks in comparison with the supervised ImageNet model. Furthermore, we present a practical approach to bridge the domain gap between natural and medical images by continually (pre-)training supervised ImageNet models on medical images. Our comprehensive evaluation yields new insights: (1) pre-trained models on fine-grained data yield distinctive local representations that are more suitable for medical segmentation tasks, (2) self-supervised ImageNet models learn holistic features more effectively than supervised ImageNet models, and (3) continual pre-training can bridge the domain gap between natural and medical images. We hope that this large-scale open evaluation of transfer learning can direct the future research of deep learning for medical imaging. As open science, all codes and pre-trained models are available on our GitHub page https://github.com/JLiangLab/BenchmarkTransferLearning.

Keywords

Transfer learning; ImageNet pre-training; Self-supervised learning

mhossei2@asu.edu .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1 Introduction

To circumvent the challenge of annotation dearth in medical imaging, fine-tuning supervised ImageNet models (i.e., models trained on ImageNet via supervised learning with the human labels) has become the standard practice [18,35,34,44,17]. As evidenced by [44], nearly all top-performing models in a wide range of representative medical applications, including classifying the common thoracic diseases, detecting pulmonary embolism (PE), identifying skin cancer, and detecting Alzheimer's Disease, are fine-tuned from supervised ImageNet models. However, intuitively, achieving outstanding performance on medical image classification and segmentation would require fine-grained features. For instance, all chest radiographs (CXR) have a relatively similar appearance; therefore, identifying abnormal conditions and diagnosing specific disorders often rely on recognition of subtle image details. Furthermore, delineating organs and isolating lesions in medical images would demand some fine-detailed features to determine the boundary pixels. In contrast to ImageNet, which was created for coarse-grained object classification, iNat2021 [21], the most recent large-scale fine-grained dataset, has recently been created. It consists of 2.7M training images covering 10K species spanning the entire tree of life. As such, the **first question** this paper seeks to answer is: *What advantages can supervised iNat2021 models offer for medical imaging in comparison with supervised ImageNet models?*

In the meantime, numerous self-supervised learning (SSL) methods have been developed. In the afore-discussed transfer learning, models are pre-trained in a supervised manner using expert-provided labels. By comparison, SSL pre-trained models use machine-generated labels. The recent advancement in SSL has resulted in self-supervised pre-training techniques that surpass gold standard supervised ImageNet models in a number of computer vision tasks [38,12,23,43,7]. Therefore, the **second question** this paper seeks to answer is: *How generalizable are the self-supervised ImageNet models to medical imaging in comparison with supervised ImageNet models?*

More importantly, there are significant differences between natural and medical images. Medical images are typically monochromatic and typically contain consistent anatomical structures [18,17]. Recently, several moderately sized medical imaging datasets have been created, including NIH ChestX-Ray14 [37], which contains 112K images, and CheXpert [22], which contains 224K images. Naturally, the **third question** this paper seeks to answer is: *Can these moderately-sized medical image datasets help bridge the domain gap between natural and medical images?*

To answer these questions, we conduct the first extensive benchmarking study to evaluate the efficacy of different pre-training techniques for diverse medical imaging tasks, covering various diseases (e.g., PE, pulmonary nodules, tuberculosis, etc), organs (e.g., lung and optic fundus), and modalities (e.g., CT, X-ray, and funduscopy). Concretely, (1) we study the impact of pre-training data granularity on transfer learning performance by evaluating the fine-grained pre-trained models on iNat2021 for various medical tasks; (2) we evaluate the transferability of 14 state-of-the-art self-supervised ImageNet models to a diverse set of tasks in medical image classification and segmentation; and (3) we investigate domain-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

adaptive (continual) pre-training [15] on natural and medical datasets to tailor ImageNet models for target tasks on chest X-rays.

Our extensive empirical study reveals the following important insights: (1) Pre-trained models on fine-grained data yield distinctive local representations that are beneficial for medical segmentation tasks, while pre-trained models on coarser-grained data yield high-level features that prevail in classification target tasks (see Fig. 1). (2) For each target task, in terms of the mean performance, there exist at least three self-supervised ImageNet models that outperform the supervised ImageNet model, an observation that is very encouraging, as migrating from conventional supervised learning to self-supervised learning will dramatically reduce annotation efforts (see Fig. 2). (3) Continual (pre-)training of supervised ImageNet models on medical images can bridge the gap between the natural and medical domains, providing more powerful pre-trained models for medical tasks (see Table 2).

2 Transfer Learning Setup

Tasks and datasets:

Table 1 summarizes the tasks and datasets, with more additional details provided in Appendix A. We considered a diverse suite of 7 common but challenging medical imaging tasks encompassing various diseases, organs, and modalities. These tasks span many common properties of medical imaging tasks, such as imbalanced classes, limited data, and small-scanning areas for pathology of interest. We use official data split of these datasets if available; otherwise, we randomly divide the data into 80%/20% for training/testing.

Evaluations:

We evaluate various models pre-trained with different methods and datasets. Therefore, we control other influencing factors such as preprocessing, network architecture, and transfer hyperparameters. In all experiments, (1) for the classification target tasks, the standard ResNet-50 backbone [20] followed by a task-specific classification head is used, (2) for the segmentation target tasks, a U-Net network with a ResNet-50 encoder is used, where the encoder is initialized with the pre-trained models, (3) all target model parameters are fine-tuned, (4) AUC (area under the ROC curve) and Dice coefficient are used for evaluating classification and segmentation target tasks, respectively, (5) mean and standard deviation of performance metrics over ten runs are reported, and (6) statistical analyses based on independent two-sample t-test are presented. More implementation details are in Appendix B and project's GitHub page.

Pre-trained models:

We benchmark transfer learning from two large-scale natural datasets, ImageNet and iNat2021, and two in-domain medical datasets, CheXpert [22] and ChestX-Ray14 [37]. We pre-train supervised in-domain models which are either initialized randomly or fine-tuned from the ImageNet model. For all other supervised and self-supervised methods, we use

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

existing official and ready-to-use pre-trained models, ensuring that their configurations have been meticulously assembled to achieve the best results in target tasks.

3 Transfer Learning Benchmarking and Analysis

1) Pre-trained models on fine-grained data are better suited for segmentation tasks, while pre-trained models on coarse-grained data prevail on classification tasks.

Medical imaging literature mostly has focused on the pre-training with *coarse-grained* natural image datasets, such as ImageNet [28,39,35,30]. In contrast to previous works, we aim to study the capability of pre-training with *fine-grained* datasets for transfer learning to medical tasks. In fine-grained datasets, visual differences between subordinate classes are often subtle and deeply embedded within local discriminative parts. Therefore, a model has to capture visual details in the local regions for solving a fine-grained recognition task [8,45,42]. We hypothesize that a pre-trained model on a fine-grained dataset derives distinctive local representations that are useful for medical tasks which usually rely upon small, local variations in texture to detect/segment pathologies of interest. To put this hypothesis to the test, we empirically validate how well pre-trained models on large-scale fine-grained datasets can transfer to a range of target medical applications. This study represents the first effort to rigorously evaluate the impact of pre-training data *granularity* on transfer learning to medical imaging tasks.

Experimental setup: We examine the applicability of iNat2021 as a pre-training source for medical imaging tasks. Our goal is to compare the generalization of the learned features from fine-grained pre-training on iNat2021 with the conventional pre-training on the ImageNet. Given this goal, we use existing official and ready-to-use pre-trained models on these two datasets, and fine-tune them for 7 diverse target tasks, encompassing multi-label classification, binary classification, and pixel-wise segmentation (see Table 1). To provide a comprehensive evaluation, we also include results for training target models from scratch.

Observations and Analysis: As evidenced in Fig. 1, fine-tuning from the iNat2021 pre-trained model outperforms the ImageNet counterpart in semantic segmentation tasks, *i.e.*, PXS, VFS, and LXS. This implies that, owing to the finer data granularity of iNat2021, the pre-trained model on this dataset yields a more fine-grained visual feature space, which captures essential pixel-level cues for medical segmentation tasks. This observation gives rise to a natural question of whether this improved performance can be attributed to the larger pre-training data of iNat2021 (2.7M images) compared to ImageNet (1.3M images). In answering this question, we conducted an ablation study on the iNat2021 mini dataset [21] with 500K images to further investigate the impact of data granularity on the learned representations. Our result demonstrates that even with fewer pre-training data, iNat2021 mini pre-trained models can outperform ImageNet counterparts in segmentation tasks (see Appendix C). This demonstrates that recovering discriminative features from iNat2021 dataset should be attributed to fine-grained data rather than the larger training data size.

Despite the success of iNat2021 models in segmentation tasks, fine-tuning of ImageNet pre-trained features outperforms iNat2021 in classification tasks, namely DXC_{14}, DXC_5,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TXC, and ECC (see Fig. 1). Contrary to our intuition (see Sec. 1), pre-training on a coarser granularity dataset, such as ImageNet, yields high-level semantic features that are more beneficial for classification tasks.

Summary: Fine-grained pre-trained models could be a viable alternative for transfer learning to fine-grained medical tasks, hoping practitioners will find this observation useful in migrating from standard ImageNet checkpoints to reap the benefits we've demonstrated. Regardless of – or perhaps in addition to – other advancements, visually diverse datasets like ImageNet can continue to play a valuable role in building performant medical imaging models.

2) Self-supervised ImageNet models outperform supervised ImageNet models.

A recent family of self-supervised ImageNet models has demonstrated superior transferability in an increasing number of computer vision tasks compared to supervised ImageNet models [12,43,23]. Self-supervised models, in particular, capture task-agnostic features that can be easily adapted to different domains [38,23], while high-level features of supervised pre-trained models may be extraneous when the source and target data distributions are far apart [43]. We hypothesize this phenomenon is more pronounced in the medical domain, where there is a remarkable domain shift [12] compared to ImageNet. To test this hypothesis, we dissect the effectiveness of a wide range of recent self-supervised methods, encompassing contrastive learning, clustering, and redundancy-reduction methods, on the broadest benchmark yet of various modalities spanning X-ray, CT, and fundus images. This work represents the first effort to rigorously benchmark SSL techniques to a broader range of medical imaging problems.

Experimental setup: We evaluate the transferability of 14 popular SSL methods with officially released models, which have been expertly optimized, including contrastive learning (CL) based on instance discrimination (*i.e.*, InsDis [40], MoCo-v1 [19], MoCo-v2 [11], SimCLR-v1 [9], SimCLR-v2 [10], and BYOL [14]), CL based on JigSaw shuffling (PIRL [27]), clustering (DeepCluster-v2 [7] and SeLa-v2 [7]), clustering bridging CL (PCL-v1 [26], PCL-v2 [26], and SwAV [7]), mutual information reduction (InfoMin [36]), and redundancy reduction (Barlow Twins [41]), on 7 diverse medical tasks. All methods are pre-trained on the ImageNet and use ResNet-50 architecture. Details of SSL methods can be found in Appendix F. As the baseline, we consider the standard supervised pre-trained model on ImageNet with a ResNet-50 backbone.

Observations and Analysis: According to Fig. 2, for each target task, there are at least three self-supervised ImageNet models that outperform the supervised ImageNet model on average. Moreover, the top self-supervised ImageNet models remarkably accelerate the training process of target models in comparison with supervised counterpart (see Appendix E). Intuitively, supervised pre-training labels encourage the model to retain more domain-specific high-level information, causing the learned representation to be biased toward the pre-training task/dataset's idiosyncrasies. Self-supervised learners, however, capture

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

low/mid level features that are not attuned to domain-relevant semantics, generalizing better to diverse sorts of target tasks with low-data regimes.

Comparing the classification (DXC_{14}, DXC_5, TXC, and ECC) and segmentation tasks (PXS, VFS, and LXS) in Fig. 2, in the latter, a larger number of SSL methods results in better transfer performance, while supervised pre-training falls short. This suggests that when there are larger domain shifts, self-supervised models can provide more precise localization than supervised models. This is because supervised pre-trained models primarily focus on the smaller discriminative regions of the images, whereas SSL methods attune to larger regions [12,43], which empowers them with deriving richer visual information from the entire image.

Summary: SSL can learn holistic features more effectively than supervised pre-training, resulting in higher transferability to a variety of medical tasks. It's worth noting that no single SSL method dominates in all tasks, implying that universal pre-training remains a mystery. We hope that the results of this benchmarking, resonating with recent studies in the natural image domain [12,43,23], will lead to more effective transfer learning for medical image analysis.

3) Domain-adaptive pre-training bridges the gap between the natural and medical imaging domains.

Pre-trained ImageNet models are the predominant standard for transfer learning as they are free, open-source models which can be used for a variety of tasks [28,4,17,39]. Despite the prevailing use of ImageNet models, the remarkable covariate shift between natural and medical images restrain transfer learning [30]. This constraint motivates us to present a practical approach that tailors ImageNet models to medical applications. Towards this end, we investigate domain-adaptive pre-training on natural and medical datasets to tune ImageNet models for medical tasks.

Experimental Setup: The domain-adaptive paradigm originated from natural language processing [15]. This is a sequential pre-training approach in which a model is first pre-trained on a massive general dataset, such as ImageNet, and then pre-trained on domain-specific datasets, resulting in domain-adapted pre-trained models. For the first pre-training step, we used the supervised ImageNet model. For the second pre-training step, we created two new models that were initialized through the ImageNet model followed by supervised pre-training on CheXpert (ImageNet→CheXpert) and ChestX-ray14 (ImageNet→ChestX-ray14). We compare the domain-adapted models with (1) the ImageNet model, and (2) two supervised pre-trained models on CheXpert and ChestX-ray14, which are randomly initialized. In contrast to previous work [4] which is limited to two classification tasks, we evaluate domain-adapted models on a broader range of five target tasks on chest X-ray scans; these tasks span classification and segmentation, ascertaining the generality of our findings.

Observations and Analysis: We draw the following observations from Table 2. (1) Both ChestX-ray14 and CheXpert models consistently outperform the ImageNet model

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in all cases. This observation implies that in-domain medical transfer learning, whenever possible, is preferred over ImageNet transfer learning. Our conclusion is opposite to [39], where in-domain pre-trained models outperform ImageNet models in controlled setups but lag far behind the real-world ImageNet models. (2) The overall trend showcases the advantage of domain-adaptive pre-training. Specifically, for DXC_{14}, fine-tuning the ImageNet→CheXpert model surpasses both ImageNet and CheXpert models. Furthermore, the dominance of domain-adapted models (ImageNet→CheXpert and ImageNet→ChestX-ray14) over ImageNet and corresponding in-domain models (CheXpert and ChestX-ray14) is conserved at LXS, TXC, and PXS. This suggests that domain-adapted models leverage the learning experience of the ImageNet model and further refine it with domain-relevant data, resulting in more pronounced representation. (3) In DXC_5, the domain-adapted performance decreases relative to corresponding ImageNet and in-domain models. This is most likely due to the lesser number of images in the in-domain pre-training dataset than the target dataset (75K vs. 200K), suggesting that in-domain pre-training data should be larger than the target data [15,33].

Summary: Continual pre-training can bridge the domain gap between natural and medical images. Concretely, we leverage the readily conducted annotation efforts to produce more performant medical imaging models and reduce future annotation burdens. We hope our findings posit new research directions for developing specialized pre-trained models in medical imaging. Our pre-trained models, in-domain (CheXpert and ChestX-ray14) as well as domain-adapted (ImageNet→CheXpert and ImageNet→ChestX-ray14), are publicly available on our GitHub page.

4 Conclusion and Future Work

We provide the first fine-grained and up-to-date study on the transferability of various brand-new pre-training techniques for medical imaging tasks, answering central and timely questions on transfer learning in medical image analysis. Our empirical evaluation suggests that: (1) what truly matters for the segmentation tasks is fine-grained representation rather than high-level semantic features, (2) top self-supervised ImageNet models outperform the supervised ImageNet model, offering a new transfer learning standard for medical imaging, and (3) ImageNet models can be strengthened with continual in-domain pre-training.

Future work:

In this work, we have considered transfer learning from the supervised ImageNet model as the baseline, on which all our evaluations are benchmarked. To compute p-values for statistical analysis, 14 SSL, 5 supervised, and 2 domain-adaptive pre-trained models were run 10 times each on a set of 7 target tasks— leading to a large number of experiments (1,420). Nevertheless, our self-supervised models were all pre-trained on ImageNet with ResNet50 as the backbone. While ImageNet is generally regarded as a strong source for pre-training [21,39], pre-training modern self-supervised models with iNat2021 and in-domain medical image data on various architectures may offer even deeper insights into transfer learning for medical imaging.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Acknowledgments

This research has been supported partially by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and partially by the NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work has utilized the GPUs provided partially by the ASU Research Computing and partially by the Extreme Science and Engineering Discovery Environment (XSEDE) funded by the National Science Foundation (NSF) under grant number ACI-1548562. We thank Nahid Islam for evaluating the self-supervised methods on the PE detection target task. The content of this paper is covered by patents pending.

Appendix

A Datasets

iNat2021 [21]:

The iNaturalist2021 dataset (iNat2021) is a recent large-scale, fine-grained species dataset with 2.7M training images covering 10k species. This dataset facilitates fine-grained visual classification problems. Compared to the more widely used dataset, ImageNet, iNat2021 contains a greater number of these fine-grained images but a narrower range of visual diversity.

iNat2021 mini [21]:

In addition to the full sized dataset, Horn *et al.* [21] created a smaller version of iNat2021, named iNat2021 mini, that contains 50 training images per species, sampled from the full train split. In total, iNat2021 mini includes 500K training images covering 10k species.

ChestX-ray14 [37]:

This hospital-scale chest X-ray dataset contains 112K frontal-view X-ray images taken from a sample of 30K unique patients. ChestX-ray14 provides an official patient-wise split for training (86K images) and test sets (25K images). In this dataset, 51K images have at least one of the 14 thorax diseases. We use the official data split and report the mean AUC score over 14 diseases for the multi-label chest X-ray classification task.

CheXpert [22]:

This large-scale publicly available dataset contains 224K high-quality chest X-ray images taken from a sample of 65K patients. The training images were annotated by a labeler to automatically detect the presence of 14 thorax diseases in radiology reports, capturing uncertainties inherent in radiograph interpretation. The test set consists of 234 images from 200 patients. The test images were manually annotated by board-certified radiologists for 5 selected diseases, i.e., Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. We use the official data split and report the mean AUC score over 5 test diseases.

SIIM-ACR Pneumothorax Segmentation [1]:

The Society for Imaging Informatics in Medicine (SIIM) and American College of Radiology provided the SIIM-ACR Pneumothorax Segmentation dataset, consisting of 10K chest X-ray images and the segmentation masks for Pneumothorax disease. We randomly

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

divided the dataset into training (80%) and testing (20%), and the segmentation performance was evaluated by using the Dice coefficient score.

RSNA PE Detection [2]:

This dataset is the largest publicly available annotated Pulmonary Embolism (PE) dataset, comprised of more than 7K CT scans with a varying number of images in each scan. Each image has been annotated for the presence or absence of the PE. Also, each scan has been labeled for additional nine patient-level labels. We randomly split the data at patient-level to training (6K) and testing (1K) sets, respectively. Correspondingly, there are 1.5M and 248K images in the training and testing sets, respectively. We report the AUC score for the PE detection task.

NIH Shenzhen CXR [24]:

The dataset contains 662 frontal-view chest X-rays, of which 326 are normal cases and 336 are cases with manifestations of Tuberculosis (TB), including pediatric X-rays (AP). We randomly divide the dataset into a training set (80%) and a test set (20%). We report the AUC score for the Tuberculosis detection task.

NIH Montgomery [24]:

The dataset contains 138 frontal-view chest X-rays from Montgomery County's Tuberculosis screening program, of which 80 are normal cases and 58 are cases with manifestations of TB. The segmentation masks for left and right lungs are provided. We randomly divided the dataset into a training set (80%) and a test set (20%) and report the mean Dice score for the lung segmentation task.

DRIVE [5]:

The dataset contains 40 retinal images, separated by its providers into a training set (20 images) and a test set (20 images). For all images, manual segmentation of the vasculature is provided. We use the official data split and report the mean Dice score for the segmentation of blood vessels.

B Implementation

We evaluate popular publicly available representations that have been pre-trained with various methods and datasets across a variety of target tasks. Therefore, we control other influencing factors such as pre-processing, network architecture, and transfer hyperparameters. We run each method ten times on all of the target tasks and report the average, standard deviation, and further present statistical analysis based on an independent two-sample *t*-test.

Architecture:

We fix the network architecture in all experiments since we seek to understand the competitiveness of representations rather than benefits from architecture. Therefore, all the pre-trained models leverage the same ResNet-50 backbone. For transfer learning to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the classification target tasks, we take the pre-trained ResNet-50 models and append a task-specific classification head. For the segmentation target tasks, we utilize a U-Net[4] network with a ResNet50 encoder, where the encoder is initialized with the pre-trained models. We have evaluated the transfer learning performance of all pre-trained models by fine-tuning all layers in the downstream networks.

Preprocessing and data augmentation:

For target tasks on X-ray modality (DXC_{14}, DXC_5, TXC, LXS, and PXS), Fundoscopic modality (VFS), and CT modality (PCC), we resize the images to 224×224, 512×512, and 576×576, respectively. For all classification target tasks, we apply standard data augmentation techniques, including random cropping, horizontal flipping, and rotating. For segmentation tasks on X-ray modality (LXS and PXS), we employ RandomBrightnessContrast, RandomGamma, OpticalDistortion, elastic transformation, and grid distortion. For segmentation task on fundoscopic modality (VFS), we use random rotation, Gaussian noise, color jittering, and horizontal, vertical and diagonal flips.

Training parameters:

Since different datasets require different optimal settings, we strive to optimize each target task with the best performing hyperparameters. In all experiments, we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use *ReduceLROnPlateau* and *cosine* learning rate decay schedulers for classification and segmentation tasks, respectively; if no improvement is seen in the validation set for a certain number of epochs, the learning rate is reduced. We employ early-stop mechanism using the 10% of the training data as the validation set to avoid over-fitting. For X-ray classification tasks (DXC_{14}, DXC_5, and TXC), segmentation tasks (VFS, LXS, and PXS), and PE detection task (ECC), we use a learning rate of $2e-4$, $1e-3$, and $4e-4$, respectively.

Table 3.

Evaluation of iNat2021 mini dataset on segmentation medical tasks. Even with less than half number of pre-training samples, iNat2021 mini achieves equal or superior performance over ImageNet counterpart. Best performance is **bolded** and second best is <u>underlined</u>.

Pre-training task		Target tasks		
Dataset	**#training data**	**PXS**	**VFS**	**LXS**
ImageNet	1.3M	67.93±1.45	79.20±0.34	98.19±0.13
iNat2021 mini	500K	<u>68.26±1.48</u>	<u>79.24±0.28</u>	<u>98.19±0.09</u>
iNat2021	2.7M	**68.43±0.92**	**79.33±0.18**	**98.25±0.07**

[4]Segmentation Models: https://github.com/qubvel/segmentation_models.pytorch

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4.

Benchmarking transfer learning from supervised iNat2021 and ImageNet models on seven medical tasks. Pre-trained models on iNat2021 are better suited for segmentation tasks (*i.e.*, LXS, VFS, and PXS), while pre-trained models on ImageNet prevail on classification tasks (*i.e.*, DXC_{14}, DXC_5, TXC, and ECC). The best model in each application is **bolded**.

Downstream task	Initialization		
	Random	ImageNet	iNat2021
Pneumothorax segmentation (PXS)	67.54±0.60	67.93±1.45	**68.43±0.92**
Lung segmentation (LXS)	97.55±0.36	98.19±0.13	**98.25±0.07**
Blood Vessels Segmentation (VFS)	78.27±0.40	79.20±0.34	**79.33±0.18**
14 thorax diseases classification (DXC_{14})	80.31±0.10	**81.70±0.15**	81.67±0.19
5 thorax disease classification (DXC_5)	86.62±0.46	**87.10±0.36**	86.26±0.69
Tuberculosis Detection (TXC)	89.03±1.82	**95.62±0.63**	94.90±0.69
Pulmonary Embolism Detection (ECC)	90.37±1.32	**94.73±0.12**	94.44±0.23

C Ablation study on iNat2021 mini dataset

We further investigate the capability of pre-trained models on fine-grained datasets in capturing fine-grained details by examining iNat2021 mini dataset for segmentation tasks. iNat2021 mini contains 500K images, which is less than half compared to ImageNet. The results in Table 3 indicate that even with fewer training data, iNat2021 achieves equal or better performance than ImageNet counterpart. This observation suggests that the superior performance of iNat2021 over ImageNet pre-trained model in segmentation tasks should be attributed to the fine-grained nature of data rather than larger pre-training size.

Table 5.

Benchmarking transfer learning from fourteen self-supervised ImageNet pre-trained models on seven medical tasks. Self-supervised ImageNet models outperform supervised ImageNet models. The best model is **bolded**, and all the other models that outperform supervised baseline are <u>underlined</u>.

Pre-training	Downstream task			
	$DXCi_4$	dxc_5	txc	ecc
Supervised	81.70±0.15	87.10±0.36	95.62±0.63	94.73±0.12
InsDis	81.09±0.17	<u>87.21±0.36</u>	94.81±0.73	91.16±1.12
MoCo-v1	81.33±0.10	86.98±0.11	94.54±0.42	90.29±1.92
PCL-v1	<u>81.79±0.22</u>	86.90±0.25	95.15±0.53	94.34±0.20
PIRL	80.60±0.13	86.79±0.35	93.34±2.72	89.17±2.62
PCL-v2	81.69±0.21	<u>87.27±0.19</u>	95.45±0.62	93.78±0.31
SimCLR-v1	81.22±0.07	**87.66±0.14**	94.45±0.76	<u>95.45±0.11</u>
MoCo-v2	80.46±0.54	86.94±0.20	95.57±0.90	89.20±2.92
SimCLR-v2	81.23±0.09	86.86±0.37	95.29±0.93	93.66±0.29
SeLa-v2	81.55±0.23	<u>87.24±0.29</u>	**96.23±0.81**	**95.68±0.05**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Pre-training	Downstream task			
	DXCi$_4$	**dxc$_5$**	**txc**	**ecc**
InfoMin	80.66±0.35	86.67±0.10	95.02±1.40	90.36±1.84
BYOL	81.58±0.11	87.09±0.40	94.69±0.78	95.63±0.05
DeepCluster-v2	81.57±0.38	87.01±0.19	96.09±0.68	**95.68±0.06**
SwAV	**81.93±0.18**	87.06±0.50	95.72±0.50	95.63±0.10
Barlow Twins	81.77±0.25	87.25±0.27	94.50±0.88	95.66±0.07
	PXS	**LXS**	**VFS**	
Supervised	67.93±1.45	98.19±0.13	79.20±0.34	
InsDis	67.84±0.55	98.25±0.03	79.03±0.34	
MoCo-v1	67.88±0.95	98.08±0.14	78.98±0.45	
PCL-v1	68.60±0.78	98.25±0.05	78.99±0.21	
PIRL	66.20±4.24	98.02±0.11	79.24±0.42	
PCL-v2	68.62±0.92	98.26±0.06	79.06±0.19	
SimCLR-v1	70.52±0.69	98.19±0.10	79.00±0.18	
MoCo-v2	67.01±1.28	97.79±0.50	79.23±0.19	
SimCLR-v2	68.26±1.39	98.16±0.20	78.72±0.37	
SeLa-v2	70.52±0.17	**98.28±0.04**	79.65±0.19	
InfoMin	67.34±2.28	97.94±0.16	79.63±0.30	
BYOL	70.04±0.62	98.20±0.08	79.39±0.22	
DeepCluster-v2	**70.59±0.55**	98.24±0.05	**79.66±0.21**	
SwAV	70.44±0.75	**98.28±0.05**	79.65±0.14	
Barlow Twins	70.42±0.15	98.23±0.05	79.48±0.16	

D Tabular results

In this section, tabulated results of different experiments are reported. The results of Fig. 1 and Fig. 2 in the main paper are presented in Table 4 and Table 5, respectively.

E Convergence Time Analysis

Transfer learning attracts great attention since it improves the target performance and accelerates the model convergence when compared to training from scratch. In that respect, a good pre-trained model should yield better target performance with less training time. Therefore, we further evaluate the pre-trained models in terms of accelerating the training process of various medical tasks. In the following, we provide the training time results for each of the three groups of experiments in the main paper. We used the early-stop technique in all target tasks, and report the average number of training epochs over ten runs for each model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6.

Fine-tuning from iNat2021 model provides higher performance in all segmentation tasks and considerably accelerates the training process in two out of three tasks in comparison to the ImageNet counterpart. The average performance and number of training epochs over ten runs is reported for each model in each target task. The best performance in each task is bolded.

Initialization	PXS		VFS		LXS	
	Dice(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)
Random	67.54±0.60	46.0±13.87	78.27±0.40	100.0±0.0	97.55±0.36	92.6±51.38
ImageNet	67.93±1.45	45.9±28.25	79.20±0.34	71±18.29	98.19±0.13	84.9±27.55
iNat2021	**68.43±0.92**	41.8±17.98	**79.33±0.18**	59.3±5.58	**98.25±0.07**	98.9±26.41

Table 7.

Fine-tuning from the best self-supervised models provide significantly better or equivalent performance and accelerate the training process in comparison to the supervised counterpart. The average performance and number of training epochs over ten runs is reported for each model in each target task. The best performance in each task is bolded.

Initialization	DXC_{14}		PXS		VFS	
	AUC(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)
Random	80.40±0.05	68.2±5.07	67.54±0.60	46.0±13.87	78.27±0.40	100.0±0.0
Supervised	81.70±0.15	34.2±4.32	67.93±1.45	45.9±28.25	79.20±0.34	71±18.29
SeLa-v2	81.55±0.23	10.0±0.71	70.52±0.17	39.40±8.99	79.65±0.19	45±10.27
DeepCluster-v2	81.57±0.38	8.80±1.92	**70.59±0.55**	37.0±17.46	**79.66±0.21**	43.6±3.58
SwAV	**81.93±0.18**	13.0±2.55	70.44±0.75	44.8±18.31	79.65±0.14	44.4±5.27
Barlow Twins	81.77±0.25	12.4±2.61	70.42±0.15	55.8±24.32	79.48±0.16	47.6±4.72

1) Supervised ImageNet model vs. supervised iNat2021 model.

We provide the training time of the segmentation tasks in which the iNat2021 model outperforms its ImageNet counterpart. The results in Table 6 indicate that fine-tuning from the iNat2021 model provides higher performance in all segmentation tasks and considerably accelerates the training process in two out of three tasks in comparison to the ImageNet counterpart.

2) Supervised ImageNet model vs. self-supervised ImageNet models.

We compare the training time of the top four self-supervised ImageNet models (based on the overall performances in different target tasks) to the supervised ImageNet model in three target tasks, including classification and segmentation. To provide a comprehensive evaluation, we also include results for training target models from scratch.

Our results in Table 7 demonstrate that fine-tuning from the best self-supervised models in each target task provide significantly better or equivalent performance and remarkably

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

accelerate the training process in comparison to the supervised counterpart. Specifically, in DXC_{14} task, SwAV and Barlow Twins achieve superior performance with significantly less number of training epochs compared to supervised ImageNet model. Similarly, in PXS task, SeLa-v2, DeepCluster-v2, and SwAV outperform supervised ImageNet model in terms of both performance and training time. Furthermore, in VFS task, all the self-supervised models yield higher performance with less training time compared to supervised ImageNet model.

Table 8.

Fine-tuning from the domain-adapted pre-trained models provides higher performance in all tasks and speeds up the training process compared to the corresponding ImageNet models in most cases. The average performance and number of training epochs over ten runs is reported for each model in each target task. The best performance in each task is bolded. "CXR14" denotes the ChestX-ray14 dataset. When pre-training and target tasks are the same, transfer learning is not applicable, denoted by "-".

Initialization	dxc_{14}		PXS		LXS	
	AUC(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)	Dice(↑)	#Epochs(↓)
Random	80.31±0.10	68.2±5.07	67.54±0.60	46.0±13.87	97.55±0.36	92.6±51.38
ImageNet	81.70±0.15	34.2±4.32	67.93±1.45	45.9±28.25	98.19±0.13	84.9±27.55
CXR14	-	-	68.92±0.98	49.0±35.48	98.18±0.06	66.2±20.58
CheXpert	81.99±0.08	15.8±4.32	69.30± 0.50	42.2±22.85	98.25±0.04	84.0±16.85
ImageNet→XR14	-	-	**69.52±0.38**	37.6±15.49	98.27±0.03	72.6±30.92
ImageNet→heXpert	**82.25±0.18**	22.2±3.49	69.36±0.49	45.8±5.93	**98.31±0.05**	110.7±40.08

Additionally, considering the principle that a good representation should generalize to multiple target tasks with limited fine-tuning [13], we fine-tuned all the models for the same number of training epochs in DXC_5 and ECC (ten and one, respectively). According to the results in Fig. 2 in the main paper and Table 5 in Appendix, with the same number of training epochs, the best self-supervised ImageNet models, such as SimCLR-v1, SeLa-v2, and Barlow Twins, achieve superior performance over supervised ImageNet models in both target tasks.

3) Supervised ImageNet model vs. domain-adapted models.

We compare the training time of the in-domain pre-trained models to ImageNet counterparts. According to the results in Table 8, (1) ChestX-ray14 and CheXpert models consistently outperform ImageNet models in terms of convergence time in most cases, and (2) The overall trend showcases the faster convergence of domain-adapted pre-trained models (*i.e.*, ImageNet→CheXpert and ImageNet→ChestX-ray14) compared to the corresponding ImageNet models.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

F Self-supervised Learning Methods

InsDis [40]:

InsDis treats each image as a distinct class and trains a non-parametric classifier to distinguish between individual classes based on noise-contrastive estimation (NCE) [16]. InsDis introduces a feature memory bank maintaining a large number of noise samples (referred to as negative samples), to avoid exhaustive feature computing.

MoCo-v1 [19] and MoCo-v2 [11]:

MoCo-v1 creates two views by applying two independent data augmentations to the same image X, referred to as positive samples. Like InsDis, the images other than X are defined as negative samples stored in a memory bank. Additionally, a momentum encoder is proposed to ensure the consistency of negative samples as they evolve during training. Intuitively, MoCo-v1 aims to increase the similarity between positive samples while decreasing the similarity between negative samples. Through simple modifications inspired by SimCLR-v1 [9], such as a non-linear projection head, extra augmentations, cosine decay schedule, and a longer training time to MoCo-v1, MoCo-v2 establishes a stronger baseline while eliminating large training batches.

SimCLR-v1 [9] and SimCLR-v2 [10]:

SimCLR-v1 is proposed independently following the same intuition as MoCo. However, instead of using special network architectures (*e.g.*, a momentum encoder) or a memory bank, SimCLR-v1 is trained in an end-to-end fashion with large batch sizes. Negative samples are generated within each batch during the training process. In SimCLR-v2, the framework is further optimized by increasing the capacity of the projection head and incorporating the memory mechanism from MoCo to provide more negative samples than SimCLR-v1.

BYOL [14]:

Conventional contrastive learning methods such as MoCo and SimCLR relies on a large number of negative samples. As a result, they require either a large memory bank (memory consuming) or a large batch size (computational consuming). On the contrary, BYOL avoids the use of negative pairs by leveraging two encoders, named online and target, and adding a predictor after the projector in the online encoder. BYOL thus maximizes the agreement between the prediction from the online encoder and the features computed from the target encoder. The target encoder is updated with the momentum mechanism to prevent the collapsing problem.

PIRL [27]:

Instead of using instance discrimination objectives like InsDis and MoCo, PIRL adapts Jigsaw and Rotation as proxy tasks. Specifically, the positive samples are generated by applying Jigsaw shuffling or rotating images by {0°, 90°, 180°, 270°}. PIRL defines a loss function based on noise-contrastive estimation (NCE) and uses a memory bank following

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

InsDis. In this paper, we only benchmark PIRL with Jigsaw shuffling, which yields better performance than its rotation counterpart.

DeepCluster-v2 [7]:

DeepCluster [6] learns features in two phases: (1) self-labeling, where pseudo labels are generated by clustering data points using the prior representation— yielding cluster indexes for each sample; (2) feature-learning, where the cluster index of each sample is used as a classification target to train a model. The two phases are performed repeatedly until the model converges. Rather than classifying the cluster index, DeepCluster-v2 explicitly minimizes the distance between each sample and the corresponding cluster centroid. DeepCluster-v2 finally applies stronger data augmentation, a MLP projection head, a cosine decay schedule, and multi-cropping to improve the representation learning.

SeLa-v2 [7]:

Similar to clustering methods, SeLa [3] requires a two-phase training (*i.e.*, self-labeling and feature-learning). However, instead of clustering the image instances, SeLa formulates self-labeling as an optimal transport problem, which can be effectively solved by adopting the Sinkhorn-Knopp algorithm. Similar to DeepCluster-v2, the updated SeLa-v2 applies stronger data augmentation, a MLP projection head, a cosine decay schedule, and multi-cropping to improve the representation learning.

PCL-v1 and PCL-v2 [26]:

PCL-v1 combines contrastive learning and clustering approaches to encode the semantic structure of the data into the embedding space. Specifically, PCL-v1 adopts the architecture of MoCo, and incorporates clustering in representation learning. Similar to clustering-based feature learning, PCL-v1 has self-labeling and feature-learning phases. In self-labeling phase, the features obtained from the momentum encoder are clustered, in where each instance is assigned to multiple prototypes (cluster centroids) with different granularity. In the feature-learning phase, PCL-v1 extends the noise-contrastive estimation (NCE) loss to ProtoNCE loss which can push each sample closer to its assigned prototypes. PCL-v2 is developed by applying the aforementioned techniques to promote the representation learning.

SwAV [7]:

SwAV takes advantages of both contrastive learning and clustering techniques. Similar to SeLa, SwAV calculates cluster assignments (codes) for each data sample with the Sinkhorn-Knopp algorithm. However, SwAV performs online cluster assignments, *i.e.*, at the batch level instead of epoch level. Compared with contrastive learning approaches such as MoCo and SimCLR, SwAV “swapped” predicts the codes obtained from one view using the other view rather than comparing their features directly. Additionally, SwAV proposes a multi-cropping strategy, which can be adopted by other methods to consistently improve their performance.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

InfoMin [36]:

InfoMin hypothesizes that good views (or positive samples) should only share label information w.r.t the downstream task while throwing away irrelevant factors, which means optimal views for contrastive representation learning are task-dependent. Following this hypothesis, InfoMin optimizes data augmentations by further reducing mutual information between views.

Barlow Twins [41]:

This method consists of two online encoders that are fed by two augmented views of the same image. The model is trained by making the cross-correlation matrix of two encoders' outputs as close to the identity matrix as possible. As a result,(1) the similarity between representations of two views is maximized, which is similar to the ultimate goal of contrastive learning, and (2) the redundancy between the components of two representations is minimized.

References

1. Siim-acr pneumothorax segmentation (2019), https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/
2. Rsna str pulmonary embolism detection (2020), https://www.kaggle.com/c/rsna-str-pulmonary-embolism-detection/overview
3. Asano YM, Rupprecht C, Vedaldi A: Self-labelling via simultaneous clusteringand representation learning (2020)
4. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A,Karthikesalingam A, Kornblith S, Chen T, Natarajan V, Norouzi M: Big self-supervised models advance medical image classification. arXiv:2101.05224 (2021)
5. Budai A, Bock R, Maier A, Hornegger J, Michelson G: Robust vessel segmentation in fundus images. International Journal of Biomedical Imaging (2013)
6. Caron M, Bojanowski P, Joulin A, Douze M: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision. pp. 132–149 (2018)
7. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A: Unsupervisedlearning of visual features by contrasting cluster assignments. arXiv:2006.09882 (2021)
8. Chang D, Ding Y, Xie J, Bhunia AK, Li X, Ma Z, Wu M, Guo J,Song YZ: The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Transactions on Image Processing 29, 4683–4695 (2020)
9. Chen T, Kornblith S, Norouzi M, Hinton G: A simple framework for contrastive learning of visual representations. In: III HD, Singh A (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020)
10. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G: Big self-supervisedmodels are strong semi-supervised learners (2020)
11. Chen X, Fan H, Girshick R, He K: Improved baselines with momentum contrastive learning (2020)
12. Ericsson L, Gouk H, Hospedales TM: How well do self-supervised models transfer? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5414–5423 (June 2021)
13. Goyal P, Mahajan D, Gupta A, Misra I: Scaling and benchmarking self-supervised visual representation learning. arXiv:1905.01235 (2019)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

14. Grill JB, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, Doersch C, Pires BA, Guo ZD, Azar MG, Piot B, Kavukcuoglu K, Munos R, Valko M: Bootstrap your own latent: A new approach to self-supervised learning (2020)
15. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D,Smith NA: Don't stop pretraining: Adapt language models to domains and tasks. arXiv:2004.10964 (2020)
16. Gutmann M, Hyvärinen A: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Teh YW, Titterington M (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 297–304. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010)
17. Haghighi F, Hosseinzadeh Taher MR, Zhou Z, Gotway MB, Liang J: Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 137–147. Springer International Publishing, Cham (2020)
18. Haghighi F, Taher MRH, Zhou Z, Gotway MB, Liang J: Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. arXiv:2102.10680 (2021)
19. He K, Fan H, Wu Y, Xie S, Girshick R: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
20. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
21. Horn GV, Cole E, Beery S, Wilber K, Belongie S, Aodha OM: Benchmarking representation learning for natural world image collections. arXiv: 2103.16483 (2021)
22. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H,Haghgoo B, Ball R, Shpanskaya K, et al. : Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031 (2019)
23. Islam A, Chen CF, Panda R, Karlinsky L, Radke R, Feris R: A broad studyon the transferability of visual representations with contrastive learning (2021)
24. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G: Twopublic chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery 4(6) (2014)
25. Kim E, Kim S, Seo M, Yoon S: Xprotonet: Diagnosis in chest radiographywith global and local explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15719–15728 (2021)
26. Li J, Zhou P, Xiong C, Hoi SCH: Prototypical contrastive learning of unsupervised representations (2021)
27. Misra I, Maaten L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
28. Mustafa B, Loh A, Freyberg J, MacWilliams P, Wilson M, McKinney SM,Sieniek M, Winkens J, Liu Y, Bui P, Prabhakara S, Telang U, Karthikesalingam A, Houlsby N, Natarajan V: Supervised transfer learning at scale for medical imaging. arXiv:2101.05913 (2021)
29. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing 437, 186–194 (2021)
30. Raghu M, Zhang C, Kleinberg J, Bengio S: Transfusion: Understanding transfer learning with applications to medical imaging. arXiv:1902.07208 (2019)
31. Rajaraman S, Zamzmi G, Folio L, Alderson P, Antani S: Chest x-ray bonesuppression for improving classification of tuberculosis-consistent findings. Diagnostics 11(5) (2021)
32. Reamaroon N, Sjoding MW, Derksen H, Sabeti E, Gryak J, Barbaro RP,Athey BD, Najarian K: Robust segmentation of lung in chest x-ray: applications in analysis of acute respiratory distress syndrome. BMC Medical Imaging 20 (2020)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

33. Reed CJ, Yue X, Nrusimha A, Ebrahimi S, Vijaykumar V, Mao R, Li B, Zhang S, Guillory D, Metzger S, Keutzer K, Darrell T: Self-supervised pretraining improves self-supervised pretraining. arXiv:2103.12718 (2021)
34. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5), 1285–1298 (2016) [PubMed: 26886976]
35. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35(5), 1299–1312 (2016) [PubMed: 26978662]
36. Tian Y, Sun C, Poole B, Krishnan D, Schmid C, Isola P: What makes forgood views for contrastive learning? (2020)
37. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
38. Wei L, Xie L, He J, Chang J, Zhang X, Zhou W, Li H, Tian Q: Cansemantic labels assist self-supervised visual representation learning? (2020)
39. Wen Y, Chen L, Deng Y, Zhou C: Rethinking pre-training on medical imaging. Journal of Visual Communication and Image Representation 78, 103145 (2021)
40. Wu Z, Xiong Y, Yu SX, Lin D: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
41. Zbontar J, Jing L, Misra I, LeCun Y, Deny S: Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230 (2021)
42. Zhao J, Peng Y, He X: Attribute hierarchy based multi-task learning for fine-grained image classification. Neurocomputing 395, 150–159 (2020)
43. Zhao N, Wu Z, Lau RWH, Lin S: What makes instance discrimination good for transfer learning? arXiv: 2006.06606 (2021)
44. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J: Models genesis. Medical Image Analysis 67, 101840 (2021) [PubMed: 33188996]
45. Zhuang P, Wang Y, Qiao Y: Learning attentive pairwise interaction for fine-grained classification. Proceedings of the AAAI Conference on Artificial Intelligence 34(07), 13130–13137 (Apr 2020)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

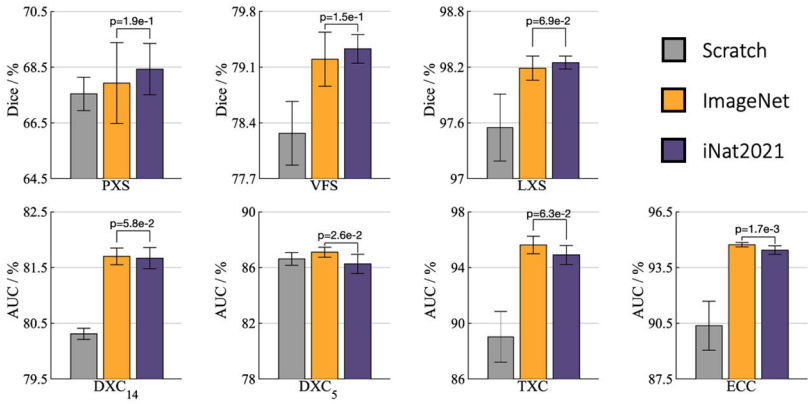

Fig.1.

For segmentation (target) tasks (*i.e.*, PXS, VFS, and LXS), fine-tuning the model pre-trained on iNat2021 outperforms that on ImageNet, while the model pre-trained on ImageNet prevails on classification (target) tasks (*i.e.*, DXC_{14}, DXC_5, TXC, and ECC), demonstrating the effect of data granularity on transfer learning capability: pre-trained models on the fine-grained data capture subtle features that empowers segmentation target tasks, and pre-trained models on the coarse-grained data encode high-level features that facilitate classification target tasks.

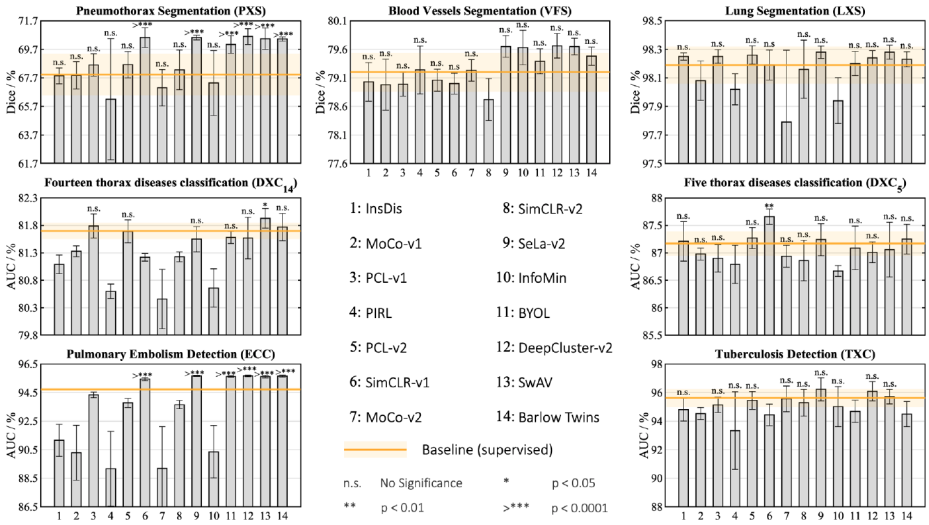

Fig.2.

For each target task, in terms of the mean performance, the supervised ImageNet model can be outperformed by at least three self-supervised ImageNet models, demonstrating the higher transferability of self-supervised representation learning. Recent approaches, SwAV [7], Barlow Twins [41], SeLa-v2 [7], and DeepCluster-v2 [7], stand out as consistently outperforming the supervised ImageNet model in most target tasks. We conduct statistical analysis between the supervised model and each self-supervised model in each target task, and show the results for the methods that significantly outperform the baseline or provide comparable performance. Methods are listed in numerical order from left to right.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Benchmarking transfer learning for seven common medical imaging tasks, spanning over different label structures (binary/multi-label classification and segmentation), modalities, organs, diseases, and data size.

Code[†]	Application	Modality	Dataset
ECC	Pulmonary Embolism Detection	CT	RSNA PE Detection [2]
DXC_{14}	Fourteen thorax diseases classification	X-ray	NIH ChestX-Ray14 [37]
DXC_5	Five thorax diseases classification	X-ray	CheXpert [22]
VFS	Blood Vessels Segmentation	Fundoscopic	DRIVE [5]
PXS	Pneumothorax Segmentation	X-ray	SIIM-ACR [1]
LXS	Lung Segmentation	X-ray	NIH Montgomery [24]
TXC	Tuberculosis Detection	X-ray	NIH Shenzhen CXR [24]

[†]The first letter denotes the object of interest ("E" for embolism, "D" for thorax diseases, etc); the second letter denotes the modality ("X" for X-ray, "F" for Fundoscopic, etc); the last letter denotes the task ("C" for classification, "S" for segmentation).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2.

Domain-adapted pre-trained models outperform the corresponding ImageNet and in-domain models. For every target task, we performed the independent two sample *t*-test between the best (bolded) vs. others. Highlighted boxes in green indicate results which have no statistically significant difference at the $p = 0.05$ level. When pre-training and target tasks are the same, transfer learning is not applicable, denoted by "-". The footnotes compare our results with the state-of-the-art performance for each task.

Initialization	Target tasks				
	DXC_{14}[a]	DXC_5[b]	TXC[c]	PXS[d]	LXS[e]
Scratch	80.31±0.10	86.60±0.17	89.03±1.82	67.54±0.60	97.55±0.36
ImageNet	81.70±0.15	87.10±0.36	95.62±0.63	67.93±1.45	98.19±0.13
ChestX-ray14 [37]	-	**87.40±0.26**	96.32±0.65	68.92±0.98	98.18±0.06
CheXpert [22]	81.99±0.08	-	97.07±0.95	69.30± 0.50	98.25±0.04
ImageNet→ChestX-ray14	-	87.09±0.44	**98.47±0.26**	**69.52±0.38**	98.27±0.03
ImageNet→CheXpert	**82.25±0.18**	-	97.33±0.26	69.36±0.49	**98.31±0.05**

[a] [25] holds an AUC of 82.00% vs. 82.25%±0.18% (ours)

[b] [29] holds an AUC of 89.40% w/ disease dependencies (DD) vs. 87.40%±0.26% (ours w/o DD)

[c] [31] holds an AUC of 95.35% ± 1.86% vs. 98.47% ± 0.26% (ours)

[d] [18] holds a Dice of 68.41% ± 0.14% vs. 69.52% ± 0.38% (ours)

[e] [32] holds a Dice of 96.94% ± 2.67% vs. 98.31% ± 0.05% (ours)