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], and Jianming Liang[1(✉)]

[1] Arizona State University, Tempe, AZ 85281, USA
{mhossei2,fhaghigh,jianming.liang}@asu.edu
[2] Stanford University, Stanford, CA 94305, USA
ruibin@stanford.edu
[3] Mayo Clinic, Scottsdale, AZ 85259, USA
Gotway.Michael@mayo.edu

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

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87722-4_1) contains supplementary material, which is available to authorized users.

S. Albarqouni et al. (Eds.): DART 2021/FAIR 2021, LNCS 12968, pp. 3–13, 2021.
https://doi.org/10.1007/978-3-030-87722-4_1

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 [9,10,23,24,31]. As evidenced by [31], nearly all top-performing models in a wide range of representative medical applications, including classifying the common thoracic diseases, detecting pulmonary embolism, 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, chest X-rays all look similar, therefore, distinguishing diseases and abnormal conditions may rely on some 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 [12], 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 [5,7,14,26,30]. 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 monochromic and consistent in anatomical structures [9,10]. Now, several moderately-sized datasets have been created in medical imaging, for instance, NIH ChestX-Ray14 [25] that contains 112K images; CheXpert [13] that consists of 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., embolism, nodule, tuberculosis, etc.), organs (e.g., lung and 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-

Table 1. We benchmark transfer learning for seven popular 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 [25]
DXC_5	Five thorax diseases classification	X-ray	CheXpert [13]
VFS	Blood vessels segmentation	fundoscopic	DRIVE [4]
PXS	Pneumothorax segmentation	X-ray	SIIM-ACR [1]
LXS	Lung segmentation	X-ray	NIH Montgomery [15]
TXC	Tuberculosis detection	X-ray	NIH Shenzhen CXR [15]

† 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).

adaptive (continual) pre-training [8] 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 details in Appendix A. We considered a diverse suite of 7 challenging and popular medical imaging tasks covering 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 of pathologies 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 [11] 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 [13] and ChestX-Ray14 [25]. 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 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 [17,19,24,27]. 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 [6,29,32]. 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, covering multi-label classification, binary classi-

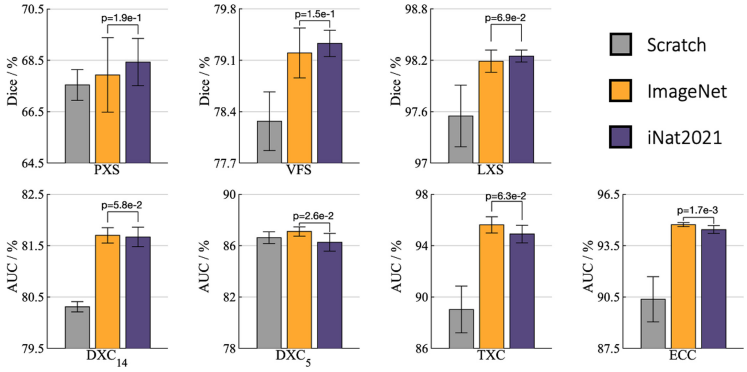

Fig. 1. For segmentation (target) tasks (*i.e.,* `PXS`, `VFS`, and `LXS`), fine-tuning the model pre-trained on iNat2021 outperforms that on ImageNet, while the model pre-trained on ImageNet prevails on classification (target) tasks (*i.e.,* $\texttt{DXC}_{14}$, $\texttt{DXC}_5$, `TXC`, and `ECC`), demonstrating 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.

fication, 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 [12] 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 $\texttt{DXC}_{14}$, $\texttt{DXC}_5$, `TXC`, and `ECC` (see Fig. 1). Contrary to our intuition (see

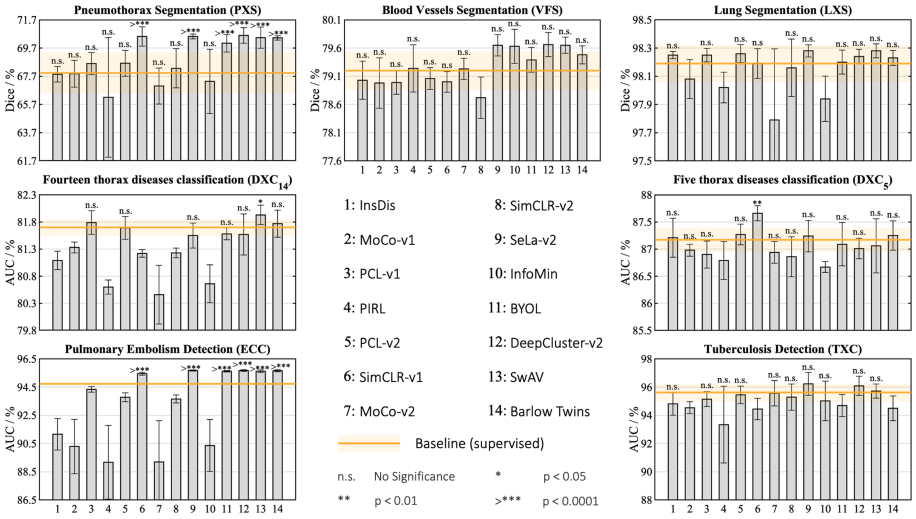

Fig. 2. For each target task, in terms of the mean performance, the supervised ImageNet model can be outperformed by at least three self-supervised ImageNet models, demonstrating the higher transferability of self-supervised representation learning. Recent approaches, SwAV [5], Barlow Twins [28], SeLa-v2 [5], and DeepCluster-v2 [5], 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.

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 [7,14,30]. Self-supervised models, in particular, capture task-agnostic features that can be easily adapted to different domains [14,26], while high-level features of supervised pre-trained models may be extraneous when the source and target data distributions are far apart [30]. We hypothesize this phenomenon is more pronounced in the medical domain, where there is a remarkable domain shift [7] compared to Ima-

geNet. 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, MoCo-v1, MoCo-v2, SimCLR-v1, SimCLR-v2, and BYOL), CL based on JigSaw shuffling (PIRL), clustering (DeepCluster-v2 and SeLa-v2), clustering bridging CL (PCL-v1, PCL-v2, and SwAV), mutual information reduction (InfoMin), and redundancy reduction (Barlow Twins), 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 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 ($\texttt{DXC}_{14}$, $\texttt{DXC}_{5}$, 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 [7,30], 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 [7,14,30], 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

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 [25]	-	**87.40±0.26**	96.32±0.65	68.92±0.98	98.18±0.06
CheXpert [13]	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] [16] holds an AUC of 82.00% vs. 82.25% ± 0.18% (ours)
[b] [18] holds an AUC of 89.40% w/ disease dependencies (DD) vs. 87.40% ± 0.26% (ours w/o DD)
[c] [20] holds an AUC of 95.35% ± 1.86% vs. 98.47% ± 0.26% (ours)
[d] [10] holds a Dice of 68.41% ± 0.14% vs. 69.52% ± 0.38% (ours)
[e] [21] holds a Dice of 96.94% ± 2.67% vs. 98.31% ± 0.05% (ours)

predominant standard for transfer learning as they are free, open-source models which can be used for a variety of tasks [3,9,17,27]. Despite the prevailing use of ImageNet models, the remarkable covariate shift between natural and medical images restrain transfer learning [19]. 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 [8]. 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 [3] 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 Ima-

geNet model 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 [27], 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 $\texttt{DXC}_{14}$, 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 $\texttt{DXC}_{5}$, 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 [8,22].

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.

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 [12,27], 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.

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.

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. Azizi, S., et al.: Big self-supervised models advance medical image classification. arXiv:2101.05224 (2021)
4. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. Int. J. Biomed. Imaging **2013** (2013). Article ID 154860
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882 (2021)
6. Chang, D., et al.: The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans. Image Process. **29**, 4683–4695 (2020)
7. Ericsson, L., Gouk, H., Hospedales, T.M.: 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
8. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks. arXiv:2004.10964 (2020)
9. Haghighi, F., Hosseinzadeh Taher, M.R., Zhou, Z., Gotway, M.B., Liang, J.: Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 137–147. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_14
10. Haghighi, F., Taher, M.R.H., Zhou, Z., Gotway, M.B., Liang, J.: Transferable visual words: exploiting the semantics of anatomical patterns for self-supervised learning. arXiv:2102.10680 (2021)
11. 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)
12. Horn, G.V., Cole, E., Beery, S., Wilber, K., Belongie, S., Aodha, O.M.: Benchmarking representation learning for natural world image collections. arXiv:2103.16483 (2021)
13. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031 (2019)
14. Islam, A., Chen, C.F., Panda, R., Karlinsky, L., Radke, R., Feris, R.: A broad study on the transferability of visual representations with contrastive learning (2021)

15. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. **4**(6), 475–477 (2014)
16. Kim, E., Kim, S., Seo, M., Yoon, S.: XProtoNet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15719–15728 (2021)
17. Mustafa, B., et al.: Supervised transfer learning at scale for medical imaging. arXiv:2101.05913 (2021)
18. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
19. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning with applications to medical imaging. arXiv:1902.07208 (2019)
20. Rajaraman, S., Zamzmi, G., Folio, L., Alderson, P., Antani, S.: Chest X-ray bone suppression for improving classification of tuberculosis-consistent findings. Diagnostics **11**(5) (2021). Article No. 840
21. Reamaroon, N., et al.: Robust segmentation of lung in chest X-ray: applications in analysis of acute respiratory distress syndrome. BMC Med. Imaging **20**, 116–128 (2020)
22. Reed, C.J., et al.: Self-supervised pretraining improves self-supervised pretraining. arXiv:2103.12718 (2021)
23. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
24. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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)
26. Wei, L., et al.: Can semantic labels assist self-supervised visual representation learning? (2020)
27. Wen, Y., Chen, L., Deng, Y., Zhou, C.: Rethinking pre-training on medical imaging. J. Vis. Commun. Image Represent. **78**, 103145 (2021)
28. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. arXiv:2103.03230 (2021)
29. Zhao, J., Peng, Y., He, X.: Attribute hierarchy based multi-task learning for fine-grained image classification. Neurocomputing **395**, 150–159 (2020)
30. Zhao, N., Wu, Z., Lau, R.W.H., Lin, S.: What makes instance discrimination good for transfer learning? arXiv:2006.06606 (2021)
31. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Med. Image Anal. **67**, 101840 (2021)
32. Zhuang, P., Wang, Y., Qiao, Y.: Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 13130–13137 (2020)