Contrastive self-supervised learning from 100 million medical images with optional supervision

Florin C. Ghesu,[a,*] **Bogdan Georgescu**,[a] **Awais Mansoor**,[a] **Youngjin Yoo**,[a] **Dominik Neumann**,[b] **Pragneshkumar Patel**,[a] **Reddappagari Suryanarayana Vishwanath**,[c] **James M. Balter**,[d] **Yue Cao**,[d] **Sasa Grbic**,[a] **and Dorin Comaniciu**[a]

[a]Siemens Healthineers, Digital Technology and Innovation, Princeton, New Jersey, United States

[b]Siemens Healthineers, Digital Technology and Innovation, Erlangen, Germany

[c]Siemens Healthineers, Digital Technology and Innovation, Bangalore, Karnataka, India

[d]University of Michigan, Department of Radiation Oncology, Ann Arbor, Michigan, United States

Abstract

Purpose: Building accurate and robust artificial intelligence systems for medical image assessment requires the creation of large sets of annotated training examples. However, constructing such datasets is very costly due to the complex nature of annotation tasks, which often require expert knowledge (e.g., a radiologist). To counter this limitation, we propose a method to learn from medical images at scale in a self-supervised way.

Approach: Our approach, based on contrastive learning and online feature clustering, leverages training datasets of over 100,000,000 medical images of various modalities, including radiography, computed tomography (CT), magnetic resonance (MR) imaging, and ultrasonography (US). We propose to use the learned features to guide model training in supervised and hybrid self-supervised/supervised regime on various downstream tasks.

Results: We highlight a number of advantages of this strategy on challenging image assessment problems in radiography, CT, and MR: (1) significant increase in accuracy compared to the state-of-the-art (e.g., area under the curve boost of 3% to 7% for detection of abnormalities from chest radiography scans and hemorrhage detection on brain CT); (2) acceleration of model convergence during training by up to 85% compared with using no pretraining (e.g., 83% when training a model for detection of brain metastases in MR scans); and (3) increase in robustness to various image augmentations, such as intensity variations, rotations or scaling reflective of data variation seen in the field.

Conclusions: The proposed approach enables large gains in accuracy and robustness on challenging image assessment problems. The improvement is significant compared with other state-of-the-art approaches trained on medical or vision images (e.g., ImageNet).

© *2022 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.9.6.064503]

Keywords: self-supervised learning; clustering; semi-supervised learning; abnormality assessment.

Paper 22133GRR received May 25, 2022; accepted for publication Nov. 14, 2022; published online Nov. 30, 2022.

1 Introduction

Self-supervised learning has enjoyed much attention in recent years in the vision research community with methods powered by large amounts of data nearing the accuracy level of

*Address all correspondence to Florin C. Ghesu, florin.ghesu@siemens-healthineers.com

2329-4302/2022/$28.00 © 2022 SPIE

state-of-the-art supervised learning strategies on well-known benchmarks, such as ImageNet.[1–3] Moreover, they demonstrate that one can use visual representations derived through self-supervised learning to guide regular downstream supervised learning and achieve increased performance (e.g., via transfer learning).

Few studies have investigated the impact of self-supervised learning in the medical image analysis domain[4–6])—a field where the development of artificial intelligence (AI) technologies is impacted by a high cost of annotations (often requiring expert radiologists precision) and scarcity of medical imaging data. Several solutions focus on architectures for segmentation (i.e., encoder–decoder) and do not support deep architectures often used for classification or detection.[7,8] In addition, the current methods do not exploit truly large datasets and are at best trained with thousands or hundreds of thousands of cases—the same range as many systems trained with supervised learning.[9] In this work, we overcome these limitations by proposing a method for self-supervised learning from medical image data, which enables the training of classification-optimized architectures. In particular, we make a first step towards truly big-data training and break the barrier of 100,000,000 training images.

The contributions of the paper are the following:

- We propose a method for self-supervised learning based on contrastive learning[10] and online feature clustering.[1] The method enables hybrid self-supervised/supervised learning from multimodality data and is applicable to two-dimensional (2D) and three-dimensional (3D) image data. As a core part of the system, we propose a new set of image transformation operations optimized for medical image data. Closest to our work is the contribution of Caron et al.[1]
- We conduct large-scale self-supervised training experiments, including a dataset of over 1,300,000 x-rays (i.e., images as 2D array; also called radiographs) and a dataset of over 105,000,000 multimodality image data [including x-ray, computed tomography (CT), magnetic resonance (MR) imaging, and ultrasonography (US)]. To the best of our knowledge, this represents the largest machine learning experiment to date focused on medical image data that has been reported in the literature.
- We perform a rigorous validation of the method on three medical computer-aided diagnosis problems: (1) chest radiography abnormality assessment, (2) brain metastasis detection in MR, and (3) brain hemorrhage detection in CT data. For this purpose, we use challenging test datasets that are reflective of real clinical practice and with highly curated annotations derived by consensus of multiple expert radiologists. This is an essential step in obtaining an accurate assessment of performance. We intentionally avoid public datasets, such as ChestX-ray8,[11] with reported suboptimal image quality and label error rates of 65% to 85% in terms of sensitivity.[9]
- We compare our approach with different state-of-the-art methods for self-supervised learning and to the common strategy of feature transfer from the ImageNet dataset. We demonstrate that, using the proposed method, one can achieve a considerable performance increase on all the previously enumerated tasks, i.e., significant accuracy increase [average of 6% to 8% area under the curve (AUC)], robustness gain, and acceleration of model training convergence (up to 85%).

The paper is organized as follows: Sec. 2 provides an overview of related work, with the last subsection focusing on recent developments for self-supervised learning in the medical imaging domain; Sec. 3 describes the proposed method followed by Sec. 4 in which we present the experiments on various abnormality detection problems based on different 2D/3D image modalities. Finally, Sec. 5 concludes the paper with a summary and outlook on future work.

2 Background and Motivation

2.1 *Self-Supervised Learning by Contrastive Learning*

Proposed as a principled approach for dimensionality reduction, contrastive learning[10] based on invariant input transformations has become a key optimization strategy for self-supervised

feature learning. Using various transformations of the input data, which determine a series of surrogate classes, Dosovitskiy et al.[12] proposed a supervised discriminative learning approach as a means to learn robust features from unlabeled data. In contrast, Bojanowski and Joulin[13] learned a supervised mapping to a set of target deep representations sampled from an uninformative distribution, referred to as noise-as-targets. Using this strategy, they argue that one can avoid learning trivial feature sets or the effects of feature collapse. One limitation of instance learning is the intractable number of classes, which is proportional to the number of instances.[12] Wu et al.[14] addressed this limitation using a nonparametric approach, which constructs a memory bank to store the target instance representations and applies noise-contrastive estimation (NCE)[15] to compare instances. A memory bank is used also by Zhuang et al.[16] for their local aggregation scheme, designed to optimize the instance representation such that similar data samples are clustered, whereas dissimilar ones become separated in the target manifold. Recently, He et al.[17] proposed to replace the memory bank with a momentum encoder coupled with a queue to generate and store representations for contrastive learning. In contrast, Hjelm et al.[18] proposed to use mutual information maximization based on NCE[15] for unsupervised feature learning—applying adversarial learning to constrain the representation according to a given prior. Bachman et al.[19] extended the approach to optimize the mutual information on multiple feature scales based on so-called multiple views, i.e., different augmentations of the input. Tian et al.[20] further extended the method proposed by Hjelm et al.[18] to support more than two views for an improved performance. Similar principles are applied by Henaff et al.[21] using contrastive predictive coding to learn deep representations from a spatial decomposition of the input image.

2.2 *Self-Supervised Learning by Clustering*

Unsupervised representation learning using clustering[1,16,22,23] is a common alternative to instance learning and contrastive learning. Caron et al.[23] proposed DeepCluster, an end-to-end unsupervised feature learning approach using the *k*-means clustering algorithm as optimization criteria. Coupled with the self-supervised learning method proposed by Gidaris et al.,[24] the method is further enhanced to effectively scale to large uncurated datasets.[25] In their approach Xueting et al.[26] also relied on the *k*-means algorithm, but in a two-stage approach: first cluster assignments are computed from a pretrained model and used as pseudolabels in the second stage for feature learning. In contrast, Huang et al.[27] introduced anchor neighborhood discovery—a divide-and-conquer strategy coupled with curriculum learning for effective sample clustering. Using this optimization criteria, they demonstrate that one can learn representative deep features in an end-to-end manner. Different from this, Asano et al.[28] proposed an effective algorithm for simultaneous feature learning and label inference by maximizing the mutual information between data indices and labels in the form of an optimal transport problem.

2.3 *Learning from Pretext Task*

Another formulation for self-supervised learning reduces the problem to learning from a supervised signal that is artificially constructed to model a pretext task, e.g., solving a Jigsaw puzzle[29,30] or Rubik's cube.[31] Agrawal et al.[32] proposed to use egomotion as supervision, demonstrating that the features learnt from movement prediction are superior to features learned from traditional image labels. Similarly, Misra et al.[33] learned feature representations by estimating the correct temporal order of frames in video captures. Inspired by early approaches for landmark detection via coordinate regression,[34] Doersch et al.[35] proposed to use visual context as supervised signal, learning to estimate the relative position of pairs of patches extracted from given unlabeled images. Noroozi and Favaro[29] proposed as pretext task and artificial Jigsaw puzzle of image tiles. They demonstrate that one can train a deep learning model (in the form of a context free network) to solve the puzzle and thereby learn semantic features. An alternative strategy is feature learning by inpainting using context encoders trained with an adversarial optimization criterion.[36] Finally, Larsson et al.[37] used colorization as pretext task, learning to estimate a per-pixel color histogram.

2.4 *Self-Supervised Learning in the Medical Domain*

Similar principles for self-supervised feature learning are applied in medical image analysis to improve the accuracy and robustness of downstream tasks, e.g., abnormality classification or anatomy segmentation.[4,38] For instance, Chen et al.[39] proposed a commonly known restoration strategy for feature learning from images with artificially swapped local patches. In contrast, Zhou et al.[4] applied various image manipulation steps (nonlinear intensity transformation, local pixel shuffling or in/out-painting) and train an encoder–decoder architecture to reconstruct the original image information thereby learning semantic features in an unsupervised way. With focus on volumetric anatomy segmentation, Chaitanya et al.[5] introduced a framework for self-supervised learning based on a hybrid contrastive loss that learns both global and local image representations. For the same application, Nguyen et al.[6] proposed to use spatial awareness as signal for self-supervised learning—learning to predict the displacement of different image slices after random swaps of image patches between slices. Jiao et al.[40,41] leveraged streams of data (i.e., ultrasound) and matched speech signal to extract features from the data. Finally, Azizi et al.[42] relied on the contrastive learning-based method proposed by Chen et al.[2] to pretrain features and improve the accuracy of various downstream classification tasks from radiography or dermatology images.

3 Proposed Method

In this work, we propose a method for self-supervised learning from medical images based on contrastive learning[10] and online feature clustering.[1] We assume that we are given a dataset $\mathcal{D} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N'}, (\boldsymbol{x}_{N'+1}, \boldsymbol{y}_{N'+1}), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)]$ of N signal samples, e.g., 2D or 3D images $\boldsymbol{x}_k = \boldsymbol{I}_k$; $1 \leq k \leq N$. A subset of $\mathcal{D}$ consists of $N - N'$ samples, which is paired with labels $\boldsymbol{y}_k$ ($N' < k \leq N$), such as: binary image labels, masks, and others. We propose to use this dataset for hybrid self-supervised/supervised model pretraining to learn representative features that can be transferred to downstream use-cases, i.e., used as initialization in a supervised training routine. Figure 1 provides an overview.

3.1 *Online Clustering - Swapped Prediction Optimization*

Following the work of Caron et al.,[1] we use an online clustering strategy coupled with principles of contrastive learning to learn image features in a self-supervised way. Given a family of image augmentation operations $\mathcal{A}$ (described later in Sec. 3.3), the goal is to estimate the visual features parametrized by $\boldsymbol{\theta}$ in the model/projector $f_{\boldsymbol{\theta}}$ (as shown in Fig. 1) via assignment to cluster codes. In particular, this assignment is optimized to be invariant to various hierarchies of augmentation operations sampled from $\mathcal{A}$ and applied to any given image $\boldsymbol{I} \in \mathcal{D}$. The workflow is as follows:

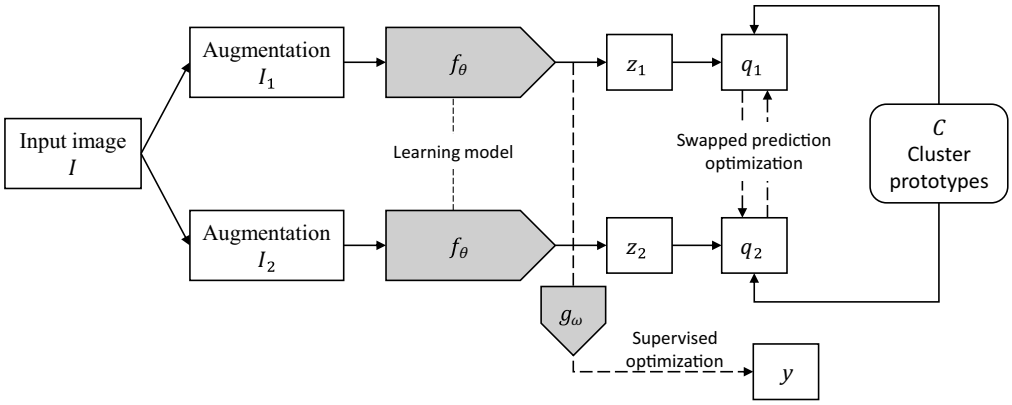

Fig. 1 Schematic overview of the training methodology.

1. Based on an arbitrary image $\boldsymbol{I} \in \mathcal{D}$, two transformed images $[\hat{\boldsymbol{I}}_1, \hat{\boldsymbol{I}}_2]$ are computed using random augmentation operations sampled from $\mathcal{A}$ and applied hierarchically.
2. The nonlinear model/projector f_θ (i.e., the model that we attempt to pretrain) is applied on $[\hat{\boldsymbol{I}}_1, \hat{\boldsymbol{I}}_2]$ to compute features $[\boldsymbol{z}_1, \boldsymbol{z}_2]$ which in turn are assigned to cluster codes $[\boldsymbol{q}_1, \boldsymbol{q}_2]$ from a set of K prototypes $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ (K is a parameter set by the user; empirically selected to ensure stable training; and our proposed method is not sensitive to the value of K, given K is sufficiently large);

Given the pair of features $[\boldsymbol{z}_1, \boldsymbol{z}_2]$ and code pair $[\boldsymbol{q}_1, \boldsymbol{q}_2]$, the self-supervised optimization criterion is formulated using a swapped prediction strategy based on the cross-entropy loss function

$$\mathcal{L}_{\text{ss}}(\boldsymbol{x}) = -\sum_i \boldsymbol{q}_2^{(i)} \log \frac{\exp\left(\frac{1}{\tau}\boldsymbol{z}_1^\top \boldsymbol{c}_i\right)}{\sum_j \exp\left(\frac{1}{\tau}\boldsymbol{z}_1^\top \boldsymbol{c}_j\right)} - \sum_i \boldsymbol{q}_1^{(i)} \log \frac{\exp\left(\frac{1}{\tau}\boldsymbol{z}_2^\top \boldsymbol{c}_i\right)}{\sum_j \exp\left(\frac{1}{\tau}\boldsymbol{z}_2^\top \boldsymbol{c}_j\right)}, \tag{1}$$

where $\tau \in \mathbb{R}$ is a temperature parameter, and $\mathcal{L}_{\text{ss}}$ refers to self-supervised loss. Without loss of generality, $\boldsymbol{z}_1 = f_{\boldsymbol{\theta}}(\hat{\boldsymbol{I}}_1)/\|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{I}}_1)\|_2$ (similar derivation also for $\boldsymbol{z}_2$) and all prototype vectors $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ are trainable. Let $\boldsymbol{C}$ denote a matrix with column vectors defined by the K prototypes. The optimization described in Eq. (1) is performed using stochastic batch-wise sampling of cases from the training set $\mathcal{D}$

$$[\boldsymbol{\theta}^*, \boldsymbol{C}^*] = \underset{\boldsymbol{\theta}, \boldsymbol{C}}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_{\text{ss}}(\boldsymbol{x}_k). \tag{2}$$

In the following sections, we describe in detail the online clustering algorithm, based on two different scenarios: (1) $\mathcal{D}$ consists only of images of one modality (e.g., radiography) and (2) the dataset contains images of multiple modalities (e.g., radiography, US, CT, MR, etc.).

3.1.1 *Single-modality clustering*

Assume a batch size of $B > 1$ samples is used for training. In the following, we focus on one branch of the processing steps depicted in Fig. 1; all projectors are shared on the other branch (without loss of generality, let that be $\boldsymbol{I} \rightarrow \boldsymbol{I}_1 \rightarrow \boldsymbol{z}_1; \boldsymbol{q}_1$). Whereas in Eq. (1), we used codes $\boldsymbol{q}_1, \boldsymbol{q}_2$ in a setting with one sample, virtually a batch-size of $B = 1$; in this case (i.e., $B > 1$) $\boldsymbol{q}_1, \boldsymbol{q}_2$ would become two matrices: $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$; each of size $K \times B$. For simplicity, in the following we use $\boldsymbol{Q}$ to refer to either $\boldsymbol{Q}_1$ or $\boldsymbol{Q}_2$ and describe how they are estimated (same process for each). The set of computed output features $\boldsymbol{z}$ is captured by matrix $\boldsymbol{Z} \in \mathbb{R}^{F \times B}$ where F denotes the size of the any given feature $\boldsymbol{z}$ (column vector). The prototypes are captured by matrix $\boldsymbol{C} \in \mathbb{R}^{F \times K}$ where K denotes the number of prototypes. Finally, the codes that enable the mapping of projected features to prototypes are captured by matrix $\boldsymbol{Q}$ of size $K \times B$. To prevent a trivial solution that would map all images in one batch to the same code, an equipartition constraint is enforced based on the entropy measure

$$\max_{\boldsymbol{Q} \in \mathcal{Q}} \text{Tr}(\boldsymbol{Q}^\top \boldsymbol{C}^\top \boldsymbol{Z}) + \epsilon H(\boldsymbol{Q}), \tag{3}$$

where H denotes the entropy and ϵ the regularization weight.[1] Inspired by the work of Asano et al.,[28] we constrain the solution space Q to ensure that each prototype is selected at least B/K times in one batch. In addition, empirical evidence indicates that using continuous codes is more effective in the online training setting, compared with discretizing the solution. Following the derivation of Caron et al.[1] and optimal transport theory, the solution $\boldsymbol{Q}^* \in \mathbb{R}^{K \times B}$ to Eq. (3) can be determined as a normalized exponential matrix using the Sinkhorn–Knopp algorithm.[43] For more details, we refer the reader to the works of Asano et al.[28] and Cuturi et al.[43]

3.1.2 *Multimodality clustering*

Training with images from multiple modalities is more challenging. Whereas in theory, it may enable the pretraining of robust modality-invariant features that would generalize to a variety of downstream tasks; in practice, simply mixing images of multiple modalities in one single batch impacts the training stability. We hypothesize that a modality specific clustering can alleviate this issue. Let $\mathcal{D}$ contain images of M different modalities. In this case, we propose to partition $\mathcal{D} = \bigcup_{m=1}^{M} \mathcal{D}^{(m)}$ where $\mathcal{D}^{(m)}$ denotes the set of all images on modality indexed by $1 \leq m \leq M$ (e.g., radiography); and M denotes the number of different modalities captured in $\mathcal{D}$. Without loss of generality, let us assume the batch-size B is a multiple of the number of modalities M. We propose to partition each batch of samples in M subsets of equal size, each subset containing only images of one modality sampled randomly from any $\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(M)}\}$. In this case, Eq. (3) can be adapted to

$$\sum_{m=1}^{M} \max_{\boldsymbol{Q}_m \in \mathcal{Q}} \mathrm{Tr}(\boldsymbol{Q}_m^{\top} \boldsymbol{C}_m^{\top} \boldsymbol{Z}^{(m)}) + \epsilon H(\boldsymbol{Q}_m), \tag{4}$$

with $\boldsymbol{Q}_m, \boldsymbol{C}_m$ conditioned on modality m; and $\boldsymbol{Z}^{(m)}$ denoting the aggregate of B/M vectors $\boldsymbol{z}$ associated with modality m [the remaining variables follow the same definition as in Eq. (3)]. The intuition behind this equation is the same as for Eq. (3) with the additional summation term over the M different modalities. The same logic can be applied to adapt Eq. (1).

3.2 *Hybrid Self-Supervised - Supervised Learning*

As we defined dataset $\mathcal{D}$, $N - N'$ cases are associated with labels, e.g., provided by human (expert) annotators, extracted via natural language processing, or other automatic means from the image, associated clinical reports, or other corresponding non-imaging data. Recall, for any arbitrary training sample $\boldsymbol{x}_k (k > N')$, we denote the corresponding label as $\boldsymbol{y}_k$. We propose to dynamically learn from these labels in a joint self-supervised/supervised strategy. Let $\boldsymbol{g}_{\boldsymbol{\omega}}$ be a deep neural network projector parametrized by $\boldsymbol{\omega}$, mapping for any such sample $\boldsymbol{x}_k$ from features of model $\boldsymbol{f}_{\boldsymbol{\theta}}$ (output features and/or intermediate layer features) to an output $\hat{\boldsymbol{y}}_k$

$$\hat{\boldsymbol{y}}_k = \boldsymbol{g}_{\boldsymbol{\omega}}(\boldsymbol{x}_k | \boldsymbol{f}_{\boldsymbol{\theta}}), \qquad \forall\ k > N'. \tag{5}$$

In this case, the goal is similar to any supervised learning problem, i.e., minimize the distance of $\hat{\boldsymbol{y}}_k$ to $\boldsymbol{y}_k$ according to a loss function $\mathcal{L}_{\text{sup}}$ (sup $\equiv$ supervised)

$$[\boldsymbol{\theta}^*, \boldsymbol{\omega}^*] = \underset{\boldsymbol{\theta}, \boldsymbol{\omega}}{\mathrm{argmin}} \frac{1}{N - N'} \sum_{k=N'+1}^{N} \mathcal{L}_{\text{sup}}[\boldsymbol{g}_{\boldsymbol{\omega}}(\boldsymbol{x}_k | \boldsymbol{f}_{\boldsymbol{\theta}}), \boldsymbol{y}_k]. \tag{6}$$

We combine Eqs. (2) and (6) to a single global optimization criterion, using the self-supervised learning signal $\mathcal{L}_{\text{ss}}$ and the supervised learning signal $\mathcal{L}_{\text{sup}}$ for input samples $\boldsymbol{x}_k$ associated with labels $\boldsymbol{y}_k$, $1 \leq k \leq N$. We rebalance the contribution of each these signals using factors $\alpha, \beta \in \mathbb{R}$

$$[\boldsymbol{\theta}^*, \boldsymbol{\omega}^*, \boldsymbol{C}^*] = \underset{\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{C}}{\mathrm{argmin}} \tfrac{1}{N} \textstyle\sum_{k=1}^{N} \alpha \mathcal{L}_{ss}(\boldsymbol{x}_k) + \mathbf{1}_{k>N'} \beta \mathcal{L}_{\text{sup}}[\boldsymbol{g}_{\boldsymbol{\omega}}(\boldsymbol{x}_k | \boldsymbol{f}_{\boldsymbol{\theta}}), \boldsymbol{y}_k]. \tag{7}$$

3.3 *Augmentation Strategies*

To determine the input to the model during training $[\boldsymbol{I}_1, \boldsymbol{I}_2]$, random augmentation operations are applied from a set of augmentation operations defined as $\mathcal{A}$. These operations are as listed below.

Image rescaling is applied to a fraction s of the size of the original image $\boldsymbol{I}$ as $\Psi_s(\boldsymbol{I})$ with Ψ denoting the rescaling function and $s \in \mathbb{R}$ sampled uniformly at random from the interval $[0.1, 0.7]$. The rescaling factor s is sampled from a large range of possible values to encourage the learning of robust scale-invariant features.

Energy-based augmentation is performed based on the image normalization algorithm proposed by Philipsen et al.[44] Following their methodology, an image $\boldsymbol{I}$ is decomposed into B energy bands $\boldsymbol{I}^{(1)}, \boldsymbol{I}^{(2)}, \ldots, \boldsymbol{I}^{(B)}$ using Gaussian filtering. For each band $1 \leq i \leq B$, the energy value $e_i(\boldsymbol{I}, \Omega)$ is computed as the brightness dispersion in a predefined image region defined by Ω (in our case, the entire image). The normalized image is computed as

$$\hat{\boldsymbol{I}}(\Omega) = \sum_{i=1}^{B} \lambda_i(\Omega) \boldsymbol{I}^{(i)} = \sum_{i=1}^{B} \frac{e_i^{\text{ref}}(\Omega)}{e_i(\boldsymbol{I}, \Omega)} \boldsymbol{I}^{(i)}, \tag{8}$$

where $e_i^{\text{ref}}(\Omega) = \frac{1}{R} \sum_{k=1}^{R} e_i(\boldsymbol{I}_k, \Omega)$ denotes the reference energy value on band i with $\boldsymbol{I}_1, \boldsymbol{I}_2 \ldots \boldsymbol{I}_R$ denoting R preselected reference images. We set $R = 1000$ and propose to augment the image using a variable reference energy $e_i^{\text{ref}}(\Omega)$ for any given band $1 \leq i \leq B$ around the mean value. Concretely, on each band, we propose to model the distribution of the R reference values using a Gaussian distribution and sample the value of the reference energy from the range $[-\sigma, +\sigma]$ around the mean energy value.

Intensity rescaling is applied in two different ways: (1) nonlinear rescaling using a gamma transform with the exponent $\gamma \in \mathbb{R}$ sampled uniformly at random from the interval [1.8, 2.6] and (2) linear rescaling of the intensity as $a * \boldsymbol{I} + b$ with $a \in [0.9, 1.1]$ a random uniform sample and b restricted to $\pm 20\%$ of the intensity range of $\boldsymbol{I}$.

Cropping from random image locations (sampled uniformly at random) is the final augmentation applied.

4 Experiments and Results

4.1 *Datasets for Self-Supervised Training*

We constructed several datasets for self-supervised training (based on 2D and 3D image modalities). The weights of models trained on these datasets using self-supervision were then transferred to initialize models that were optimized using supervision on downstream tasks. The datasets are the following:

- 2D x-ray dataset $\mathcal{D}_X$ containing 1,297,699 x-ray images capturing various anatomies, including chest, spine, back, arm, leg, and more. The data are acquired from both public[11,45–53] (accessed as of September 1, 2021) and internal sources.
- 2D mixed modality dataset $\mathcal{D}_M$ containing 105,006,320 images/slices of various anatomies (head, abdomen, chest, legs, etc.) and from various imaging modalities, including the x-ray dataset $\mathcal{D}_X$. Except the public data contained in $\mathcal{D}_X$ as described in the previous section, this dataset contains only internal data. The proportions per modality are: 72% CT slices, 25% MRI slices, 1% x-ray, and 2% US images. Around 0.6% of $\mathcal{D}_M$ is acquired from public sources (the sources as listed in the definition of $\mathcal{D}_X$).
- 3D CT dataset $\mathcal{D}_{\text{CT}}$ containing 24,440 3D CT volumes coming from 1,345,040 DICOM slices of noncontrast CT (NCCT) head scans aquired from internal sources.

4.2 *Training Hyperparameters, Infrastructure, and Scaling*

Different architectures have been investigated as part of our experiments. In the 2D context we investigated variants of residual networks[7] (including ResNet-152 and ResNet-50 with several variants denoted as ResNet-50w2/w4 as described by Caron et al.[1]). Training hyperparameters are defined in Table 1. Several parameters are inherited from swapping assignments between multiple views (SwAV),[1] and, for model details, we refer the reader to that reference. The training infrastructure consists of 4 nodes (each with 8 Volta GPUs with 16GB GPU memory, 80 cores, and 512 GB main memory). All nodes are connected via InfiniBand. The system uses the Quobyte file system for parallel and distributed IO operation, and PyTorch distributed functionality is applied to scale the training to multiple nodes.

Table 1 Hyperparameters for self-supervised training in 2D/3D. For the 2D use-case, the number/size of crops can be interpreted as: two random crops at 224×224 pixels and six random crops at 92×92 pixels based on the multicrop strategy introduced by Caron et al.[1] In comparison, for 3D, the crop sizes are $224 \times 224 \times 192$ and $112 \times 112 \times 96$ voxels. Image size range denotes the possible sizes of the image, i.e., number of rows/columns after rescaling (at least 600 and at most 1200 in 2D). Up to 10% of the image border may be cropped on all sides. The assigned crops variable denotes the indices of the so-called standard resolution crop as defined the the multicrop strategy[1] (in this case, the first two crops at indices 0 and 1). The interpretation of these variables is similar for the 3D use-case.

Parameter	2D experiments	3D experiments
Number of crops	[2, 6]	[1, 4]
Size of crops	$[(224 \times 224), (92 \times 92)]$	$[(224 \times 224 \times 192), (112 \times 112 \times 96)]$
Image size range	[600, 1200]	[(112, 112, 96), (250, 250, 200)]
Cropped image border	10%	10%
Assigned crops	[0, 1]	[0, 1]
Temperature	0.1	0.1
Epsilon	0.03	0.03
Number of Sinkhorn iterations	3	3
Feature dimension	128	128
Number of prototypes	3000	1500
Queue length	3840	1920
Epoch queue starts	10	3
Epochs	100	100
Batch size	32	4
Start learning rate	0.6	0.6
Final learning rate	0.0006	0.0006
Freeze prototypes	10,000	4000
Weight decay	10^{-6}	10^{-6}

4.3 *Abnormality Detection in Chest Radiography*

We focused on the detection of lung lesions (i.e., nodules/masses) and pneumothorax from frontal chest radiographs. These are critical findings, lesions with potential long-term relevance (e.g., pulmonary malignancy, cancer), whereas pneumothorax as acute finding often immediately endangers the life of the patient. As proposed in our previous work,[54,55] we approach the problem as a multiclass detection problem using bounding boxes to isolate the abnormalities. The model architecture is fully convolutional and multiscale and is inspired by the work of Tian et al.[20] The entire content of the image is processed in one single forward pass and labeled bounding boxes (with an associated probability) around relevant abnormalities are predicted (see Fig. 2). A total of 11730 chest radiographs were used for training. The data were acquired from internal data sources. Various abnormalities (including lesion and pneumothorax) were annotated by expert radiologists on each image using bounding boxes (see Figs. 3 and 4). Further details related to the training data and training routine can be found in our previous work.[54,55]

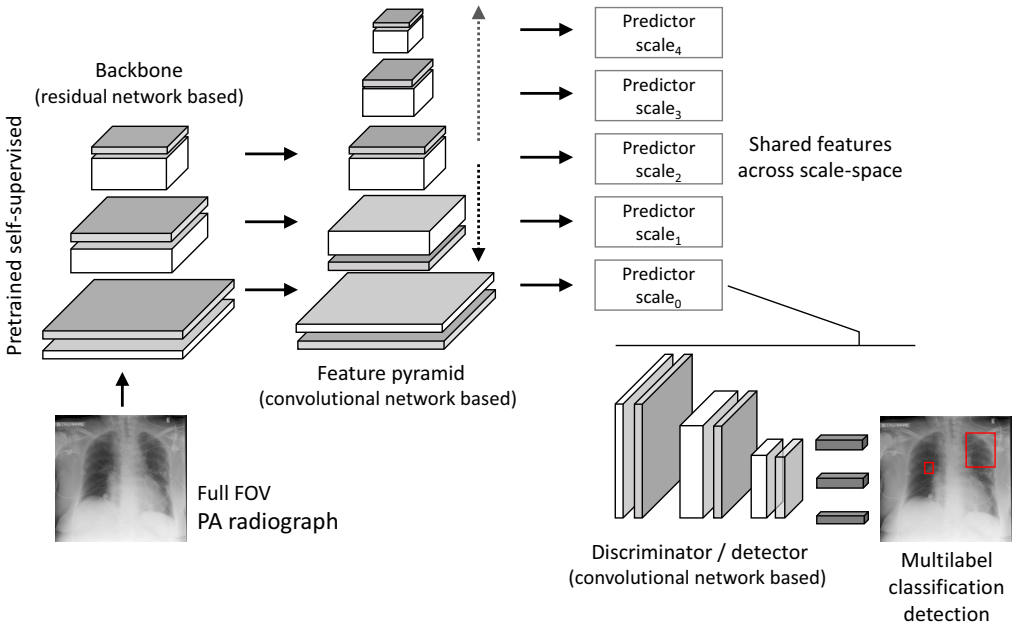

Fig. 2 Architecture used for classification and detection of lesion and pneumothorax in chest radiographs. The backbone (ResNet-50w2) is pretrained using self-supervision.

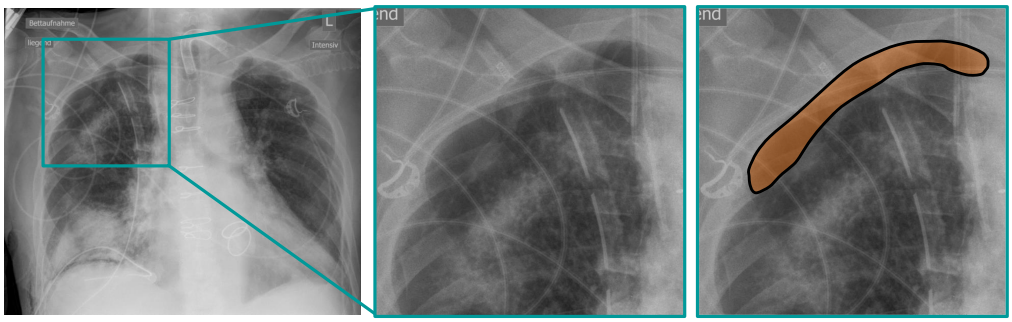

Fig. 3 Example case from the pneumothorax test set. Left: chest radiograph with right upper zone showing a lucent area towards the apex with vaguely defined pleural line suggestive of pneumothorax. ICD tube along with numerous lines make the distinction of the pneumothorax difficult; middle: image subregion, capturing the pneumothorax; and right: curve highlighting the dehiscent visceral pleural separated from the thoracic wall—indicating the pneumothorax.

For testing the pneumothorax detection feature, we use a test set of 321 chest radiographs, all acquired at bedside in anterior–posterior view—capturing severely ill patients, covered by tubes and/or wires with 34 cases acquired in the intensive care unit. The ground truth is determined by a consensus of three expert radiologists. Each image is first read independently by each reader, followed by a joint discussion to determine the final annotation. As pneumothoraces can be very small (to a few millimeters in sectional width) or obscured by other structures, e.g., the ribs, they can easily be overlooked. Using three readers, we intend to minimize that risk. Of the 321 radiographs, 125 are identified as pneumothorax positive using 148 bounding boxes to isolate the abnormal anatomy. Figure 3 shows an example.

For testing the lesion detection feature, we use a test set of 288 radiographs from the LIDC dataset.[50] Each radiograph is paired with a CT scan acquired in close time proximity. The information from this additional modality is used to improve the ground truth quality. As not all lesions captured in CT are also visible in chest radiography,[56] we propose two protocols to derive two versions of the ground truth:

1. Synchronous reading of chest radiograph and corresponding CT by an expert radiologist, marking on the radiograph only lesions that are visible. We denote this version of the ground truth as LIDC-synch.
2. A staged approach is applied. In the first stage, three expert radiologists read independently all chest radiographs and mark lesions (without looking at CT). Subsequently all marks are aggregated by an additional radiologist, removing any duplicate marks of the same lesion. These candidate marks are then assessed in a second stage using CT as point of reference—for each mark on the radiograph, if the CT displays a lesion at that location that mark is positive; if not, the mark is removed. We denote this version of the ground truth as LIDC-staged.

With 146 (out of 288) positive radiographs and 187 lesions marked, LIDC-synch is a significantly more sensitive ground truth than LIDC-staged (111 positive radiographs with 133 lesions marked). However, LIDC-staged has the benefit of removing any bias related to the assessment of visibility on radiographs (a step performed in the derivation of LIDC-synch). Example images and annotations are shown in Fig. 4.

Receiver operating characteristic (ROC)-AUC is used to assess the classification performance, the accuracy of the bounding box detection is assessed using free-response receiver operating characteristic (fROC). In particular we report the sensitivity at instance level (captured in the y-axis in fROC) averaged at the following average numbers of false-positive per image (captured in the x-axis in fROC): [0.01, 0.05, 0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0]. We compare our solution with five alternative approaches:

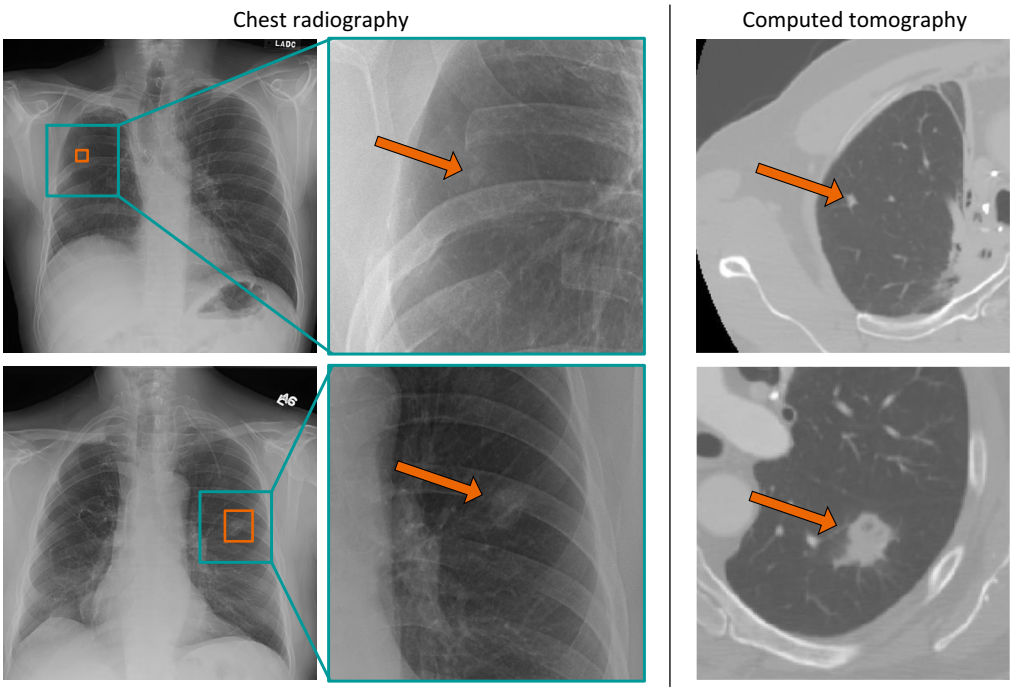

Fig. 4 Example cases from lesion test set. Left: chest radiographs with lesions highlighted using a red bounding box and middle: image subregions with arrows indicating the lesions. Right: axial slices of the corresponding CT scan with arrows indicating the same lesions. The first row shows an image of a patient who has sustained a right thoracotomy. In the resection area, there is a very subtle lesion that was marked positive only in LIDC-synch. In the generation of LIDC-staged, it was missed by all three readers when reading the radiograph to generate candidates for CT-confirmation, and as such the case is marked as negative in LIDC-staged. In contrast, the lesion captured in the second row is much larger and easier to see. It is marked as positive in both LIDC-synch and LIDC-staged.

1. Supervised learning[7] - training on the ImageNet dataset.[3]
2. Self-supervised learning[1] (denoted as SwAV) - training on the ImageNet dataset.
3. Sim contrastive learning of visual representations (SimCLR) method for self-supervised learning[2] - training on the internal medical image data.
4. Self-supervised learning[1] (denoted as SwAV-med) - training on the internal medical image data (in our result comparisons between methods 3 and 4 and our proposed method, we use the same dataset for self-supervised training and downstream training, and test under similar conditions).
5. Using no pretraining, relying on a random initialization of the network weights (in our experience, often used for medical image analysis applications).

For more simplicity in the interpretation of the results, we often focus the comparison of our method with method (1) proposed by He et al.[7]—the previously best strategy for this task. Tables 2 and 3 give an overview of the performance of all reference methods. Please note, we used the same pretrained model for lesion and pneumothorax experiments.

Table 2 AUC performance for lesion detection on LIDC-staged when using 100%, 50%, 25%, or 10% of the training data. Selection of the subsets is done randomly.

	AUC performance (LIDC-staged)[a]			
Method	100%	50%	25%	10%
No pretraining	0.77	0.73	0.65	0.53
SimCLR[2]	0.90	0.88	0.82	0.79
SwAV[1]	0.90[b]	0.89	0.85	0.80
SwAV-med[1]	0.91[b]	0.88	0.84	0.82
Supervised ImageNet[7]	0.91[b]	0.89	0.82	0.80
Ours	0.94[b]	0.91	0.85	0.85

[a]Average of five models selected with highest AUC on validation set. For each entry, the downstream training was executed five times, using every time the exact same options as specified for this task. For each training round, we selected the model that maximizes the AUC on the validation set.The final entry is the average AUC performance of these five models on the test set, specifically LIDC-staged.
[b]Over five training rounds standard deviation of AUC ≤ 0.003 for each, indicating the high stability of the performance measurement

Table 3 Performance for lesion detection on LIDC-staged using different pretraining methods. After each performance number, we report in parenthesis the 95% confidence interval. The operating point is selected to achieve a specificity close to 95%.

Method	AUC	Precision	Sensitivity	Specificity
SimCLR[2]	0.896 (0.858 to 0.930)	0.850 (0.754 to 0.932)	0.459 (0.367 to 0.551)	0.949 (0.916 to 0.978)
SwAV[1]	0.901 (0.862 to 0.934)	0.857 (0.769 to 0.946)	0.486 (0.393 to 0.577)	0.949 (0.916 to 0.983)
SwAV-med[1]	0.908 (0.873 to 0.942)	0.887 (0.810 to 0.958)	0.568 (0.471 to 0.657)	0.951 (0.922 to 0.983)
Supervised ImageNet[7]	0.913 (0.877 to 0.942)	0.878 (0.798 to 0.947)	0.586 (0.496 to 0.670)	0.949 (0.913 to 0.978)
Ours	0.944 (0.913 to 0.967)	0.902 (0.832 to 0.961)	0.748 (0.664 to 0.825)	0.949 (0.912 to 0.982)

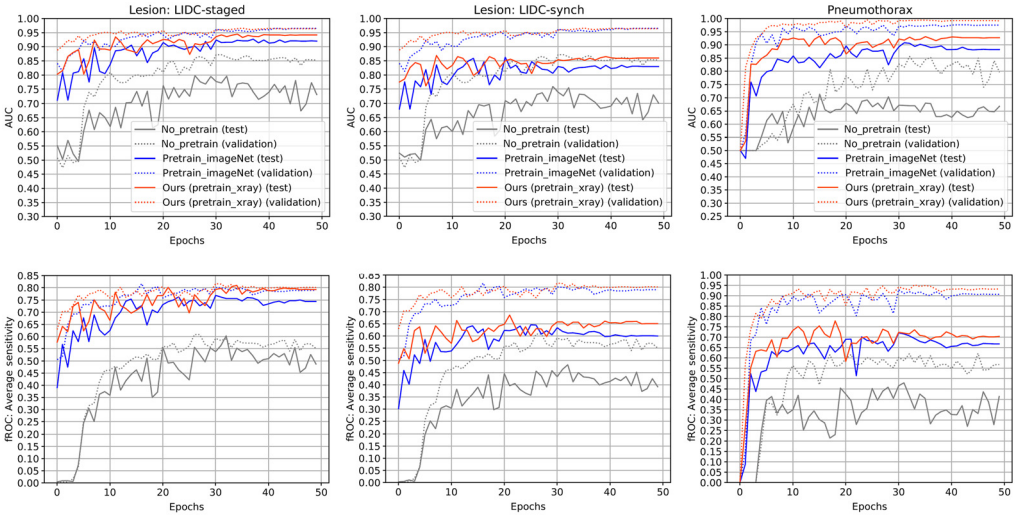

Fig. 5 Performance evolution of lesion and pneumothorax detection in chest radiographs during training (left: lesion on LIDC-staged, middle: lesion on LIDC-synch, and right: pneumothorax). Our solution (initializing the model with self-supervised pretrained weights on the x-ray dataset $\mathcal{D}_X$) significantly outperforms both in terms of AUC and average instance detection sensitivity the previously best pretraining strategy, i.e., supervised pretraining on ImageNet.[7] The difference is significant, ranging between 3% and 5%. The difference is much larger when compared with using no-pretraining—ranging between 20% and 25% on the lesion test and almost 30% on the pneumothorax test data.

Figure 5 shows the performance evolution for lesion and pneumothorax detection on the test and validation sets during training. We highlight the AUC and average instance level fROC sensitivity as a function of the epochs. A significant increase in average performance is achieved for the test set.

Figure 6 shows the acceleration in convergence speed when using our method compared to conventional supervised pretraining or no pretraining. The speedup is at least 50% (also on both lesion test sets); both when analyzing evolution of performance on test data and validation data. Finally, Fig. 7 highlights the increased robustness of self-supervised pretrained models after down-stream fine-tuning with respect to certain image variations, which are typically observed in practice (e.g., image rotations/scaling or intensity variations).

4.4 *Brain Metastases Detection in MRI*

Automated detection and segmentation of brain metastases in 3D MRI scans could support therapy workflows. In this study, we conducted another experiment on slice-wise brain metastasis on contrast-enhanced magnetization-prepared rapid acquisition with gradient echo (MPRAGE) scans, which can be used for treatment selection, planning and longitudinal monitoring by guiding radiosurgery protocols and other treatment decisions. However, this task remains challenging due, in part, to the scarcity of training data containing metastatic tissue in an MRI volume, which makes learning clinically meaningful image features from scratch challenging. A reliable pretrained model may have the potential to mitigate this limited data problem. Thus we focused on analyzing the impact of self-supervised training as pretraining on classifying 2D slices with metastases in MPRAGE volumes. We utilized a 2.5D encoder-decoder network to first obtain a segmentation mask showing potential areas of suspected metastases.[57] The output segmentation mask was subsequently used as an attention channel along with the 5 input slices to train a 2.5D classification network to perform a slice-wise classification. The architecture of the classification network followed the concept of ResNet-50w2. The six-channel input was compressed to make three channels by a 1×1 convolution. Our dataset included a training set (341 cases), a validation set (36 cases), and a test set (43 cases). The details about the data and preprocessing can be found in our previous work.[58]

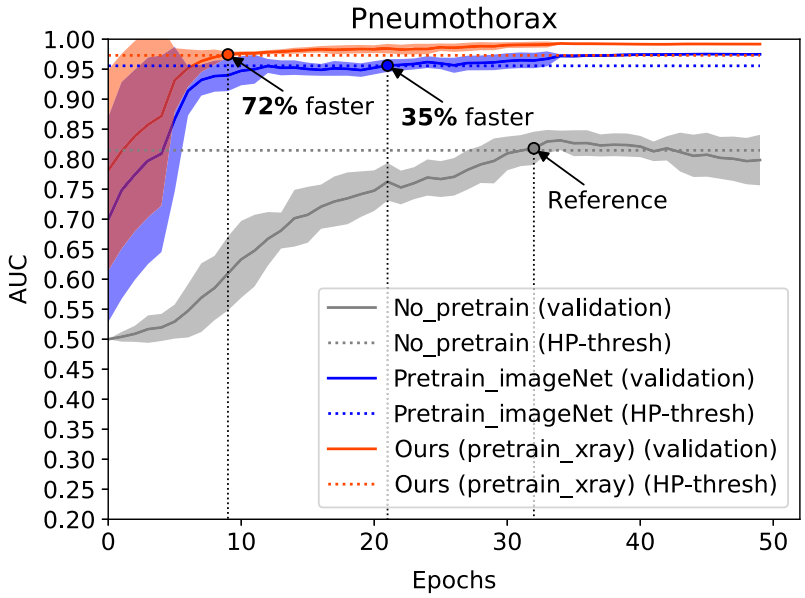

Fig. 6 Visualization of the training speed-up of the model for pneumothorax. We denote the convergence point as the earliest epoch during training where 98% of the final best performance is achieved (denoted as HP-thresh, i.e., high performance threshold). Using our method, one can achieve convergence 72% faster than using no pretraining. The transparent area along each curve denotes the standard deviation.

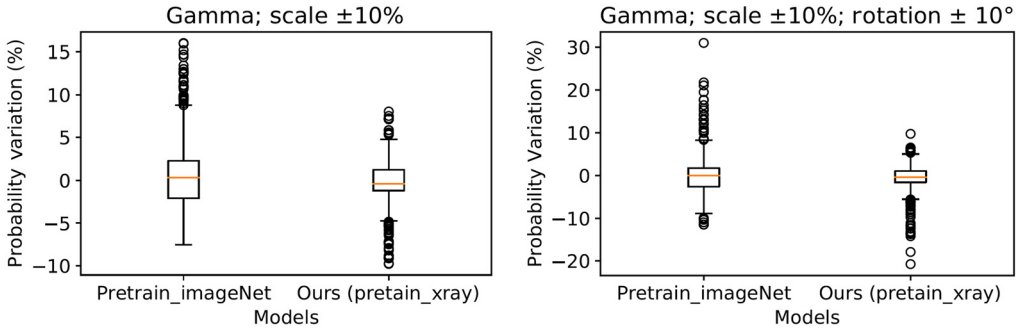

Fig. 7 Using self-supervised pretraining on $\mathcal{D}_X$ leads to an increase in robustness compared with using features pretrained on ImageNet.[7] The box plot shows the relative deviation in probability for lesion (at case level) when applying various augmentations, such as gamma transform with $\gamma \in [1.8, 2.6]$ and/or random image rotation/scaling. The distribution is shown for a random set of 50 radiographs with 50 random transformations per image (i.e., 2500 data points).

The metastatic slice detection performance was evaluated by ROC-AUC and mean average precision. Figure 8 shows the training evolution using detection AUC measured on the validation dataset for each epoch. Training the model from scratch reached AUC 90% after 300 epochs. On the other hand, training the model initialized with the pretrained ResNet-50w2 achieved AUC 92% to 93% within 10 epochs. Also the pretrained model consistently outperformed the model without pretraining by AUC 2% to 3%. The testing AUC with our pretraining method was 93.2% which was higher than both the model without pretraining and the model pretrained with SwAV[1] by 2% as shown in Table 4. To evaluate with other metrics, we binarized the prediction score with thresholds selected to provide 80% in sensitivity. At the same sensitivity level, precision with our pretraining method was 0.702 which was higher than the model without pretraining by about 0.19 and than the model pretrained with SwAV[1] by about 0.12. Accuracy with our pretraining method was 0.912 that was higher over the model without pretraining and the model pretrained

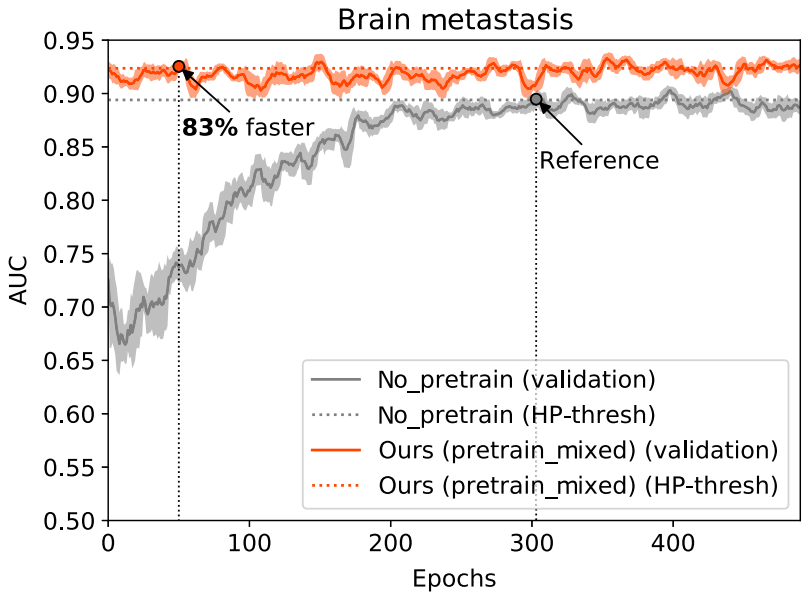

Fig. 8 Validation AUC evolution over training epochs for brain metastatic slice detection. Using our self-supervised pretraining leads to an increase in training convergence rate and validation AUC.

Table 4 Performance comparison for slice-wise brain metastasis detection on 8145 testing MPRAGE slice images (1336 metastatic images). AUC: area under the ROC curve. For the metrics requiring binarization, operation points were selected to have about 80% sensitivity. 95% confidence interval for each metric is shown in parenthesis.

Method	AUC	Precision	Accuracy	Sensitivity	Specificity
No pretraining	0.913 (0.906 to 0.921)	0.515 (0.497 to 0.533)	0.843 (0.837 to 0.850)	0.796 (0.778 to 0.812)	0.853 (0.845 to 0.860)
SwAV[1]	0.913 (0.903 to 0.921)	0.578 (0.561 to 0.597)	0.872 (0.865 to 0.878)	0.804 (0.786 to 0.821)	0.885 (0.879 to 0.891)
Ours	0.932 (0.924 to 0.939)	0.702 (0.683 to 0.722)	0.912 (0.907 to 0.917)	0.804 (0.785 to 0.823)	0.933 (0.928 to 0.938)

with SwAV (0.843 and 0.872, respectively). The model with our pretraining also achieved the highest specificity (0.933). Figure 9 shows examples of metastatic slice detection in a patient with brain metastasis.

4.5 *Brain Hemorrhage Detection in CT*

Current standard of care for evaluation of patients presenting with stroke symptoms or following head trauma involves assessment of NCCT scans for presence of hemorrhage. Accurate and timely detection of head bleeding is critical to start the appropriate treatment as soon as possible such as starting administration of thrombolytics for stroke patients or surgical intervention for trauma patients. Automated detection of hemorrhage in NCCT scans[59,60] has the potential to minimize the time it takes for a patient to receive the appropriate treatment.

In this work, we investigate the gains of automated AI-based detection of hemorrhage in 3D NCCT scans using both auxiliary tasks as well as self-supervised pretraining of a 3D network. The input dicoms are stacked together in a 3D volume, we then reformat the input volume to be in axial orientation, and a set of head landmarks are used to crop the brain region and scale it in a box of dimensions $40 \times 224 \times 192$. The Hounsfield units are normalized to (0,1) using a

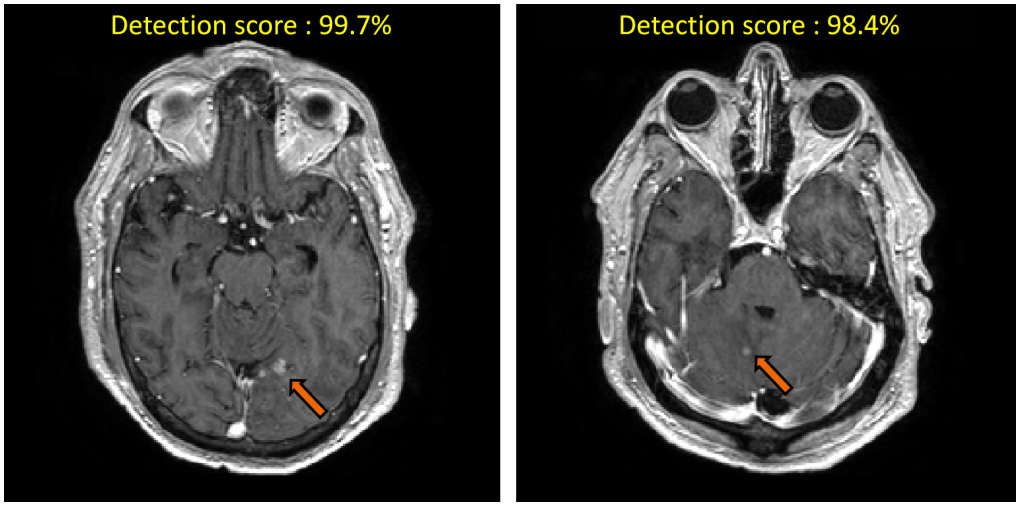

Fig. 9 Example postcontrast MPRAGE slices from a patient with brain metastases. The images show metastatic slices and their detection scores by the model pretrained with our self-supervised learning. Image courtesy: University of Michigan.

transformation with a (center, window) = (55, 200). For feature extraction we employ a 3D densely connected network with variable input size and five dense blocks with (1,2,3,3,3) units, each unit having 3D convolution, 3D batch normalization and LeakyReLU activation layers with 16 initial features and a growth rate of 7. The first two blocks only process data in-plane with (1,3,3) kernels and downsample only in (x, y) plane and the last three process data with full (3,3,3) 3D kernels. The final features of fixed dimension 1024 are computed through adaptive pooling (both max and average pooling). The self-supervised pretraining is performed on dataset $\mathcal{D}_{\text{CT}}$ using the hyper-parameters specified in Table 1.

A set of 3017 3D volumes are used for validation and model selection, and a set of 2945 volumes are used for testing (both datasets coming from patients not included in pretraining). The main task is hemorrhage detection for each 3D volume and we have used as auxiliary tasks training with hemorrage types labels (subarachnoid, subdural, epidural, intraventricular, and intraparenchymal hemorrhages) and with presence/absence of hemorrhage for each slice. Figure 10 illustrates the AUC and accuracy of the base network and the improvement by training with auxiliary tasks as well as with self-supervised pretraining. The network trained with only

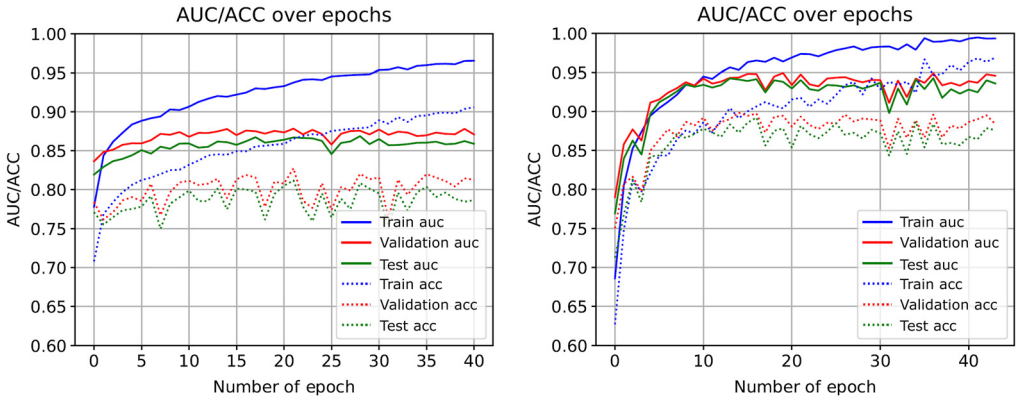

Fig. 10 Performance of hemorrhage detection on 3D NCCT: using auxiliary tasks and self-supervised pretraining of a 3D network on $\mathcal{D}_{\text{CT}}$ leads in an increase in performance. The auxiliary tasks include the use of hemorrhage types and hemorrhage labels on each slice. The left figure illustrates AUC and accuracy for the base model trained only on 3D hemorrhage labels for training, validation and testing data splits. The right figure illustrates the performance for the model trained using auxiliary tasks and self-supervision. It shows a significant increase of the AUC from 0.88/0.87 (validation/test) to 0.95/0.94.

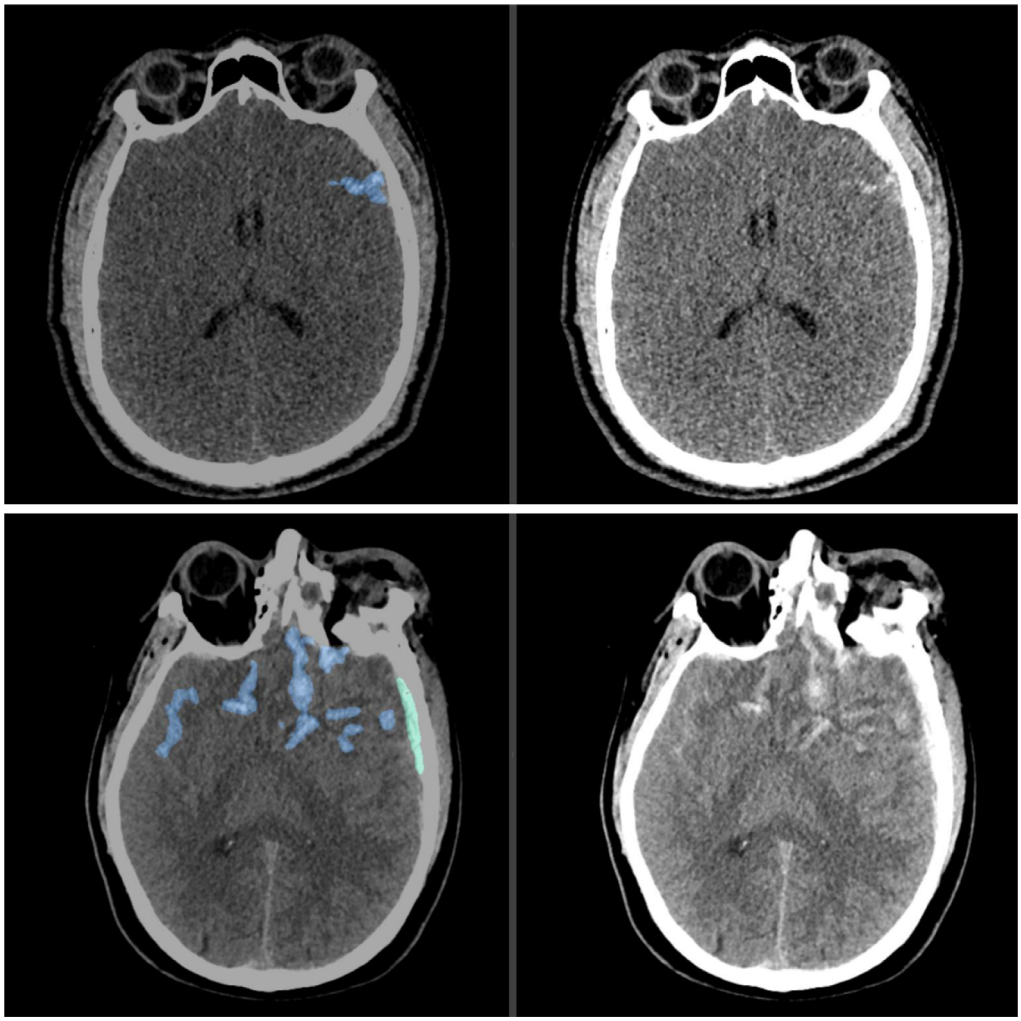

Fig. 11 Examples of hemorrhage detection on 3D NCCT on various types of hemorrhage. On the left, it is illustrated with color that overlays the hemorrhage region and on the right, the NCCT image slice.

presence/absence of hemorrhage for the whole 3D volume achieves an AUC of 0.88/0.87 for the validation/test sets, whereas the network trained also with the auxiliary tasks and the self-supervision achives and AUC of 0.95/0.94, respectively, that is an 8% increase in performance. Furthermore, if a segmentation head is added to the feature space and both axial and coronal NCCT acquisitions are used the best performance reaches an AUC of 0.97 with a sensitivity of 92% and a specificity of 94% on the testing set. Figure 11 illustrates different types of hemorrhage successfully detected by the system. Table 5 shows the performance gains by sequentially using each of the auxiliary tasks and self-supervision—the best performance is obtained when all are used with self-supervision.

4.6 *Limitations and Directions of Future Research*

Further optimization and research is required in different directions: (1) with a training time of 6.5 to 14 days (depending on the training dataset - $\mathcal{D}_X$, $\mathcal{D}_M$, or $\mathcal{D}_{\text{CT}}$) further optimization and better scalability of the training is required to execute more training rounds, and perform more ablative analysis. This has limited the amount of experiments and analysis; (2) once the previous point is addressed, more work is needed to investigate the effectiveness of the proposed training technique on more diverse models and (3) more dedicated focus is needed to investigate the utility of self-supervised learning in tracking and registration tasks in which often models are very small and shallow to ensure high efficiency.

Table 5 Performance of hemorrhage detection on 3D NCCT showing the AUC, sensitivity, specificity and 95% confidence intervals for validation/test when only 3D labels are used and adding types, slice labels, and the self-supervision pretraining of the same 3D network. Operating point has been selected such that sensitivity and specificity are equal.

Labels	3D	3D and types	3D, types, and slice	3D, types, slice, and self-supervision
AUC	0.879/0.867 (0.865 to 0.891)/ (0.853 to 0.881)	0.890/0.875 (0.878 to 0.902)/ (0.861 to 0.887)	0.928/0.918 (0.918 to 0.938)/ (0.908 to 0.929)	0.949/0.943 (0.941 to 0.957)/ (0.934 to 0.952)
Sensitivity	0.805/0.770 (0.782 to 0.826)/ (0.743 to 0.795)	0.817/0.783 (0.795 to 0.840)/ (0.758 to 0.809)	0.853/0.827 (0.832 to 0.875)/ (0.803 to 0.849)	0.886/0.876 (0.867 to 0.904)/ (0.855 to 0.896)
Specificity	0.805/0.803 (0.788 to 0.821)/ (0.786 to 0.821)	0.817/0.803 (0.801 to 0.834)/ (0.786 to 0.821)	0.852/0.865 (0.835 to 0.868)/ (0.850 to 0.880)	0.885/0.876 (0.870 to 0.899)/ (0.861 to 0.890)

A limitation of this study is the fact that 99.4% of the data used for self-supervised training is internal and cannot be shared with the community. We tried to compensate for this limitation by evaluating the method on public data from the LIDC collection—to allow the open comparison of future methods with ours in terms of performance. It is our hope that this study will inspire the future collection of such large datasets that can be openly shared with the community to drive research and innovation.

5 Conclusion

In conclusion, we propose an effective technique for self-supervised learning based on contrastive learning and online clustering, with support for hybrid self-supervised/supervised learning and multi-modality training data (2D and 3D). In addition, we demonstrate the scalability of the method on a large dataset of over 105,000,000 images, and highlight the impact of the learned image representations in improving the accuracy (average of 6% to 8% AUC), robustness and training speed (up to 85%) on various downstream tasks. Using our method, one can achieve higher performance compared with using other state-of-the-art techniques for self-supervised learning or using supervised pretraining on natural images (e.g., ImageNet).

Disclosures

The authors declare no conflicts of interest.

Acknowledgments

Data were obtained from the TB portals,[61] which is an open-access TB data resource supported by the National Institute of Allergy and Infectious Diseases (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) in Bethesda, Maryland, United States. These data were collected and submitted by members of the TB Portals Consortium.[62] Investigators and other data contributors that originally submitted the data to the TB portals did not participate in the design or analysis of this study. The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health and their critical role in the creation of the free publicly available LIDC/IDRI database used in this study. This work was supported in part by the National Institutes of Health (Grant No. R01 CA262182-01). The concepts and information presented in this paper are based on research results that are not commercially available.

References

1. M. Caron et al., "Unsupervised learning of visual features by contrasting cluster assignments," in *Adv. Neural Inf. Process. Syst., 33*, Curran Associates, Inc. (2020).
2. T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, PMLR, Vol. **119**, pp. 1597–1607 (2020).
3. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 248–255 (2009).
4. Z. Zhou et al., "Models genesis: generic autodidactic models for 3D medical image analysis," *Lect. Notes Comput. Sci.* **11767**, 384–393 (2019).
5. K. Chaitanya et al., "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Adv. Neural Inf. Process. Syst., 33*, Curran Associates, Inc. (2020).
6. X.-B. Nguyen et al., "Self-supervised learning based on spatial awareness for medical image analysis," *IEEE Access* **8**, 162973–162981 (2020).
7. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
8. Z. Tian et al., "FCOS: fully convolutional one-stage object detection," in *IEEE Int. Conf. Comput. Vision*, IEEE Computer Society, pp. 9627–9636 (2019).
9. S. Gündel et al., "Robust classification from noisy labels: integrating additional knowledge for chest radiography abnormality assessment," *Med. Image Anal.* **72**, 102087 (2021).
10. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Vol. **2**, pp. 1735–1742 (2006).
11. X. Wang et al., "ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3462–3471 (2017).
12. A. Dosovitskiy et al., "Discriminative unsupervised feature learning with convolutional neural networks," in *Adv. Neural Inf. Process. Syst., 27*, Curran Associates, Inc. (2014).
13. P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in *Int. Conf. Mach. Learn.*, JMLR.org, pp. 517–526 (2017).
14. Z. Wu et al., "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3733–3742 (2018).
15. M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: a new estimation principle for unnormalized statistical models," in *Int. Conf. Artif. Intell. and Stat.*,Y. W. Teh and M. Titterington, Eds., Vol. **9**, pp. 297–304, PMLR (2010).
16. C. Zhuang, A. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *IEEE Int. Conf. Comput. Vision*, IEEE Computer Society, pp. 6001–6011 (2019).
17. K. He et al., "Momentum contrast for unsupervised visual representation learning," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 9726–9735 (2020).
18. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Int. Conf. Learn. Represent.* (2019).
19. P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Adv. Neural Inf. Process. Syst., 32*, Curran Associates, Inc. (2019).
20. Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *Lect. Notes Comput. Sci.* **12356**, 776–794 (2020).
21. O. Henaff, "Data-efficient image recognition with contrastive predictive coding," *in Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, PMLR, Vol. **119**, pp. 4182–4192 (2020).
22. M. A. Bautista et al., "CliqueCNN: deep unsupervised exemplar learning," in *Adv. Neural Inf. Process. Syst., 29*, Curran Associates, Inc. (2016).
23. M. Caron et al., "Deep clustering for unsupervised learning of visual features," *Lect. Notes Comput. Sci.* **11218**, 139–156 (2018).
24. S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Int. Conf. Learn. Represent.* (2018).

25. M. Caron et al., "Unsupervised pre-training of image features on non-curated data," in *IEEE Int. Conf. Comput. Vision*, IEEE Computer Society, pp. 2959–2968 (2019).
26. X. Yan et al., "ClusterFit: improving generalization of visual representations," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 6508–6517 (2020).
27. J. Huang et al., "Unsupervised deep learning by neighbourhood discovery," in *Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, PMLR, Vol, **97**, pp. 2849–2858 (2019).
28. Y. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *Int. Conf. Learn. Represent.* (2020).
29. M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," *Lect. Notes Comput. Sci.* **9910**, 69–84 (2016).
30. D. Kim et al., "Learning image representations by completing damaged jigsaw puzzles," in *IEEE Winter Conf. Appl. Comput. Vision*, IEEE Computer Society, pp. 793–802 (2018).
31. J. Zhu et al., "Rubik's Cube+: a self-supervised feature learning framework for 3D medical image analysis," *Med. Image Anal.* **64**, 101746 (2020).
32. P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *IEEE Int. Conf. Comput. Vision*, IEEE Computer Society, pp. 37–45 (2015).
33. I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," *Lect. Notes Comput. Sci.* **9905**, 527–544 (2016).
34. S. K. Zhou, J. Zhou, and D. Comaniciu, "A boosting regression approach to medical anatomy detection," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1–8 (2007).
35. C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE Int. Conf. Comput. Vision*, IEEE Computer Society, pp. 1422–1430 (2015).
36. D. Pathak et al., "Context encoders: feature learning by inpainting," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2536–2544 (2016).
37. G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," *Lect. Notes Comput. Sci.* **9908**, 577–593 (2016).
38. F. Navarro et al., "Evaluating the robustness of self-supervised learning in medical imaging," (2021).
39. L. Chen et al., "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.* **58**, 101539 (2019).
40. J. Jiao et al., "Self-supervised representation learning for ultrasound video," in *IEEE Int. Symp. Biomed. Imaging*, IEEE Computer Society, pp. 1847–1850 (2020).
41. J. Jiao et al., "Self-supervised contrastive video-speech representation learning for ultrasound," *Lect. Notes Comput. Sci.* **12263**, 534–543 (2020).
42. S. Azizi et al., "Big self-supervised models advance medical image classification," (2021).
43. M. Cuturi, "Sinkhorn distances: lightspeed computation of optimal transport," in *Adv. Neural Inf. Process. Syst., 26*, Curran Associates, Inc. (2013).
44. R. H. H. M. Philipsen et al., "Localized energy-based normalization of medical images: application to chest radiography," *IEEE Trans. Med. Imaging* **34**(9), 1965–1975 (2015).
45. A. E. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data* **6**, 317 (2019).
46. A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation* **101**(23), e215–e220 (2000).
47. A. Rosenthal et al., "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *J. Clin. Microbiol.* **55**(11), 3267–3282 (2017).
48. D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Inf. Assoc.* **23**(2), 304–310 (2016).
49. A. Bustos et al., "Padchest: a large chest x-ray image dataset with multi-label annotated reports," *Med. Image Anal.* **66**, 101797 (2020).
50. S. G. Armato, III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Med. Phys.* **38**(2), 915–931 (2011).

51. S. Jaeger et al., "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantum Imaging Med. Surg.* **4**(6), 475–477 (2014).
52. S. S. Halabi et al., "The RSNA pediatric bone age machine learning challenge," *Radiology* **290**(2), 498–503 (2019).
53. P. Rajpurkar et al., "MURA: large dataset for abnormality detection in musculoskeletal radiographs," (2018).
54. J. Rudolph et al., "Artificial intelligence in chest radiography reporting accuracy: added clinical value in the emergency unit setting without 24/7 radiology coverage," *Investig. Radiol.* **57**, 90–98 (2021).
55. J. Rueckel et al., "Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training," *Eur. Radiol.* **31**, 7888–7900 (2021).
56. E. J. M. Barbosa, Jr. et al., "Automated detection and quantification of COVID-19 airspace disease on chest radiographs: a novel approach achieving expert radiologist-level performance using a deep convolutional neural network trained on digital reconstructed radiographs from computed tomography–derived ground truth," *Investig. Radiol.* **56**, 471–479 (2021).
57. F. C. Ghesu et al., "Quantifying and leveraging predictive uncertainty for medical image assessment," *Med. Image Anal.* **68**, 101855 (2021).
58. Y. Yoo et al., "Evaluating deep learning methods in detecting and segmenting different sizes of brain metastases on 3D post-contrast T1-weighted images," *J. Med. Imaging* **8**(3), 037001 (2021).
59. M. Arbabshirani et al., "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration," *NPJ Digit. Med.* **1**, 2398–6352 (2018).
60. W. Kuo et al., "Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning," *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22737–22745 (2019).
61. https://tbportals.niaid.nih.gov
62. https://tbportals.niaid.nih.gov/Partners

Biographies of the authors are not available.