Zeng *et al. BioMedical Engineering OnLine* (2024) 23:107
https://doi.org/10.1186/s12938-024-01299-9

BioMedical Engineering OnLine

REVIEW Open Access

Self-supervised learning framework application for medical image analysis: a review and summary

Xiangrui Zeng[1*], Nibras Abdullah[2*] and Putra Sumari[1]

*Correspondence:
xavierzeng@student.usm.my;
n.faqera@arabou.edu.sa

[1] School of Computer Sciences, Universiti Sains Malaysia, USM, 11800 Pulau Pinang, Malaysia
[2] Faculty of Computer Studies, Arab Open University, Jeddah, Saudi Arabia

Abstract

Manual annotation of medical image datasets is labor-intensive and prone to biases. Moreover, the rate at which image data accumulates significantly outpaces the speed of manual annotation, posing a challenge to the advancement of machine learning, particularly in the realm of supervised learning. Self-supervised learning is an emerging field that capitalizes on unlabeled data for training, thereby circumventing the need for extensive manual labeling. This learning paradigm generates synthetic pseudo-labels through pretext tasks, compelling the network to acquire image representations in a pseudo-supervised manner and subsequently fine-tuning with a limited set of annotated data to achieve enhanced performance. This review begins with an overview of prevalent types and advancements in self-supervised learning, followed by an exhaustive and systematic examination of methodologies within the medical imaging domain from 2018 to September 2024. The review encompasses a range of medical image modalities, including CT, MRI, X-ray, Histology, and Ultrasound. It addresses specific tasks, such as Classification, Localization, Segmentation, Reduction of False Positives, Improvement of Model Performance, and Enhancement of Image Quality. The analysis reveals a descending order in the volume of related studies, with CT and MRI leading the list, followed by X-ray, Histology, and Ultrasound. Except for CT and MRI, there is a greater prevalence of studies focusing on contrastive learning methods over generative learning approaches. The performance of MRI/Ultrasound classification and all image types segmentation still has room for further exploration. Generally, this review can provide conceptual guidance for medical professionals to combine self-supervised learning with their research.

Keywords: Self-supervised, Medical image, Computer vision, CNN, Transformer

Introduction and background

Medical imaging is an interdisciplinary field that harnesses specific substances—like X-rays, electromagnetic fields, and ultrasound—to interact with the human body, capturing its structure and density through visual representations. These diagnostic images are invaluable in clinical settings, offering insights that aid physicians and surgeons in diagnosing and assessing medical conditions. Over the past few decades, the

exponential growth in the volume of medical images has outstripped the availability of human resources, highlighting an urgent need for the integration of automated diagnostic and analytical systems into the medical imaging workflow. Such systems are poised to enhance diagnostic efficiency and ensure the precision of medical evaluations, thereby fortifying the quality of patient care.

Advancements in the field of computer vision (CV) have provided a novel and promising solution for medical image analysis and diagnosis to help doctors in patient care and treatment. Since the release of AlexNet [1] in 2011, CV has begun to enter the stage of deep neural networks structured such as convolutional neural networks (CNNs), which makes deep learning a popular tool for medical image analysis. Especially after the release of Vision Transformer (ViT) [2] in 2020, CV research entered a new era where two backbone architectures—convolutional neural networks and transformers—coexist as mainstream approaches. These deep network-based Computer-Aided Diagnosis (CAD) systems have gradually demonstrated accuracy comparable to or even exceeding that of human experts, including X-ray [3–9], computed tomography (CT) [10–20], magnetic resonance imaging (MRI) [21–27], histology [28–36], ultrasound [37], and hybrid [38–48].

However, many successful studies have adopted the pattern of supervised learning in the past, which requires tens of thousands of data and corresponding labels for training. These labels seriously affect the model performance, and the majority require manual annotation. Annotating training data costs the labor of numerous imaging experts, which is expensive and time-consuming. In pathological images, subjective bias can lead to the low quality of labels [49], and manual labeling of 3D images can pose security issues of privacy exposure [23]. In addition, the performance and generalization ability are limited by the size of the dataset, leading to a lack of practical significance in training models using small datasets. Overall, the annotation and quantity limitations of data when using supervised learning to train models have become the main challenges for deep neural networks in medical image diagnosis applications, limiting research on constructing effective models in different clinical use cases.

Transfer learning has become a new attempt to alleviate these challenges. Transfer learning first pre-trains the neural network on a large dataset (such as ImageNet [50]), then adds an adaptation layer [51] to the network and fine-tunes the whole with relatively smaller labeled data to improve performance [52]. Transfer learning can reuse pre-trained neural networks in different tasks, saving training time and shrinking the need for labeled data what is more. The key to this mechanism working is that the image features and low-level statistics learned by the network during pre-training can be reused [53]. Transfer learning has achieved good results in the natural image field but performs poorly in transferring from the natural to the medical image field [54]. This may be because the features of natural images are not similar to those of medical images. Besides, the number of labels in natural is much larger than that of medical images (e.g., imagine-1 k [55] has a thousand categories), which may result in the medical image field not requiring such many classes of pre-trained models [56].

To overcome the generalization problem of transfer learning in the medical field, and driven by the goal of reducing manual labeling, self-supervised learning (SSL) is starting to be the research preamble for the pre-selected connectivity paradigm in CV research.

SSL uses unlabeled datasets and the pretext task to learn semantic features in images and does not require manual labeling [24]. A small amount of labeled data is used to fine-tune the pre-trained network and achieve satisfactory performance. This upstream–downstream training mechanism (pre-training and fine-tuning) allows the use of a large number of unlabeled datasets (as close as possible to the type of target task) from the Internet to participate in the pre-training to improve the performance while requiring only a small amount of labeled data (the same type with the target task) with supervised learning to improve the performance further. With the success of SSL in the natural image field, more and more researchers deem SSL as a promising way to utilize large amounts of unlabeled data in clinical practice and medical research. In recent years, many studies have demonstrated the effectiveness of SSL in medical image research tasks, such as classification, segmentation, false-positive reduction, and image enhancement. Hence, it is valuable to summarize these SSL studies in the medical image field and the most appropriate implementation strategies.

This review will look back at the recent works on SSL and the results of its role in medical image diagnosis, and discuss the shortcomings and further directions of current research in medical imaging. The review aims to provide a general reference for imaging practitioners and researchers in the CV field, or readers interested in both fields, to understand SSL. This review will first provide a background on the characteristics and development of SSL and then implement a comprehensive review of 59 recent works in which SSL has been applied to medical image diagnosis, covering X-rays, CT, MRI, histology, and ultrasound images. Similar SSL reviews have been conducted before us, such as Huang et al. [57] and VanBerlo et al. [58]. Compared to their work, this review updates the research that emerged after 2022 and includes multiple tasks, such as segmentation, image enhancement, and false-positive reduction besides classification. Compared to VanBerlo et al., this review contains histological image studies and excludes articles that partially or all use unpublished datasets, so all included studies have a reproducible basis.

Materials and methods

Study outline

This review mainly focuses on the research of SSL in computer vision, involving the fields of natural and medical images. Since SSL nowadays uses deep neural networks almost as the basic algorithm, the research covered is based on CNN or ViT as backbones, and some complex models may also have both. The research on SSL will be summarized into two categories: contrastive learning and generative learning. The initial section delved into studies about the natural image domain, establishing a foundation for the discussion, followed by a summary of the application of SSL in medical images. After that, the enumeration of the respective strengths and limitations of each study is accompanied by a comparative analysis. In the conclusive segment, the review discusses previous research and draws conclusions and future recommendations.

Data acquisition

The data for the review mainly selected peer-reviewed journal and conference articles publicly published between January 1, 2018, and September 10, 2024. For SSL papers in

the natural image domain, this review is only cited to demonstrate the categories and developments of SSL, so that would only include representative articles with milestone significance. For the medical image, Google Scholar and PubMed are selected as databases, and the search keywords used are "self-supervised learning" + "Medical image", "X-ray", "CT", "MRI", "Histology", or "Ultrasound". The search results were checked by the title and abstract to determine their relevance, and irrelevant articles were excluded. In the scrutiny of SSL research about medical images, the review process deliberately excluded studies that relied solely or partially on proprietary datasets, and those that partially utilized public datasets without clear delineation of the utilized segments, due to the infeasibility of comparative analysis. Ultimately, 46 articles were retained for detailed examination and discussion within this review.

Review

Background: deep neural network and self-supervised learning

In the past decade, computer vision has been firmly bound to deep neural networks, and almost all research is based on a specific deep neural network algorithm—CNN or ViT. These two structures were created for computer vision tasks, and many improved algorithms have been derived based on them. This section reviews these two basic neural network structures to provide readers unfamiliar with deep learning with a more specific understanding of them. Afterward, the working mechanism of SSL will also be introduced.

CNN

CNN is a neural network architecture suitable for image tasks, and its conception emerged in 1989 [59]. Unlike a Multilayer Perceptron (MLP), a CNN consists mainly of convolutional and pooling layers. The sensory field of the topmost convolutional layer already covers the whole image, so the CNN can extract any helpful attribute and has translational invariance (i.e., the target can be recognized no matter where it is located in the image). Each convolutional layer contains more than one filter, which can be tuned as hyperparameters. The pooling layer aims to reduce the size of the feature map by iterating and retaining only the maximum or average values in the latent space. Convolutional and pooling layers are stacked on top of each other hierarchically, allowing the CNN to extract basic visual features that can ultimately be used for a specific target task. The more layers of CNN, the more refined the abstract information that represents the image. This gives CNN greater robustness to image variations, scaling, and distortion relative to MLP and greater applicability in CV. Its advanced performance has been achieved in datasets such as ImageNet, which is known as a typical representative technique in supervised learning.

CNN models usually have multilayer networks, the last few layers being MLP with BackPropagation (BP). The convolutional and pooling layers scan the entire pixel matrix in left-to-right and top-to-bottom directions, where the pooling layer can be in the form of Max, Min, and Average. As an example of a pixel matrix, the computation of the convolution layer and the maximum pooling layer is shown in Fig. 1, where the convolution layer uses a dot product. The convolution kernel size is not fixed and is usually an odd number, such as a 3×3 matrix [60]. Besides, the stride of

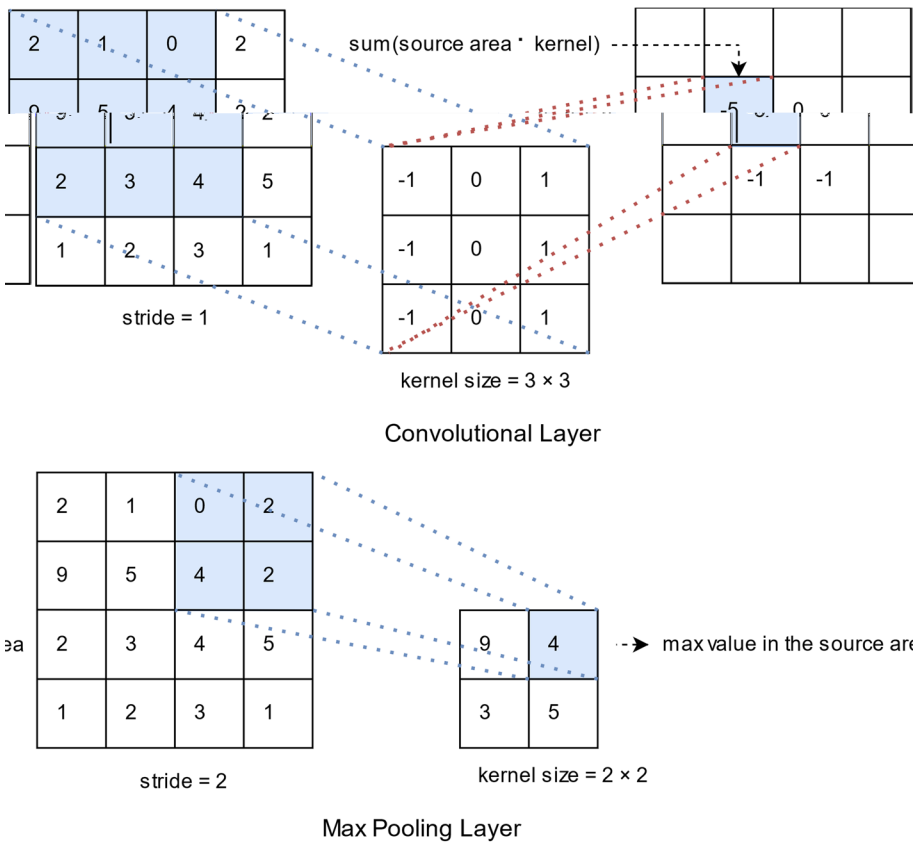

Fig. 1 Computational procedure for convolutional and pooling layers

the convolution and pooling kernel scans is not fixed, as long as it does not exceed the side length of the convolution/pooled kernel.

Since the emergence of AlexNet, CNN has skyrocketed to several times that of classical machine learning algorithms in the image field. Following this, a series of classical models arose based on AlexNet. In medical image diagnostics, there have been countless studies based on CNNs from 2012 to the present, generating tens of thousands of papers annually. The mainstream CNN models are divided into two groups called Visual Geometry Group (VGG) [61] (which focuses on deepening the network) and NetWork In NetWork (NIN) [62] (which focuses on enhancing the function of the convolutional module) and combined to produce the Inception ResNet [63] eventually. However, a problem covering all CNN models can be identified from the computational principles of the CNN convolutional layer in the above paragraph, i.e., it is difficult to establish connections between pixels beyond the range of the convolutional kernel. For example, the pixel regions in the upper left and lower right corners cannot overlap each other's perceptual fields regardless of the convolutional layers, which makes it difficult for CNNs to focus on global information. This difficult-to-improve defect leads to the limitations of CNNs in image semantic understanding.

Vision transformer

ViT is a model for applying the transformer [64] to image classification proposed by the Google team in 2020 [2]. Although it is not the first paper that uses a transformer for vision tasks, it still became a landmark work in applying the transformer in CV since its model is simple, effective, and scalable. When the training dataset is not large enough, ViT usually performs worse than ResNet [60] of the same parameters scale because it lacks two inductive biases compared to CNN. One is localization, i.e., neighboring regions on the image have similar features, and the other is a translation equivariance. However, the most central conclusion in the ViT paper is that when there is enough data for pre-training, ViT could break through the limitation of inductive biases and achieve better migration results in downstream tasks than CNN [2]. Meanwhile, the self-attention mechanism of ViT can overcome the limitations of CNN, extracting the context dependency relationship in the picture easily, which brings a revolutionary change in the CV field.

Transformer distinguishes itself from CNN through its self-attention mechanism, similar to looking up a dictionary. Taking the standard ViT structure as an example, ViT divides the input image into multiple patches ($n \times 16 \times 16$), and then projects each patch into a fixed-length vector (Fig. 2), adds a class token to these vectors, and then adds random position vectors to all the patches. The internal structure of the encoder is shown in Fig. 3, in which the whole encoder can be concatenated as needed, and the number of concatenated encoders is a significant parameter of the transformer. In Multi-head Attention, the number of heads is also an important parameter, which determines the size of the transformer together with the number of encoders.

Transformer has led to significant advances in several areas of vision, such as target detection [65], video classification [66], image classification [67], and image generation [2]. In addition to the standard ViT, improved ViT algorithms such as Swin Transformer

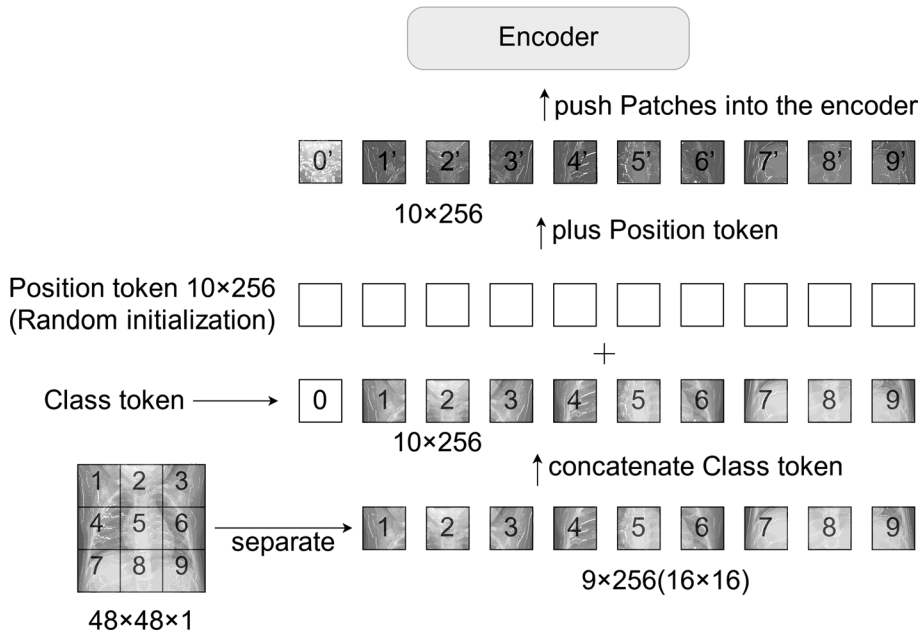

Fig. 2 Example of input patches generation for ViT

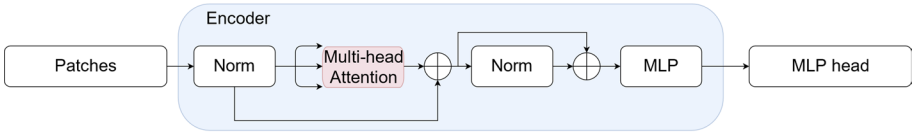

Fig. 3 The workflow of ViT

[68], which draws inspiration from CNN ideas, and DETR [65], which can better focus on the relationships within images, have also emerged one after another. Some of these models applying ViT techniques for recognition can meet or even exceed the results of SOTA solutions in this field. In the field of medical images, some researchers have directly discarded CNNs and used ViT only to build networks [69] toward one force breaking all laws direction. On CT datasets, the transformer retains more spatial information than CNNs, and models using it not only outperform CNNs [70, 71] but are also faster to train [72]. Transformer also performs well on 2D types like X-ray images and even outperform human experts on anomaly detection tasks [73]. It is far from reaching a performance bottleneck today and has no insurmountable drawbacks. Transformer is better suited for CV tasks than CNN during sufficient data, and it represents the future direction and a broad path.

Self-supervised learning

Supervised learning requires data labels to train some or all of the model, which requires a certain amount of labeled data to ensure initialization. SSL neither requires labeled data nor is an extensive learning paradigm like unsupervised learning. It creates an artificial pretext task that allows the model to learn feature extraction as accurately as supervised learning. SSL has been broadly categorized into two paths, called generative learning and comparative learning.

Generative learning The easiest way to understand generative learning is to treat it as a destruction/reconstruction effort, with the most common form being generative adversarial networks (GAN). The model crops or distorts an intact image X to Z and restores it to its original state ($G(Z) = X$). As the model is training, it learns the most important image features (e.g., the difference between a cat and a dog is in the ears, mouth, and eyes) and thus can parse the emerging image. Generative learning networks usually contain encoders and decoders, with which they are competing with each other. The encoder refines the input into high-level information, and the decoder reduces this information to the original input. The two networks work together to optimize until they are stable, and the encoder is usually utilized for some task of interest after training is complete, such as image style conversion [74].

Contrastive learning Contrastive learning usually aims to enable network learning to parse high-level representations of images, in which similar instances are more tightly coupled together in latent space because of akin features, while different instances are relatively far away. (For example, any photograph of a person is characteristic of that person, while photos of others can be recognized even if they have the same pose in

the same place.) Contrastive learning does not need to focus on the details of instances and only abstractly distinguishes various representations, which are more generalizable. Furthermore, contrastive learning represents a framework that allows different backbone types. In the case of the CNN backbone, two instances are passed to two networks of generally the same structure to output $U(I_1)$ and $V(I_2)$ for comparison. The two networks share the same weights or use momentum updating [75] to update one network by the other. The contrast loss of the outputs forces two networks learning, i.e., the smaller the loss of the same class or the larger the loss of the different class, the better (Loss $= U(I_1) - V(I_2) \rightarrow 0$).

Typical categories of self-supervised learning

There is a giant amount of unlabeled data in the world, and ways to harness it can truly realize the potential of the internet. SSL is a subfield of machine learning and does not require human markup for training, unlike supervised learning. The regular workflow of SSL is shown in Fig. 4, which learns the representation from an upstream pretext task and transfers the representation parse ability to the downstream to solve the target task of interest. The pretest task replaces the role of labels in model training, so unlabeled data from any source not necessarily relevant to the target task can be utilized. The upstream of SSL pre-trains the network, and then the pre-trained weights are fine-tuned using specific data in the downstream. The pretext and the target task domains are not necessarily the same, similar to transfer learning. However, SSL is best pre-trained using the same type of data. Networks pre-trained using natural images are usually inferior to upstream networks pre-trained directly using medical resources, even after fine-tuning. The visual and semantic differences between natural and medical images may be why [3].

Generative learning in SSL

Autoregressive Autoregressive (AR) models are usually Bayesian network structures in SSL. PixelRNN [76] and PixelCNN [77] separately use RNN and CNN as backbone constructs. Their general idea is to model the discrete probability distribution of an image and encode the dependencies of different pixels in the image. The neural network pre-

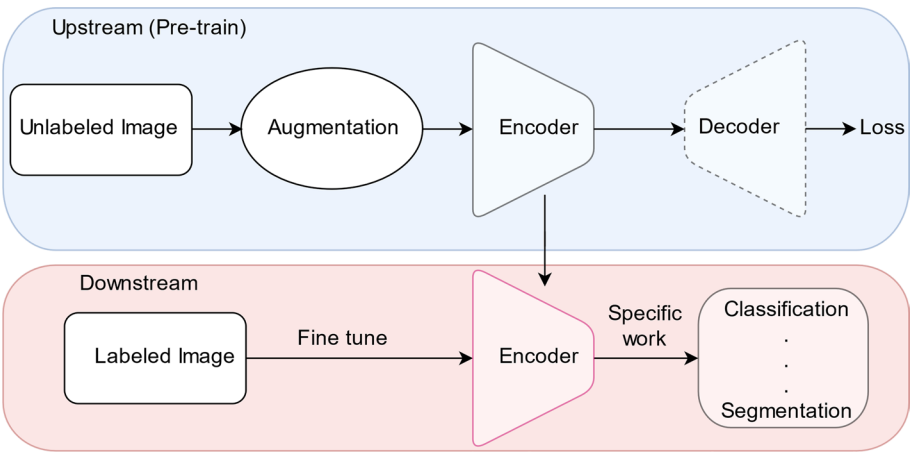

Fig. 4 The typical workflow of self-supervised learning

dicts the conditional distribution by scanning each pixel and then generates the next pixel sequentially. For example, PixelRNN generates the next pixel using all previous pixels and the specific unidirectional decomposition probabilities in the 2D image, its pretext task compares the predicted pixel and the original pixel. The subsequent PixelCNN relies on multi-channel convolution to acquire pixel sequences of different lengths of the image at once to learn the probability distribution, which saves a lot of computation (The training phase leverages parallel processing for efficiency, while the inference phase remains a sequential pixel-wise operation). This mechanism relies on masked convolutional kernels to mask or blur parts of the image, where each kernel contains lots of masked channels with different sizes. Figure 5 represents the main implementation idea of the two studies.

Flow-based model The flow-based model is capable of accurate likelihood estimation and inference, generally estimates the high-dimensional density $p(x)$ of the input x and then infers the z that can describe x through $p(x)$. The model is trained by reducing the loss between z and x, i.e., using an abstract representation of itself to reconstruct itself as a pretexting task. RealNVP [78] devised novel affine coupling layers to perform the estimation. Further, Glow [79] introduced reversible 1×1 convolution to simplify RealNVP's work and achieved good results. The loss function of this type can be represented by Formula (1).

$$\text{Loss} = \min\left(x_i - p\left(f(x_i)\right)\right) \tag{1}$$

Autoencoding model Autoencoder (AE) models typically consist of two components: an encoder and a decoder. The encoder maps the input data into latent space for sampling, which is an order of magnitude faster than sampling in pixel space. The decoder reconstructs the input from the latent space and collaboratively updates the encoder and decoder by comparing the output to the original input. Only the encoder will be used for downstream tasks once the collaborative update has stabilized. Comparison of the

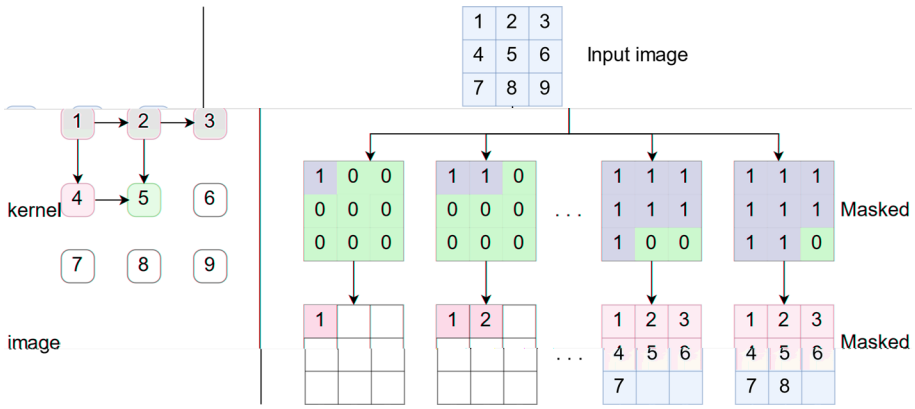

Fig. 5 Left: PixelRNN utilizes all previous pixels to predict the current pixel. Each forward propagation is computed from the top left to the bottom right, and the predicted value of the current pixel is derived from the likelihood product of all previously predicted values; Right: PixelCNN uses masked convolutional kernels to mask the image in parallel. Each convolution (mask) can obtain all the required pixel regions and predict the target pixel accordingly in time-saving

generated image with the original image is the pretext task. It may become the most popular generative model due to its flexibility. For example, VQ-VAE-2 [80] uses a multi-scale hierarchical organization mechanism that maps local and global information to the latent space through different hierarchical encoders and then converts these continuous features into discrete ones. The decoder reconstructs the discrete features to images and compares them to the original to push training forward. Figure 6 shows the workflow of its multilevel encoder/decoder. Typically, this type of model is deployed in three ways, such as denoising the image (input the blurred image during training and forcing the model to restore the original), classifying by looking for similarities through Euclidean distance in latent space, and leveraging the trained encoder for downstream tasks, including but not limited to classification and segmentation.

Hybrid generative model Hybrid generative models are structures that combine AE and flow-based models [81] or AR and AE models [82, 83], typically used in the natural language domain and graph generation. Such models generally can combine the advantages of different model types, but complex steps leading to more processing time are hard to avoid.

The different typical types of generative learning mentioned in this section are listed in Table 1.

Contrastive learning in SSL

Contrast learning in SSL has evolved through four phases: Exploration stage, Merging stage, Maturity stage, and Extension stage. A representative working relationship of these is schematically shown in Fig. 7. Contrast learning using SSL has outperformed supervised learning already on ImageNet [84, 85]. One important reason is that the network has mastered higher-level features in the learning process of multi-view images [86], thereby recognizing images in higher dimensions.

Exploration stage—no fixed training paradigm In this period, methods, models, objective functions, and pretext tasks are not unified, and all the work is in the exploration. The first representative work in this period is InstDise [87], where the authors consider each

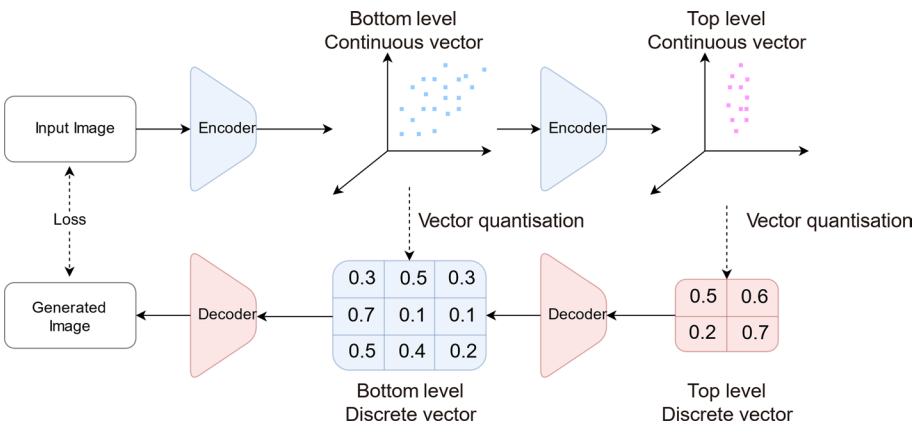

Fig. 6 The typical workflow of self-supervised learning

Table 1 Summary of typical types in generative learning

Type	Mechanism	Pros and Cons
Autoregressive	Predicts the current image by all previous pixels	Numerous amounts of data are not required as the model can predict, but the data needs to be autocorrelated, and the model cannot easily utilize the context simultaneously
Flow-based	The data itself is estimated using the likelihood of the abstracted representation of the data (high-dimensional features). Half of the features are returned to the previous step after each abstraction to continue participating in the abstraction	The trained encoder can be used directly in reverse to become a decoder, but it also limits the performance of the whole model
Autoencoding	Encoding the image as a high-level feature (shrinking) and then recovering it forces the encoder–decoder to learn the representation of the image	Good generalization ability and can also utilize context at the same time. However, it is sensitive to anomalous data and does not recognize the labels of the data itself

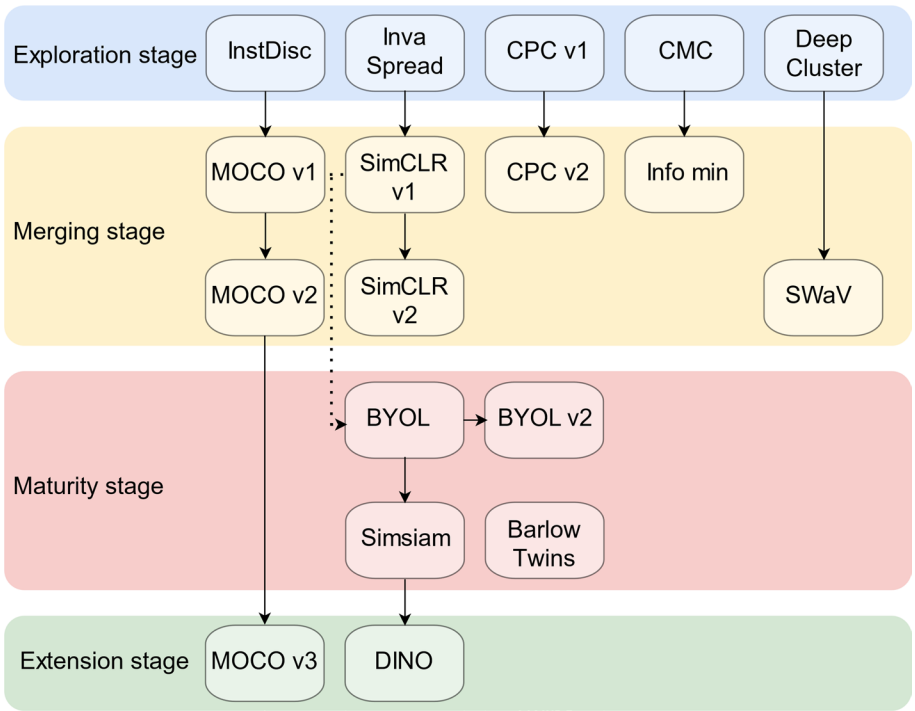

Fig. 7 The representative works of contrastive learning in each stage

image as one class and hope that the model can learn the image representation to distinguish all kinds of classes. Each image with different views is positive, and other images are negative samples. A memory bank stores all negative samples and updates them every epoch. The authors set the representation vector of each image to 128 dimensions to control the size of the memory bank since there are 1.28 million images in the ImageNet dataset. Considering the time consumption, only 4096 negative samples were randomly selected to participate in comparison in the forward propagation. The second representative work is InvaSpread [88], which abandons the memory bank and only utilizes the data in the mini-batch. Each instance in the mini-batch is augmented once into $2N$ positive

samples, the rest of the *2(N−1)* as negative samples. Similarly, CMC [89] augments an image four times and uses it as its positive sample, and the rest *4(N−1)* samples are negative samples.

Merging stage—two studies determine the future direction The most significant representations of the second stage are MOCO v1 [75] and SimCLR v1 [90]. MOCO v1 summarized contrastive learning as a dictionary query problem, a memory queue, and momentum encoder technology have been added to the workflow to form a giant and constant dictionary for negative sample saving. The memory queue dynamically updates the representation of negative samples in each small batch training, which ensures low time consumption and stable update representation. SimCLR v1 augments samples twice to form N positive and *2(N−1)* negative samples and adds a projector layer (i.e., MLP). Its structure only added a projector layer compared with nvaSpread [88] but achieved a better performance.

Maturity stage—training only using positive samples The third stage of comparative learning discards negative samples completely, avoiding the overhead of storing and updating a negative library. The purpose of using negative samples in previous stages was to prevent the models from falling into shortcut solutions. It is envisioned that without constraints, all models use the same output for any input so that models will fall into a situation where the loss is 0 but meaningless, becoming a model collapse. However, starting from BYOL [91], a representative work of this phase, the model was trained without negative samples and can achieve satisfactory results. Simsiam [92] streamlined many modules in the past work to verify the root cause affecting the model performance. It did not need negative samples, a large batch size, and a momentum encoder but instead ended up with fantastic modeling results. In conclusion, the third phase did not make structural changes but achieved good results and avoided negative samples as an influencing factor. The BYOL and Simsiam architectures are shown in Fig. 8 [92].

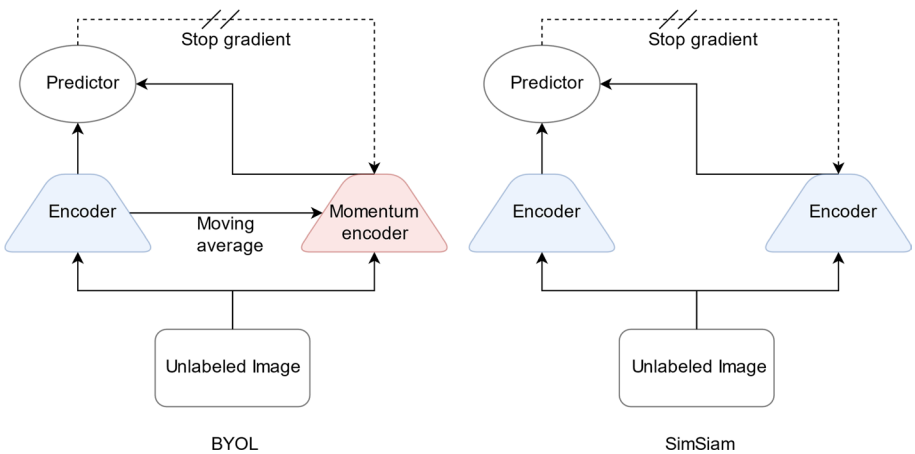

Fig. 8 Architecture bet1ween BYOL and Simsiam

Extension stage—pursue higher performance with vision transformer Since the emergence of ViT, many contrastive learning tasks have begun to use ViT as a backbone due to its powerful ability. MOCO v3 [93] and DINO v1 [85] are the representations in this stage. The structure of MOCO v3 is similar to Simsiam, except its CNN is replaced by ViT. In addition to ViT, DINO v1 uses a centering normalization operation in the teacher encoder to avoid model collapse. That is, average the data in the batch and then subtract this average from all instances in the batch.

The typical works of generative and contrastive approaches are listed in Table 2.

Self-supervised learning in medical image application

Most of the representations of SSL development have arisen in the natural image field, and one reason may be that natural images are more intuitive, i.e., the visualization of a model can be easily judged without prior knowledge. In relative terms, medical images generated by medical devices contain different semantic information, contextual structures, and representation types, which require independent research and evaluation. In most cases, SSL works in medical images have a similar workflow with differences in pretext tasks, embedded modules, and encoder types. The rest of this section summarizes the achievements of SSL on different image types, and Table 3 summarizes some valuable works.

X-ray

SSL has had many successful studies in medical imaging. In the X-ray images field, Azizi et al. [3] demonstrated that in self-supervised frameworks such as SimCLR v1, first using ImageNet for unlabeled self-supervised pre-training, followed by routine medical image pre-training and fine-tuning can effectively improve performance. This conclusion is based on X-ray and dermatology images, providing an optimization approach for

Table 2 Summary of SSL approaches for different types or stages

Category	Type or stages	Refs.	Pretext Task	Dataset	Perf
Generative learning	Autoregressive	PixelRNN [76]	Self-pixel prediction	MNIST [94] CIFAR-10 [95] ImageNet [50]	80.75 (NLL) 3 (NLL) 3.86 (NLL)
		PixelCNN [77]		CIFAR-10 [95] ImageNet [50]	3.14 (NLL) 3.57 (NLL)
	Flow-based	RealNVP [78]	Self-pixel abstract-reconstruction	CIFAR-10 [95] ImageNet [50]	3.49 (NLL) 4.28 (NLL)
	Autoencoding	VQ-VAE-2 [80]	Local/global information mapping-reconstruction	FFHQ [96]	3.41 (NLL)
Contrastive learning	Exploration	InstDisc [87]	Positive and negative samples comparison	ImageNet [50]	54% (ACC)
	Merging	SimCLR v1 [90]			69.3% (ACC)
		MOCO v1 [75]			60.6% (ACC)
	Maturity	BYLO [91]	Different image view comparison		74.3% (ACC)
		Simsiam [92]			71.3% (ACC)
	Extension	MOCO v3 [93]	Different image view comparison (ViT)		77% (ACC)
		DINO v1 [85]			77% (ACC)

NLL Negative log-likelihood, the smaller the better, *ACC* Top 1 accuracy, the higher the better The research in contrastive learning is based on ResNet50 or ViT-S as the backbone

Table 3 Summary of representative SSL works that contain downstream tasks in the medical image

Refs.	Type	Pretext task	Dataset	Metrics	Perf
Taleb et al. [23]	G & C	Contrastive predictive Rotation prediction Jigsaw puzzles Relative patch location Exemplar networks	BraTS 2018 (MRI) [115]	DSC	85.27 86.04 83.79 84.86 84.71
Haghighi et al. [40]		Discrimination Image restores	11 datasets include ChestX-ray14 [106] CheXpert [97] And Montgomery (X-ray) [142]	AUC [106] IOU [142]	81.12 ± 0.17 87.59 ± 0.28 98.24 ± 0.09
Cox et al. [117]	C	Masked volume in-painting Random rotation Contrastive coding	T1w and T2-FLAIR from the UK Biobanks (MRI) [162] BraTS 2021 (MRI) [115, 128, 129] ATLAS v2 (MR) [163]	DSC	For pre-train 0.9115 0.712
Qi et al. [116]		Reconstructing original images based on partial observations	BraTS 2021 (MRI) [115, 128, 129] Amos 2022 (MRI) [164]		86.75 ± 0.75 57.77 ± 13.93
Zeng et al. [109]		Generative adversarial	FS-CAD (X-ray) XCAD (X-ray)		0.828 0.755
Yu et al. [114]		Positive and negative pairs Neighboring crops contrast	COPDGene (CT) [151] MosMed (CT) [152]	ACC	86.3 ± 0.7 65.4 ± 2.5

Table 3 (continued)

Refs.	Type	Pretext task	Dataset	Metrics	Perf
Zhu et al. [38]	C	Cube disorder Cube spin Cube mask	BraTS 2018 (MRI) [115]	DSC	81.70
Spitzer et al. [29]		3D distance	BigBrain (His.) [123]		0.80
Lin et al. [35]		Class comparison Coloration	MoNuSeg (His.) [124] CPM (His.) [125]		74.41 73.73
Taleb et al. [41]	G	Multimodal data generation Puzzle solving	BraTS 2018 (MRI) [115] Prostate (MRI) [126] CHAOS (MRI/CT) [127]		81.89 (10%) 80.69 (10%) 93.85 (10%)
Chen et al. [24]		Context restoration	BraTS 2017 (MRI) [115, 128, 129]		84.27 25%)
Tajbakhsh et al. [11]		Rotation Reconstruction	LIDC-IDRI (CT) [130]	DSC ROC	0.909 (10%) 0.645 (10%)
Wu et al. [14]		Pixel disorder Nonlinear transformation Patches exchange Internal cropping External cropping		AUC	97.17 ± 0.32
Zhao et al. [42]		Pixel shuffling	OCT (CT) & Chest X-ray [131]		0.9642 0.8265
Wang et al. [21]		Multi-input correspondence Geometric transformation	PROSTATEx (MRI) [132]		0.753 (10%)
Zhou et al. [12]		Nonlinear local exchange, external cropping Internal cropping	LUNA2016 (CT) [133] ChestX-ray 14 [106]		98.34 ± 0.44
Haghighi et al. [47]		Self-discover Self-classification Self-restoration			98.47 ± 0.22
Tao et al. [45]		Rubik's cube restoration	NIH Pancreas computed tomography (CT) [134] MRBrainS18 (MRI) [135]	DSC	73.3 (10%) 77.56
Chaitanya et al. [25]	C	Positive and negative pairs	ACDC (MRI) [136] Prostate (MRI) [137] MMWHS (MRI) [138]		0.912 0.697 0.787
Srinidhi et al. [28]		Spatial correspondence of resolution predicting	BreastPathQ (His.) [139] Camelyon 16 (His.) [140] Kather multiclass (His.) [141]	ICC AUC ACC	0.701 (10%) 0.836 (10%) 0.976 (10%)

Table 3 (continued)

Refs.	Type	Pretext task	Dataset	Metrics	Perf
Nguyen et al. (2020) [10]	G	Confirm replacement Replacement sources	StructSeg2019 (CT) [143]	DSC	91.02 (50%)
Fischer et al. [27]	C	Multiple views comparison	ACDC (MRI) [136] MMWHS (MRI) [144] Prostate (MRI) [137]		89.4 ± 03.6 (6) 80.4 ± 04.2 (6) 76.3 ± 05.6 (6)
Gazda et al. [4]		Positive and negative pairs	Cell (X-ray) [145] ChestX-ray 14 [106] C19-Cohen (X-ray) [146] COVIDGR (X-ray) [147]	AUC	96.9 78.1 91.5 86.0
Zhou et al. [46]			ChestX-ray 14 [106] LUNA2016 (CT) [133]		76.2 (10%) 84.4 (10%)
Li et al. [36]			Camelyon 16 (His.) [140] TCGA (His.) [148, 149]		0.9165 0.9815
Dong et al. [9]			DBT (X-ray) [150] Cell (X-ray) [145]		0.7848 0.8831
Yan et al. [100]		Multiple views comparison	Chest X-ray 15 K [157] RSNA2018 [158, 159] COVID-19 Radiography Database[160]		0.96 (10%) 0.92 (10%) 0.925 (10%)
Sheng et al. [101]			VinDr-CXR [161] ChestX-ray 14 [106]	mAP50 AUC	0.2502 0.7756 (10%)
Liu et al. [17]		Rotated degree prediction	LUNA2016 (CT) [133]	FROC	0.906 (1/8)
Sun et al. [16]		Patch contrast Graph contrast	COPDGene (CT) [151] MosMed (CT) [152] COVID-19 CT [153–156]	ACC	80.8 ± 0.17 65.3 96.3

G and C represent generative and contrastive learning, separate. The table only contains classification, segmentation, and false-positive reduction task studies. If works use similar frameworks and the same dataset, only record the ones with better performance. The metrics they use include Dice Similarity Coefficient (DSC), Receiver Operating Characteristic Curves (ROC), Area Under Curve (AUC), Intraclass Correlation Coefficient (ICC), Accuracy (ACC), Intersection Over Union (IOU), and Free-response Receiver Operating Characteristic Curves (FROC). About parentheses contents after the number in the Perf., the percentage indicates the proportion of fine-tuning data used, the number between brackets represents the amount of fine-tuning annotations (if any), and the blank means that the article does not mention it

situations where the number of medical images is insufficient. Azizi et al. [5] and Li et al. [6] conducted the same type of research and reached the same conclusion.

Gazda et al. [4] tested the SSL transferability of X-ray images using a framework similar to SimCLR v1. The study used CheXpert [97] as the training set and tested it on four other X-ray datasets. The results indicate that even with the same image type, images from different sources can lead to unstable performance of the model. Furthermore, Xing et al. [7] tested the masking rate of the Masked AutoEncoders (MAE) framework [98] on X-ray images and finally obtained the best rate of 0.4. Reversely, Imagawa et al. [99] found that fine-tuning with 60% labels on the SimCLR framework can achieve performance equivalent to supervised learning, and better performance can be achieved by retaining a higher proportion.

Tiu et al. [8] used images and corresponding clinical reports to train the model that the performance obtained can surpass supervised learning. However, clinical reporting means images have been reviewed and diagnosed, which requires as much human labor as labeling.

Dong et al. [9] used a sliding window mechanism to feed high-resolution images. Each image will be divided into many patches, and two augmented views of one patch become a positive sample pair, while the adjacent patches are negative samples. This framework

adopts the pretext task of positive and negative sample pairs and the structure of an encoder and a projector, where only the encoder participates in downstream tasks.

Yan et al. [100] trained the encoder using the MOCO v1 paradigm, forcing the model to learn losses from the output features of both encoders. This study mainly focuses on protecting the privacy of fine-tuned datasets. It factorizes each residual layer of the backbone into two low-rank layers called LoRA, and after freezing the pre-training weights, adds Gaussian noise for parallel training. The essence of this study is to add Gaussian blur to the original image for confidentiality and freeze pre-training parameters to reduce the impact of noise on performance. However, increasing noise would hurt model training, and the idea of encrypting data needs to have a smaller impact on performance.

Sheng et al. [101] adopted a two-step training strategy: Barlow Twins [102] pre-training and Faster R-CNN [103] fine-tuning (ResNet50). Two different datasets were used in two steps to apply the model for abnormal area localization in lung X-ray images. This study provides new ideas for different framework combinations, without significant changes to the original structure.

Zhang et al. [103] used the MAE framework to validate the transferability of adults to young children on four datasets: MIMIC-CXR [104], CheXpert [97], COVIDx [105], ChestX-ray 14 [106], and PediCXR [107, 108]. After pre-training on adult X-rays and fine-tuning on young data, its results are positive. Although this study does not have an innovation framework or model, it once again confirms the usefulness of a broader pre-training dataset.

Zeng et al. [109] transformed photos of the same region in different states to perform generative adversarial tasks and finally fine-tuned the trained model for vascular segmentation. This approach can be used for any task that involves before and after states, besides vascular segmentation.

CT

CT can be seen as a 3D X-ray image and has been widely used in clinical practice. In the study of the CT field, Nguyen et al. [10] established an analytical framework for organ segmentation and intracranial hemorrhage detection using semantic and spatial features using a standard self-supervised upstream and downstream pattern. This framework utilizes the spatial information of 3D data and the semantic information in 2D slices to train pretext tasks. Specifically, a portion of a 3D data slice will be replaced by the same region in another slice. The two labels of the task are whether the specific slice has been replaced and which slice the replaced region comes from. The entire framework is concise and clear, and the only problem is that the lowest proportion of fine-tuning data used is 50%.

Tajbakhsh et al. [11] used rotation and reconstruction pretext tasks for the false-positive reduction in lung lobe segmentation and nodule detection, respectively. Two tasks were conducted independently with different network structures but the same dataset. The results of this study indicate that a self-supervised framework relying on a single simple task (single loss function) and CNN backbone may be insufficient performance for classification tasks.

Zhou et al. [12, 13] set up four schemes to deform images, i.e., non-linear, local exchange, outer cropping, and inner cropping. The outer cropping preserves the selected patches, while the inner deletes the selected patches. The input data set will be randomly transformed through three schemes, with 12 conversion paths in total. The transformed image will undergo loss calculation with the original image after passing through the encoder–decoder structure, and the encoder is used for downstream false-positive reduction classification of pulmonary nodules.

Wu et al. [14] and Huang et al. [15] combined image perturbation with self-supervision. The pretext task designed by it first divides the image into different crops, which undergo pixel rearrangement, monotonic Bessel curve nonlinear transformation [12], local patches exchange within the crop, and data generation through inner and outer cropping. The upstream network structure adopts an encoder–decoder structure, which is the difference between the restored and original image as a loss. Huang et al. [15] further designed a Domain Adaptation Block but achieved relatively lower performance.

Sun et al. [16] used the anatomical features of medical images to help establish pseudo-labels that similar regions in images can be represented as similar nodes (graphs). Each node and its enhanced image are positive sample pairs, while other nodes at the same position are negative sample pairs. All nodes in a graph can form a complete image representation, with different views of the image forming secondary positive sample pairs and different images forming secondary negative sample pairs. This study innovatively combines deep learning and graphs, belonging to the category of contrastive learning. Downstream tasks include future exercise detection, severity of lung tissue abnormalities, and classification of COVID-19 patients.

Liu et al. [17] directly rotated each 3D image three times (90°, 180°, 270°) and formed four positive samples with the original image to train the ability to predict rotation angles. The framework only uses one encoder, which is used for classification tasks after fine-tuning.

Shabani et al. [18] build a segmentation model through two steps. First, the pseudo-mask is generated over SSL, and then traditional supervised learning is performed using U-Net [110] for conventional supervised learning. In detail, they first use a vanilla encoder–decoder to turn unhealthy images into healthy images. The generated image is compared with the original image to obtain the changed area outline, i.e., the pseudo-mask. Although the performance is quite different from supervised learning, it provides a new direction for pseudo-label generation.

Tan et al. [19] combined MAE and self-distillation and found that the performance can be effectively improved where the best masking ratio of MAE is 0.4. Another unconventional study is by Yu et al. [20]. This study used normal images and artificial images containing metal artifacts to train the network, constructing a contrastive learning framework that can effectively remove metal artifacts in CT images.

Another innovative application of the MAE framework is Kumar et al. [111] using it to encrypt images. The specific method is to encrypt the masked image using TPM and transmit it to the user. In addition to possessing the key, the user must have a trained decoder to restore the original image from the damaged image.

Guo et al. [112] proposed a stepwise incremental pre-training strategy, which first trains a discriminative encoder through discriminative learning, and then connects

the pre-trained discriminative encoder to a recovery decoder to form a skip-connected encoder–decoder for further joint discriminative and recovery. And then, associating the pre-trained encoder–decoder with the adversarial encoder for complete discrimination, recovery, and adversarial learning. This strategy improved the AUC performance of frameworks, such as MoCo, BYOL, PCRL [46], and Swin UNETR [113], for different disease classifications on the LUNA16 dataset by 0.1%–10.6%, but the improvement effect for most classification tasks did not exceed 1%.

Yu et al. [114] improved MOCO v1 using the same image region of different patients as positive sample pairs and the remaining regions as negative pairs based on anatomical structure. At the same time, feature similarity was also calculated in the nearby regions of this region. The overall approach and pretext tasks are relatively primitive, so the performance is not outstanding.

MRI

Wang et al. [21] proposed an SSL framework called MI SelfL for prostate cancer classification in MRI images. The self-supervised framework in the article designs two pretext tasks, namely Multi-input correspondence and Geometric transformation. Specifically, this framework randomly replaces the local regions of some input images with those of the same image batch to generate pseudo-labels (modified images are abnormal images). Secondly, the framework randomly rotates and flips images. The image features that have been randomly replaced should have significant differences from the original image, while images that have only undergone geometric changes should have similar features. As a prerequisite, this framework needs to work in a patient with multiple different modalities or parameters images, and its performance still has room for further improvement.

Kalapos et al. [22] did not attempt to achieve better downstream task performance. They discovered a new pre-training mechanism that enables downstream tasks to converge faster/more stably and reduces the need for fine-tuning data. Concisely, this study conducted weight transfer on the online (teacher) and target (student) encoders of the BYOL framework, with the transferred weights coming from the supervised and self-supervised networks trained with ImageNet. Although this approach requires pre-training the initial weights of BYOL, it is meaningful when various publicly available ImageNet weights can be directly used.

Taleb et al. [23] designed five models and their respective pretext tasks (contrastive prediction, rotation prediction, jigsaw puzzle, relative position, and exemplar networks), respectively. Contrastive prediction first divides the 3D image into several patches (such as $3\times3\times3$), then uses pyramid-shaped data as input to predict a specific patch (e.g., if predicting the patches $P_{i,j,k+1}$ and $P_{i,j,k+2}$, the input data is $P_{i\pm1,j\pm1,k}$ and $P_{i\pm2,j\pm2,k-1}$). The second training task is to rotate the image by any angle and force the model to predict the angle. The puzzle task also requires dividing the image into several patches, shuffling the order, and having the model restore it. Relative position prediction is to give any patch to the model, and the model predicts its position throughout the entire image. Exemplar networks use a positive and negative sample pairing mechanism to train the model, where the feature Z_i of the original image and the augmented Z_{i+} form

a positive sample pair. In contrast, Z_i and the feature Z_j of other images form a negative sample pair.

Chen et al. [24] used image restoration as a pretext task and validated it on three downstream tasks (classification, localization, and segmentation). Detailed, the study divides an image into several regions and then randomly exchanges the positions of some regions. The original image is used as the label of the scrambled image for training. The self-supervised framework uses a single encoder structure and calculates the loss with $L2$ compared between the original and the restored image. This study belongs to the early stages of SSL, in which the concise structure results in suboptimal performance.

Chaitanya et al. [25] innovated two loss functions, namely global and local comparison loss functions. Then, the training process adopts the paradigm of contrastive learning, sampling N images as a batch, enhancing a single image once as a positive sample pair, and enhancing the remaining images once as $2N - 2$ negative sample pairs. The new loss function is a framework that can achieve 8% of baseline performance with only 4% of data.

Zhou et al. [26] developed a self-supervised multi-modal image (different perspectives to one region) synthesis framework to enhance MRI image quality. The framework combines autoencoding and the self-attention mechanism to implement SSL on 2D and 3D images. Region images in different perspectives undergo a destruction–restoration pretext task to train their region-dedicated feature extraction SSL model. The intermediate features of various models are superimposed on each other and sent to the generation network to obtain an enhanced image. This framework was tested on the BraTS 2018 [115] data set, and the peak signal-to-noise ratio (PSNR) was greatly improved compared to the baseline.

Fischer et al. [27] combined random walks with SSL and used multi-stage contrast to enhance training. This SSL framework uses an encoder–decoder structure (like U-Net) to parse different augmented views, trained by reducing the gap between views of the same original image. Uniquely, the decoder stage performs loss calculations in each step to enforce constraints and fine-tune with only 6 annotations.

Qi et al. [116] randomly extracted different crops from MRI images and attempted to restore the original images using a Siamese ViT network. The model loss function is the MSE value between the restored image and the original image, as well as the similarity coefficient between the outputs of the Siamese network. This is a novel pretext task, but the current performance is not stable.

Cox et al. [117] designed a novel two-stage pre-training method using the Swin Transformer for segmenting brain lesion locations. During model pre-training, only healthy images and Swin UNETR are used to train the model, and then abnormal are used for fine-tuning. This study achieved good performance, however, the pre-training of the model requires healthy images, and the tolerance for noisy data has not been tested. At the same time, the training strategy for fully healthy images also requires some annotation labor.

Histology

Srinidhi et al. [28] proposed a new pretext task to predict and sort resolution sequences in images, which essentially predicts the corresponding spatial relationships between

images of different resolutions that are enlarged or reduced. The model used the ResNet-18 network as the backbone and constructed a classic teacher-student structure. During the experiment, there was a significant difference in performance between adding or not adding unlabeled test sets to the training set.

Spitzer et al. [29] used SSL to segment 3D histological images of the human brain. They designed a Siamese network to encode two patches separately, with pseudo-labels representing their 3D distance of on the human brain. Compared to the baseline, the performance has improved by 8%.

Lu et al. [30] used self-supervised methods to extract abstract features of proteins in cells from cell microscopy images, which can be used for tasks such as determining the impact of drugs on cells. The pretext task of the framework utilizes different channels of cells, such as ordinary protein images and corresponding stained protein images, as data and label pairs. The data and label images are processed through an encoder, concatenation, and decoder to obtain protein features. The entire study does not have a conventional classification or segmentation downstream task. However, the feature extraction ability can save a lot of labor, providing a novel reference for the application direction of self-supervision.

The study by Ciga et al. [31] aims to verify the factors that affect the performance of SSL and experiment with their hypotheses on 57 tissue pathology datasets with SimCLR v1. In the end, they found that pre-training with data similar to downstream tasks is better than directly using the pre-training weights of ImageNet. Secondly, the smaller the dataset, the greater the pre-training effect. The other two findings are that the more pre-trained images, the larger the size, which is more beneficial for downstream tasks.

Veeling et al. [32] used image rotation as the pretext task and two encoders with the same network structure for feature extraction, with the distance between the two output features used as the loss. The experimental results indicate that overly simple networks and self-supervised structures make it difficult to achieve satisfactory performance, and a deep neural network with sufficient parameters and training constraints may be the performance guarantee.

Zhao et al. [33] built a framework that combines one online (teacher) and two target (student) encoders. Its process is similar to BYOL and aims to classify and diagnose endometrial cancer. However, the performance increased only slightly compared with BYOL.

Yang et al. [34] designed a two-step contrastive learning framework. First, the H and E channel views are entered into the encoder–decoder structure for contrastive learning. Following this, the encoder continues to perform contrastive learning by different augmented images and then combines the decoder (frozen parameters) to perform MSE loss calculation with the original image. The two encoders of contrastive learning share parameters and participate in specific downstream tasks. The work of Lin et al. [35] also separated H and E channel images, but they only used the H channel for training. The two images (H channel) generate pseudo-labels through clustering and then send them to the segmentation net (encoder). The generated features are compared with each other (task 1) and compared with the original image after coloring (task 2).

Li et al. [36] used the characteristics of whole slide images to design a classification framework based on pyramid shape features. This framework first uses contrastive

learning (the image and its augmented view form a positive sample pair and form negative pairs with other images) to train the model by different enlarged images (such as 5x, 20x) and then concatenate the outputs of multiple encoders in order along the pyramid shape. These features represent the characteristics of images and are used to distinguish image class. Finally, the framework outperforms supervised learning and mitigates memory requirements.

Despotovic et al. [118] tested self-supervised learning frameworks, such as SimCLR, BYOL, and MOCO v3, in glioma subtype classification tasks, and the final experimental results showed that BYOL performed the best in this task. Another similar study was carried out on breast cancer images by Ye et al. [119]. They found that the original iBOT model performed best.

Ultrasound

Ultrasound images are formed by collecting the echoes generated by ultrasound passing through the human body. About ultrasound images, Liu et al. [37] developed an enhancement framework for images. They used CNN to multiscale low-resolution ultrasound images into high and super-resolution and then converted these images into low-resolution. The process of enlarging/reducing the original image is completed by the Generative Adversarial Network (GAN), and the loss function is the adversarial and discriminator losses. Compared to conventional GANs, the same images with different resolutions form a self-supervised training pair that enhances the training, ultimately resulting in the enhanced original resolution image.

Hybrid

Hybrid works have validated its framework on multiple image types, demonstrating hyper-universality. Zhu et al. [38] proposed a contrastive learning framework using a combination of three sub-tasks (three loss functions): cube sorting, orientation, and masking recognition, as a 3D pretext task. The framework performs two tasks: cerebral hemorrhage classification and brain tumor segmentation in CT (private) and MRI brain images. In detail, this framework divides 3D images into several cubes of the same volume and then shuffles them in order, randomly rotates, and masks the partial volume of the cubes. Then, two identical and synchronously updated CNN Siamese networks are used to train the model. The three sub-tasks mainly rely on the translation and rotation invariance of 3D images for work, and the entire framework has a certain tolerance for noisy data. Zhuang et al. [39] adopted the same network structure, but only cube sorting and rotation were used as pretext tasks. Due to the fewer loss functions for constraint training, their model performance is relatively lower.

Taleb et al. [41] innovatively set up a multimodal puzzle task for images. It divides the image into several crops and then exchanges some crops between different modalities to generate a ground truth. These exchanged ground truth images will be sent to the model for training after shuffling the cropping order, and the loss function is the mean squared error between the model output and the unordered ground truth. The multimodality of images was generated using cycle GAN [78], which solved the bottleneck of using the model on datasets with insufficient modalities. Finally, this framework was applied to CT and MRI datasets for segmentation and achieved good results.

Haghighi et al. [40] combined discriminative, restorative, and adversarial learning tasks to calculate losses and train the model, applying them to data types, such as CT, MRI, and X-ray. First, the model uses a separate discriminator to determine the differences between images through hyper-features which is the first training route. The second route establishes an encoder–decoder structure, which updates the model by generating adversarial paradigms and the distance loss between the generated and the original image. This framework can be combined with other SSL methods, such as MoCo v2 [120], SinSiam [92], etc., to improve the performance of downstream tasks.

Zhao et al. [42] proposed an anomaly detection SSL framework called SALAD for diagnostic analysis of CT and X-ray images. The entire pretext task relies on a mechanism called random pixel shuffling, which swaps some pixels in the image to form pseudo-labeled data with the original image. The encoder processes the pseudo-labeled data, and then the decoder restores the encoded features. The loss function is the output comparison consisting of the encoder and the output with the original image. For the dataset, the training of this model needs to exclude all abnormal and only use healthy ones.

An exploratory study on CT, X-ray, and MRI by Haghighi et al. [43] combined multiple SSL paradigms, including Discriminative, Restorative, and Adversarial Learning. This study used three routes for combined training, with the total loss function being the weighted sum of the three paths. The discriminative route separately uses two encoders f_θ and f_γ to encode two inputs X_1 and X_2, and compare the distance of the encoded features. The two inputs may be different images or unlike views of the same image. The restorative route compares image X_1 with reconstructed image X_1', which is processed by encoder f_θ and decoder g_θ and then calculates the loss. Finally, the adversarial route added a discriminator after encoder f_θ and decoder g_θ to utilize X_1 and X_1 for adversarial evolution. This scheme works successfully on frameworks, such as MoCov2 and SimSiam, but has performance improvements of 0.08% -13.6% under different frameworks and fine-tuning data ratios.

Ghesu et al. [44] used four types of medical images: X-ray, CT, MRI, and Ultrasound, totaling over 100 million images, to pre-train the model. The self-supervised framework adopts a Siamese encoder setting, and the difference in features after encoding is the training loss. Where input images will be scaled and Energy-based augmented [121]. The experimental task is divided into three parts: abnormal chest X-ray detection, MRI detection of brain tumor metastasis, and CT detection of cerebral hemorrhage. The research results indicate that pre-training mechanisms can effectively improve accuracy, robustness, and model convergence speed. Like this study, Anand et al. [122] prove the same effectiveness of pre-training in SSL based on cardiac ultrasound images.

Haghighi et al. [47] constructed a framework including self-discovery, self-classification, and self-recovery tasks. In the three tasks, self-discovery utilizes the anatomical features of medical images to divide them into different patches, followed by self-classification to classify these patches according to their types. For example, they could be used to determine whether a patch belongs to the left or right lung in an X-ray. The final self-recovery is a typical self-supervised method, which borrowed the method of Zhou et al. [13] to establish a pipeline and validate the idea on X-ray, CT, and MRI images.

Tao et al. [45] explored the performance of self-supervised frameworks in segmentation tasks on CT and MRI images. They divide the 3D image into volumes like a Rubik's cube so the image can rotate in three directions: sagittal, coronal, and axial. Randomly rotate the Rubik's Cube and restore to the original image as the pretext task. The framework adopts a GAN structure and calculates losses in both discriminator and image restoration directions.

Zhou et al. [46] added two encoders to improve the performance of X-ray and CT contrastive learning, forming a structure of three encoders and one decoder. One of the two newly added encoders learns image representations as a regular encoder, while the other hybrid encoder mixes the outputs of the regular encoder and momentum encoder. The image is processed through one of the augmentation methods in the random crop, random flip, random rotation, painting, outputting, and Gaussian blur before entering the regular and momentum encoders. The outputs of the three encoders are then compared to the original image, resulting in mean squared error loss (the encoder has a U-Net-like structure). The framework also stores images in the memory bank, where the same image and its enhanced images are positive samples and form negative sample pairs with other images. The update mechanism is similar to InstDisc.

Qayyum et al. [48] built a new transformer-based SSL framework on MRI and CT images, using different image views for contrastive learning. The innovation is that the encoder uses a transformer, and the decoder is a CNN, and then the cascade becomes a U-shaped structure.

Discussion

Medical images are different from natural images in that they are only composed of a single-color channel and have specific position/intensity/size features. These features are crucial components of medical images and may serve as potential indicators of health conditions. Owing to these unique features, the exploration of medical images necessitates a specialized approach that diverges from the methodologies applied to natural images. This review encapsulates the recent scholarly work on self-supervision within the medical imaging sphere, encompassing X-ray, CT, MRI, Histology, and Ultrasound. The previous SSL works have shown a clear division of stages, i.e., the nascent and the flourishing stages. The nascent stage is usually a single pipeline, wherein researchers harness unlabeled data for training after designing an encoder/decoder. This encoder/decoder is then directly deployed for specific diagnostic tasks. A flourishing stage study typically involves upstream and downstream work, with upstream using unlabeled data to pre-train models and downstream using labeled data to fine-tune the pre-trained network (usually the encoder in the upstream). Post-fine-tuning, the network is augmented with additional layers to execute the targeted medical imaging tasks.

The performance of SSL frameworks has clearly shown characteristics of catching up with or exceeding the supervised learning frameworks in recent years. Especially at the current stage, researchers using combination loss functions and multi-pipeline structures will significantly improve the performance of SSL frameworks, indicating that SSL research has emerged as a mature paradigm and crossed the initial stage. The pretext task is the most critical and influential part of the SSL framework. It determines the basic structure of the framework, helping the model train and learn image features

without labels. Pretext tasks in medical images can usually be divided into three types: Destruction Recovery, Sample Pair Comparison, and Information Prediction (such as relative position prediction), among which Destruction Recovery and Sample Comparison are the most commonly used types. Destruction Recovery and Information Prediction focus more on global features. In contrast, Sample Pair Comparison focuses more on local features (many studies combine destruction recovery and sample pair comparison to focus on both features). Using different pretext task strategies can make the model generative or contrastive. For example, Destruction Recovery is a generative task, while combining Destruction Recovery and Sample Pair Comparison would be a contrastive model. For the model explanation and implementation, the complexity of Information Prediction, Sample Pair Comparison, and Destruction Recovery increases in sequence, which usually leads to an increase in the structural complexity of the model. Information Prediction usually requires images in the dataset to have the same anatomical structure, sampling specifications, or other almost invariant information, while the other two types do not. Sample Pair Comparison has unique advantages in multimodal training, as it can compare the outputs of different modal encoders. However, this task usually means training multiple networks simultaneously, and more computational overhead cannot be avoided. Some studies have pointed out that medical image tasks focus on extracting fine-grained features [40], so Destruction Restoration is more suitable. The fact that the most commonly used task type in existing research is the Destruction Recovery task indirectly proves this point, especially with the good performance of ViT, Destruction Recovery tasks have also shown more target tolerance and excellent performance results, but designing the areas or mechanisms of Destruction Recovery tasks requires more time. The correct choice of the pretext task is often a guarantee for the model's performance. However, the selection process of the pretext task is inspiring and difficult, and there are almost no paradigms for all goals. Choosing the appropriate task based on different tasks is still worth exploring.

Due to the differences between medical and natural images, not all pretext tasks in natural images can be directly used. Usually, tasks, such as destruction recovery, patch order recovery, rotation, scaling, etc., can be directly utilized. However, some pretext tasks that rely on anatomical knowledge are unique to medical images, such as relative position prediction, distance prediction, etc. According to the included research, the performance of using a single task is generally worse than using multiple tasks, indicating that combining multiple tasks can enable the model to learn more accurate representations. The multiple loss functions can also constrain the model to make it more robust.

In some cases, the transferability of SSL remains a challenge, as test sets from the same distribution as the training data mostly outperform those from different sources, highlighting the need for improved generalization in SSL models. Models that perform well on one type of image often have unstable performance on another, with poor surface model universality. In addition, existing research still faces issues such as excessively high fine-tuning data or performance not significantly improving with an increase in the proportion of fine-tuning data. Most existing classification studies have not reported the false-positive rate, an important indicator. Due to unlabeled training of SSL, it cannot focus on the target facts and may learn too much information, resulting in excessive false-positive reports. Nevertheless, SSL has demonstrated significant benefits in

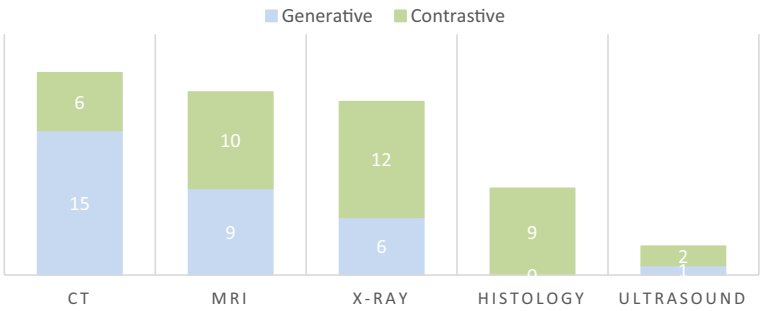

Fig. 9 Study amount statistics

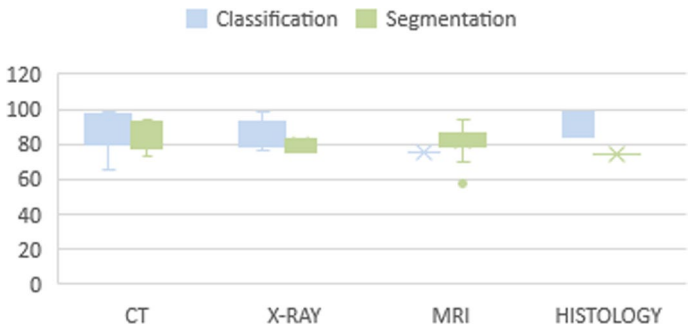

Fig. 10 Study performance statistics

medical imaging, effectively mitigating the challenges associated with limited labeled data and reducing the impact of labeling biases.

Figure 9 illustrates that SSL-related studies exist in various medical image types, highlighting a notable focus on CT imaging studies. MRI, X-ray, and Histology are also well-represented, whereas ultrasound imaging has garnered the least attention. In these SSL studies, except for the research on CT images that mostly use generative learning, contrastive learning in other image types is equally prevalent or significantly more than generative learning. The number of public datasets and the difficulty of diagnosis may be the reasons for different studied quantities. Moreover, the collected research also covers a comprehensive range of organs, including lungs, liver, kidneys, brain, heart, and prostate. Regarding the specific clinical challenges (downstream tasks), the previous works involve classification, localization, segmentation, false-positive reduction, and image enhancement. Among them, classification and segmentation account for the majority. Figure 10 shows segmentation and classification performance statistics. The performance metrics are the DSC and the AUC; however, the datasets utilized across various studies are not uniform. Therefore, the comparative figure is a general guide rather than a precise benchmark for the specific image type. In Fig. 10, MRI classification and Histology segmentation tasks are worthy of future exploration due to the small study size, and low or concentrated performance. Overall, after upstream pre-training, the current SSL framework can schedule any challenging downstream task with satisfactory results.

This review paper collects most of the research on medical images but does not include studies on Skin Lesions, Retina, Endoscopy, and Optical Coherence Tomography

images. In addition, this review cannot provide recommendations for optimal pretext tasks and their combination strategies. The impact of different tasks and self-supervised approaches on performance also cannot be answered.

In future, people could incorporate ablation experiments as much as possible. For example, comparing the studies of Zhu et al. [38] and Zhuang et al. [39] showed that frameworks with almost identical structures but using more modules achieved lower performance. In addition, future research should incorporate performance comparisons with supervised learning as much as possible to help readers understand intuitively. The computing overhead of the entire framework also needs to be provided to facilitate evaluation. This is because the deep network has more and more parameters and more complex structures. It is necessary to intuitively evaluate whether the network scale suits edge computing and other fields. Whether using different types of medical data before pre-training can achieve the same performance improvement as using natural images, researchers could test and find its effective mechanisms. For the pretext task, existing medical image-related research mostly uses mature frameworks or pretext tasks from natural images, and pretext tasks that utilize unique anatomical structures (such as predicting relative positions of regions) have not shown significant advantages. Therefore, it has been found that pretext tasks that are more suitable for medical images can also be further explored in future. Reducing the proportion of fine-tuning data, using synthetic data for training, multimodal training, and specific distillation of large models is also worth exploring in future. The proportion of labeled data in the pre-trained data in downstream tasks is better not exceeding 10% or even 1%, which will make the model more practical. Otherwise, it should be compared with the baseline without pre-training to verify the value of the SSL strategy. Using generated data for training can increase the diversity of training data and decrease the difficulty of collecting data, thereby increasing the model universality. However, bridging the domain gap between synthetic and real data is crucial. On the other hand, multimodal training can improve model accuracy and reduce false positives and misdiagnosis rates, which is also worth exploring. Finally, using large models for specific distillation can reduce pre-training time and obtain more feature parsing ability than conventional pre-training.

Conclusions

Machine learning diagnostic analysis methods for medical images past required extensive labeled data, which was often a challenge to satisfy in the medical field. SSL has emerged as a transformative solution, significantly diminishing the reliance on labeled data and achieving performance comparable to supervised learning with as little as 10% of the total labels. This paper aims to equip researchers with the foundational knowledge of SSL in medical imaging by reviewing the latest research in the field. The review commences with an in-depth exploration of contrast learning and generative learning within SSL, followed by a categorized presentation of the image types already being extensively studied under the SSL framework, e.g., X-ray, CT, MRI, histology, and ultrasound. While SSL in natural images has seen a multitude of pretext tasks, the unique anatomical structures of medical images may necessitate the development of proprietary pretext tasks. This paper summarizes the pretext tasks utilized in medical images, encompassing damage recovery, patching disorder, rotation, scaling, relative position prediction, coloring comparison, path distance

prediction, and Rubik's cube recovery. The downstream tasks covered in the review include classification, localization, segmentation, false-positive reduction, and image enhancement. The analysis reveals certain patterns in impacting performance: complex structures and increased task/loss constraints tend to be advantageous. However, it is imperative to consider model size and training costs. The ablation studies to evaluate the contribution of each module are also beneficial. In conclusion, current SSL research has demonstrated the potential to match or even surpass the performance of supervised learning. Future directions include reducing the reliance on fine-tuning data, constructing more generalizable models, and leveraging clinical or insurance data for multimodal learning initiatives.

Abbreviations

ACC	Accuracy
AE	Auto-encoding
AR	Auto-regressive
AUC	Area under curve
BP	Back propagation
CAD	Computer-aided diagnosis
CNN	Convolutional Neural Network
CT	Computer tomography
CV	Computer vision
DSC	Dice similarity coefficient
FROC	Free-response receiver operating characteristic curves
GAN	Generative adversarial network
ICC	Intraclass correlation coefficient
IOU	Intersection over union
MAE Masked	AutoEncoders
mAP50	Mean average precision at an IOUof 50%
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
NIN	NetWork In NetWork
NLL	Negative log-likelihood
PSNR	Peak signal-to-noise ratio
ROC	Receiver operating characteristic curves
SSL	Self-supervised learning
VGG	Visual geometry group
ViT	Vision transformer

Author contributions

ZENG wrote the main manuscript text, with ABDULLAH and SUMARI serving as supervisor and project support, respectively.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable (this manuscript does not involve human/animal studies).

Competing interests

The authors declare no competing interests.

Received: 10 May 2024 Accepted: 17 October 2024
Published online: 27 October 2024

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing system. Curran Associates, Inc.; 2012. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html. Accessed 8 Mar 2024.

2. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale; 2020. https://openreview.net/forum?id=YicbFdNTTy. Accessed 25 Nov 2023.
3. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T, Natarajan V, Norouzi M. Big self-supervised models advance medical image classification; 2021. p. 3478–88. https://openaccess.thecvf.com/content/ICCV2021/html/Azizi_Big_Self-Supervised_Models_Advance_Medical_Image_Classification_ICCV_2021_paper.html. Accessed 13 Mar 2024.
4. Gazda M, Plavka J, Gazda J, Drotar P. Self-supervised deep convolutional neural network for chest X-ray classification. IEEE Access. 2021;9:151972–82. https://doi.org/10.1109/ACCESS.2021.3125324.
5. Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, Chen T, Tomasev N, Mitrović J, Strachan P, Mahdavi SS, Wulczyn E, Babenko B, Walker M, Loh A, Chen P-HC, Liu Y, Bavishi P, McKinney SM, Winkens J, Roy AG, Beaver Z, Ryan F, Krogue J, Etemadi M, Telang U, Liu Y, Peng L, Corrado GS, Webster DR, Fleet D, Hinton G, Houlsby N, Karthikesalingam A, Norouzi M, Natarajan V. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat Biomed Eng. 2023;7:756–79. https://doi.org/10.1038/s41551-023-01049-7.
6. Li G, Togo R, Ogawa T, Haseyama M. COVID-19 detection based on self-supervised transfer learning using chest X-ray images. Int J Comput Assist Radiol Surg. 2023;18:715–22. https://doi.org/10.1007/s11548-022-02813-x.
7. Xing X, Liang G, Wang C, Jacobs N, Lin A-L. Self-supervised learning application on COVID-19 chest X-ray image classification using masked autoencoder. Bioengineering. 2023;10:901. https://doi.org/10.3390/bioengineering10080901.
8. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng. 2022;6:1399–406. https://doi.org/10.1038/s41551-022-00936-9.
9. Dong H, Zhang Y, Gu H, Konz N, Zhang Y, Mazurowski MA. SWSSL: sliding window-based self-supervised learning for anomaly detection in high-resolution images. IEEE Trans Med Imaging. 2023;42:3860–70. https://doi.org/10.1109/TMI.2023.3314318.
10. Nguyen X-B, Lee GS, Kim SH, Yang HJ. Self-supervised learning based on spatial awareness for medical image analysis. IEEE Access. 2020;8:162973–81. https://doi.org/10.1109/ACCESS.2020.3021469.
11. Tajbakhsh N, Hu Y, Cao J, Yan X, Xiao Y, Lu Y, Liang J, Terzopoulos D, Ding X. Surrogate supervision for medical image analysis: effective deep learning from limited quantities of labeled data. In: 2019 IEEE 16th international symposium biomedical imaging ISBI 2019; 2019. p. 1251–5. https://doi.org/10.1109/ISBI.2019.8759553.
12. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models genesis. Med Image Anal. 2021;67: 101840. https://doi.org/10.1016/j.media.2020.101840.
13. Zhou Z, Sodha V, Siddiquee MMR, Feng R, Tajbakhsh N, Gotway MB, Liang J. Models genesis: generic autodidactic models for 3D medical image analysis. In: International conference on medical image computing and computer-assisted intervention MICCAI. 2019; 11767: 384–93. https://doi.org/10.1007/978-3-030-32251-9_42.
14. Wu R, Liang C, Li Y, Shi X, Zhang J, Huang H. Self-supervised transfer learning framework driven by visual attention for benign–malignant lung nodule classification on chest CT. Expert Syst Appl. 2023;215: 119339. https://doi.org/10.1016/j.eswa.2022.119339.
15. Huang H, Wu R, Li Y, Peng C. Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic CT. IEEE J Biomed Health Inform. 2022;26:3860–71. https://doi.org/10.1109/JBHI.2022.3171851.
16. Sun L, Yu K, Batmanghelich K. Context matters: graph-based self-supervised representation learning for medical images. In: Proceedings of the AAAI conference artificial intelligence. 2021; 35: 4874–82.
17. Liu J, Cao L, Akin O, Tian Y. Robust and accurate pulmonary nodule detection with self-supervised feature learning on domain adaptation. Front Radiol. 2022. https://doi.org/10.3389/fradi.2022.1041518.
18. Shabani S, Homayounfar M, Vardhanabhuti V, Nikouei Mahani MA, Koohi-Moghadam M. Self-supervised region-aware segmentation of COVID-19 CT images using 3D GAN and contrastive learning. Comput Biol Med. 2022;149: 106033. https://doi.org/10.1016/j.compbiomed.2022.106033.
19. Tan Z, Yu Y, Meng J, Liu S, Li W. Self-supervised learning with self-distillation on COVID-19 medical image classification. Comput Methods Progr Biomed. 2024;243: 107876. https://doi.org/10.1016/j.cmpb.2023.107876.
20. Yu L, Zhang Z, Li X, Ren H, Zhao W, Xing L. Metal artifact reduction in 2D CT images with self-supervised cross-domain learning. Phys Med Biol. 2021;66: 175003. https://doi.org/10.1088/1361-6560/ac195c.
21. Wang Y, Song D, Wang W, Rao S, Wang X, Wang M. Self-supervised learning and semi-supervised learning for multi-sequence medical image classification. Neurocomputing. 2022;513:383–94. https://doi.org/10.1016/j.neucom.2022.09.097.
22. Kalapos A, Gyires-Tóth B. Self-supervised pretraining for 2D medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K, editors. Computer vision—ECCV 2022 workshop. Cham: Springer Nature Switzerland; 2023. p. 472–84. https://doi.org/10.1007/978-3-031-25082-8_31.
23. Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, Lippert C. 3D self-supervised methods for medical imaging. In: Proceedings of the 34th international conference on neural information processing system. Curran Associates Inc., Red Hook, NY, USA; 2020. p. 18158–72.
24. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. Med Image Anal. 2019;58: 101539. https://doi.org/10.1016/j.media.2019.101539.
25. Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Proceedings of the 34th international conference on neural information processing system. Curran Associates Inc., Red Hook, NY, USA; 2020. p. 12546–58.
26. Zhou Q, Zou H. A layer-wise fusion network incorporating self-supervised learning for multimodal MR image synthesis. Front Genet. 2022. https://doi.org/10.3389/fgene.2022.937042.
27. Fischer M, Hepp T, Gatidis S, Yang B. Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations. Comput Med Imaging Graph. 2023;104: 102174. https://doi.org/10.1016/j.compmedimag.2022.102174.

28. Srinidhi CL, Kim SW, Chen F-D, Martel AL. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med Image Anal. 2022;75: 102256. https://doi.org/10.1016/j.media.2021.102256.
29. Spitzer H, Kiwitz K, Amunts K, Harmeling S, Dickscheid T. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical image computing and computer assisted intervention—MICCAI 2018, Springer International Publishing, Cham; 2018. p. 663–71. https://doi.org/10.1007/978-3-030-00931-1_76.
30. Lu AX, Kraus OZ, Cooper S, Moses AM. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLOS Comput Biol. 2019;15: e1007348. https://doi.org/10.1371/journal.pcbi.1007348.
31. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. Mach Learn Appl. 2022;7: 100198. https://doi.org/10.1016/j.mlwa.2021.100198.
32. Veeling BS, Linmans J, Winkens J, Cohen T, Welling M. Rotation equivariant CNNs for digital pathology. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical image computing and computer assisted intervention—MICCAI 2018, Springer International Publishing, Cham; 2018. p. 210–8. https://doi.org/10.1007/978-3-030-00934-2_24.
33. Zhao F, Wang Z, Du H, He X, Cao X. Self-supervised triplet contrastive learning for classifying endometrial histopathological images. IEEE J Biomed Health Inform. 2023;27:5970–81. https://doi.org/10.1109/JBHI.2023.3314663.
34. Yang P, Yin X, Lu H, Hu Z, Zhang X, Jiang R, Lv H. CS-CO: a hybrid self-supervised visual representation learning method for H&E-stained histopathological images. Med Image Anal. 2022;81: 102539. https://doi.org/10.1016/j.media.2022.102539.
35. Lin Y, Qu Z, Chen H, Gao Z, Li Y, Xia L, Ma K, Zheng Y, Cheng K-T. Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. Med Image Anal. 2023;89: 102933. https://doi.org/10.1016/j.media.2023.102933.
36. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Conference on computing vision and pattern recognition. Workshop IEEE computer society conference on computing vision and pattern recognition. Workshop 2021; 2021. 14318. https://doi.org/10.1109/CVPR46437.2021.01409.
37. Liu H, Liu J, Hou S, Tao T, Han J. Perception consistency ultrasound image super-resolution via self-supervised CycleGAN. Neural Comput Appl. 2023;35:12331–41. https://doi.org/10.1007/s00521-020-05687-9.
38. Zhu J, Li Y, Hu Y, Ma K, Zhou SK, Zheng Y. Rubik's Cube+: a self-supervised feature learning framework for 3D medical image analysis. Med Image Anal. 2020;64: 101746. https://doi.org/10.1016/j.media.2020.101746.
39. Zhuang X, Li Y, Hu Y, Ma K, Yang Y, Zheng Y. Self-supervised feature learning for 3D medical images by playing a Rubik's cube. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap PT, Khan A, editors. Medical image computing and computer assisted intervention—MICCAI 2019, Springer International Publishing, Cham; 2019. p. 420–8. https://doi.org/10.1007/978-3-030-32251-9_46.
40. Haghighi F, Taher MRH, Gotway MB, Liang J. Self-supervised learning for medical image analysis: discriminative, restorative, or adversarial? Med Image Anal. 2024. https://doi.org/10.1016/j.media.2024.103086.
41. Taleb A, Lippert C, Klein T, Nabi M. Multimodal Self-supervised learning for medical image analysis. In: Feragen A, Sommer S, Schnabel J, Nielsen M, editors. Information processing in medical Imaging, Springer International Publishing, Cham; 2021. p. 661–73. https://doi.org/10.1007/978-3-030-78191-0_51.
42. Zhao H, Li Y, He N, Ma K, Fang L, Li H, Zheng Y. Anomaly detection for medical images using self-supervised and translation-consistent features. IEEE Trans Med Imaging. 2021;40:3641–51. https://doi.org/10.1109/TMI.2021.3093883.
43. Haghighi F, Taher MRH, Gotway MB, Liang J. DiRA: discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: 2022. p. 20824–34. https://openaccess.thecvf.com/content/CVPR2022/html/Haghighi_DiRA_Discriminative_Restorative_and_Adversarial_Learning_for_Self-Supervised_Medical_Image_CVPR_2022_paper.html. Accessed 13 Mar 2024.
44. Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Neumann D, Patel P, Vishwanath RS, Balter JM, Cao Y, Grbic S, Comaniciu D. Contrastive self-supervised learning from 100 million medical images with optional supervision. J Med Imaging Bellingham Wash. 2022;9: 064503. https://doi.org/10.1117/1.JMI.9.6.064503.
45. Tao X, Li Y, Zhou W, Ma K, Zheng Y. Revisiting Rubik's cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical image on computing and computer assisted intervention—MICCAI 2020, Springer International Publishing, Cham; 2020. p. 238–248. https://doi.org/10.1007/978-3-030-59719-1_24.
46. Zhou HY, Lu C, Yang S, Han X, Yu Y. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In: 2021 IEEECVF international conference on computer vision ICCV; 2021. p. 3479–89. https://doi.org/10.1109/ICCV48922.2021.00348.
47. Haghighi F, Hosseinzadeh Taher MR, Zhou Z, Gotway MB, Liang J. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical on image computing and computers assisted intervention—MICCAI 2020. Springer International Publishing, Cham; 2020. p. 137–47. https://doi.org/10.1007/978-3-030-59710-8_14.
48. Qayyum A, Razzak I, Mazher M, Khan T, Ding W, Niederer S. Two-stage self-supervised contrastive learning aided transformer for real-time medical image segmentation. IEEE J Biomed Health Inform. 2023. https://doi.org/10.1109/JBHI.2023.3340956.
49. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging. 2016;35:1299–312. https://doi.org/10.1109/TMI.2016.2535302.

50. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.
51. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: 2014 IEEE conference on computer vision and pattern recognition; 2014. p. 1717–24. https://doi.org/10.1109/CVPR.2014.222.
52. Goodfellow I, Bengio Y, Courville A. Deep Learning. London: MIT Press; 2016.
53. Zhang C, Zheng H, Gu Y. Dive into the details of self-supervised learning for medical image analysis. Med Image Anal. 2023;89: 102879. https://doi.org/10.1016/j.media.2023.102879.
54. Bau D, Zhou B, Khosla A, Oliva A, Torralba A. Network dissection: quantifying interpretability of deep visual representations. In: 2017 IEEE conference on computer vision and pattern recognition CVPR; 2017. p. 3319–27. https://doi.org/10.1109/CVPR.2017.354.
55. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52. https://doi.org/10.1007/s11263-015-0816-y.
56. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Comput Sci. 2022;8: e1045. https://doi.org/10.7717/peerj-cs.1045.
57. Huang S-C, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. Npj Digit Med. 2023;6:1–16. https://doi.org/10.1038/s41746-023-00811-0.
58. VanBerlo B, Hoey J, Wong A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. BMC Med Imaging. 2024;24:79. https://doi.org/10.1186/s12880-024-01253-0.
59. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989;1:541–51. https://doi.org/10.1162/neco.1989.1.4.541.
60. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference computer vision and pattern recognition CVPR; 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
61. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, CoRR; 2014. https://www.semanticscholar.org/paper/Very-Deep-Convolutional-Networks-for-Large-Scale-Simonyan-Zisserman/eb42cf88027de515750f230b23b1a057dc782108. Accessed 19 Jan 2024.
62. Lin M, Chen Q, Yan S. Network in network. In: International conference learning represents; 2013.
63. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence; 2017. 31. https://doi.org/10.1609/aaai.v31i1.11231.
64. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing system. Curran Associates Inc., Red Hook, NY, USA; 2017. p. 6000–10.
65. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer vision—ECCV 2020. Springer International Publishing, Cham; 2020. p. 213–29. https://doi.org/10.1007/978-3-030-58452-8_13.
66. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: 2018 IEEECVF conference on computer vision and pattern recognition; 2018. p. 7794–803. https://doi.org/10.1109/CVPR.2018.00813.
67. Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, Tran D. Image transformer. In: Proceedings of the 35th international conference on machine learning, PMLRl 2018. p. 4055–64. https://proceedings.mlr.press/v80/parmar18a.html. Accessed 17 Dec 2023.
68. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEECVF international conference computer vision ICCV; 2021. p. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
69. Wang D, Fan F, Wu Z, Liu R, Wang F, Yu H. CTformer: convolution-free Token2Token dilated vision transformer for low-dose CT denoising. Phys Med Biol. 2023;68: 065012. https://doi.org/10.1088/1361-6560/acc000.
70. Wu Y, Qi S, Sun Y, Xia S, Yao Y, Qian W. A vision transformer for emphysema classification using CT images. Phys Med Biol. 2021;66: 245016. https://doi.org/10.1088/1361-6560/ac3dc8.
71. Gao X, Khan MHM, Hui R, Tian Z, Qian Y, Gao A, Baichoo S. COVID-VIT: classification of Covid-19 from 3D CT chest images based on vision transformer model. In: 2022 3rd international conference generation computing applications NextComp; 2022. p. 1–4. https://doi.org/10.1109/NextComp55567.2022.9932246.
72. Islam MN, Hasan M, Hossain MK, Alam MGR, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Sci Rep. 2022;12:11440. https://doi.org/10.1038/s41598-022-15634-4.
73. Krishnan KS, Krishnan KS. Vision transformer based COVID-19 detection using chest X-rays. In: 2021 6th international conference on signal processing, computing and control ISPCC; 2021. p. 644–8. https://doi.org/10.1109/ISPCC53510.2021.9609375.
74. Gatys L, Ecker A, Bethge M. A neural algorithm of artistic style. J Vis. 2016;16:326. https://doi.org/10.1167/16.12.326.
75. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: 2020 IEEECVF conference on computer vision and pattern recognition CVPR; 2020. p. 9726–35. https://doi.org/10.1109/CVPR42600.2020.00975.
76. van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. In: Proceedings of the 33rd international conference on machine learning, PMLRl 2016. p. 1747–56. https://proceedings.mlr.press/v48/oord16.html. Accessed 26 Nov 2023.
77. van den Oord A, Kalchbrenner N, Espeholt L, Kavukcuoglu K, Vinyals O, Graves A. Conditional image generation with PixelCNN decoders. In: Advances in neural information processing system. Curran Associates, Inc.; 2016. https://papers.nips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html. Accessed 26 Nov 2023.

78. Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using Real NVP. In: 2016. https://openreview.net/forum?id=HkpbnH9lx. Accessed 26 Nov 2023.
79. Kingma DP, Dhariwal P. Glow: generative flow with invertible 1x1 convolutions. In: Advances in neural information processing system. Curran Associates, Inc.; 2018. https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html. Accessed 26 Nov 2023.
80. Razavi A, van den Oord A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. In: Advances in neural information processing system. Curran Associates, Inc.; 2019. https://papers.nips.cc/paper/2019/hash/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Abstract.html. Accessed 26 Nov 2023.
81. Shi C, Xu M, Zhu Z, Zhang W, Zhang M, Tang J. GraphAF: a flow-based autoregressive model for molecular graph generation; 2020. https://iclr.cc/virtual_2020/poster_S1esMkHYPr.html. Accessed 26 Nov 2023.
82. Germain M, Gregor K, Murray I, Larochelle H. MADE: masked autoencoder for distribution estimation. In: Proceedings of the 32nd international conference on machine learning, PMLR; 2015. p. 881–9. https://proceedings.mlr.press/v37/germain15.html. Accessed 26 Nov 2023.
83. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing system. Curran Associates, Inc.; 2019. https://papers.nips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html. Accessed 26 Nov 2023.
84. Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, Kong T. iBOT: image BERT pre-training with online Tokenizer, ArXiv; 2021. https://www.semanticscholar.org/paper/iBOT%3A-Image-BERT-Pre-Training-with-Online-Tokenizer-Zhou-Wei/9653c070724e44f023e8cc3ec79f0b9e6d59480d. Accessed 12 Nov 2023.
85. Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision transformers. In: 2021 IEEECVF interational conference on computer vision ICCV; 2021. p. 9630–40. https://doi.org/10.1109/ICCV48922.2021.00951.
86. Bachman P, Hjelm RD, Buchwalter W. Learning representations by maximizing mutual information across views. In: advances in neural information processing system. Curran Associates, Inc.; 2019. https://papers.nips.cc/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html. Accessed 9 Mar 2024.
87. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: 2018 IEEECVF confernece on computer vision and pattern recognition; 2018. p. 3733–42. https://doi.org/10.1109/CVPR.2018.00393.
88. Ye M, Zhang X, Yuen PC, Chang SF. Unsupervised embedding learning via invariant and spreading instance feature. In: 2019 IEEECVF confernece on computer vision and pattern recognition CVPR; 2019. p. 6203–12. https://doi.org/10.1109/CVPR.2019.00637.
89. Tian Y, Krishnan D, Isola P. Contrastive multiview coding. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer vision—ECCV 2020. Springer International Publishing, Cham; 2020. p. 776–94. https://doi.org/10.1007/978-3-030-58621-8_45.
90. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th international conference on machine learning, JMLR.org; 2020. p. 1597–607.
91. Grill JB, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, Doersch C, Pires BA, Guo ZD, Azar MG, Piot B, Kavukcuoglu K, Munos R, Valko M. Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th international conference on neural information processing system. Curran Associates Inc., Red Hook, NY, USA; 2020. p. 21271–84.
92. Chen X, He K. Exploring simple siamese representation learning. In: 2021 IEEECVF conference computer vision and pattern recognition CVPR; 2021. p. 15745–53. https://doi.org/10.1109/CVPR46437.2021.01549.
93. Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. In: 2021 IEEECVF international conference on computer vision ICCV, IEEE, Montreal, QC, Canada; 2021. p. 9620–29. https://doi.org/10.1109/ICCV48922.2021.00950.
94. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–324. https://doi.org/10.1109/5.726791.
95. Krizhevsky A. Learning multiple layers of features from tiny images; 2009. https://api.semanticscholar.org/CorpusID:18268744.
96. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: 2019 IEEECVF conference on computer vision and pattern recognition CVPR; 2018. p. 4396–405.
97. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison; 2019. https://doi.org/10.48550/arXiv.1901.07031.
98. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners; 2022 IEEECVF conference on computer vision and pattern recognition CVPR; 2022. p. 15979–88. https://doi.org/10.1109/CVPR52688.2022.01553.
99. Imagawa K, Shiomoto K. Evaluation of effectiveness of self-supervised learning in chest X-ray imaging to reduce annotated images. J Imaging Inform Med. 2024;37:1618–24. https://doi.org/10.1007/s10278-024-00975-5.
100. Yan C, Yan H, Liang W, Yin M, Luo H, Luo J. DP-SSLoRA: a privacy-preserving medical classification model combining differential privacy with self-supervised low-rank adaptation. Comput Biol Med. 2024;179: 108792. https://doi.org/10.1016/j.compbiomed.2024.108792.
101. Sheng H, Ma L, Samson J-F, Liu D. BarlowTwins-CXR: enhancing chest X-ray abnormality localization in heterogeneous data with cross-domain self-supervised learning. BMC Med Inform Decis Mak. 2024;24:126. https://doi.org/10.1186/s12911-024-02529-9.
102. Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow Twins: self-supervised learning via redundancy reduction. In: Proceedings of the 38th international conference on machine learning. PMLR; 2021. p. 12310–20. https://proceedings.mlr.press/v139/zbontar21a.html. Accessed 13 Sept 2024.

103. Zhang Y, Kohne J, Wittrup E, Najarian K. Three-stage framework for accurate pediatric chest X-ray diagnosis using self-supervision and transfer learning on small datasets. Diagnostics. 2024;14:1634. https://doi.org/10.3390/diagnostics14151634.
104. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6:317. https://doi.org/10.1038/s41597-019-0322-0.
105. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep. 2020;10:19549. https://doi.org/10.1038/s41598-020-76550-z.
106. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases; 2017. p. 2097–106. https://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html. Accessed 14 Feb 2024.
107. Nguyen NH, Pham HH, Tran TT, Nguyen TNM, Nguyen HQ. VinDr-PCXR: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children; 2022. https://doi.org/10.1101/2022.03.04.22271937.
108. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101:e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.
109. Zeng Y, Liu H, Hu J, Zhao Z, She Q. Pretrained subtraction and segmentation model for coronary angiograms. Sci Rep. 2024;14:19888. https://doi.org/10.1038/s41598-024-71063-5.
110. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computer and computing assisted intervention—MICCAI 2015. Springer International Publishing, Cham; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
111. Kumar K, Tanwar S, Kumar S. MAN—C: a masked autoencoder neural cryptography based encryption scheme for CT scan images. MethodsX. 2024;12: 102738. https://doi.org/10.1016/j.mex.2024.102738.
112. Guo Z, Islam NU, Gotway MB, Liang J. Stepwise incremental pretraining for integrating discriminative, restorative, and adversarial learning. Med Image Anal. 2024;95: 103159. https://doi.org/10.1016/j.media.2024.103159.
113. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Springer International Publishing, Cham; 2022. p. 272–84. https://doi.org/10.1007/978-3-031-08999-2_22.
114. Yu K, Sun L, Chen J, Reynolds M, Chaudhary T, Batmanghelich K. DrasCLR: a self-supervised framework of learning disease-related and anatomy-specific representation for 3D lung CT images. Med Image Anal. 2024;92: 103062. https://doi.org/10.1016/j.media.2023.103062.
115. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber M-A, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp Ç, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin H-C, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34:1993–2024. https://doi.org/10.1109/TMI.2014.2377694.
116. Qi L, Jiang Z, Shi W, Qu F, Feng G. GMIM: self-supervised pre-training for 3D medical image segmentation with adaptive and hierarchical masked image modeling. Comput Biol Med. 2024;176: 108547. https://doi.org/10.1016/j.compbiomed.2024.108547.
117. Cox J, Liu P, Stolte SE, Yang Y, Liu K, See KB, Ju H, Fang R. BrainSegFounder: towards 3D foundation models for neuroimage segmentation. Med Image Anal. 2024;97: 103301. https://doi.org/10.1016/j.media.2024.103301.
118. Despotovic V, Kim S-Y, Hau A-C, Kakoichankava A, Klamminger GG, Borgmann FBK, Frauenknecht KBM, Mittelbronn M, Nazarov PV. Glioma subtype classification from histopathological images using in-domain and out-of-domain transfer learning: an experimental study. Heliyon. 2024;10: e27515. https://doi.org/10.1016/j.heliyon.2024.e27515.
119. Ye J, Kalra S, Miri MS. Cluster-based histopathology phenotype representation learning by self-supervised multi-class-token hierarchical ViT. Sci Rep. 2024;14:3202. https://doi.org/10.1038/s41598-024-53361-0.
120. Chen X, Fan H, Girshick RB, He K. Improved baselines with momentum contrastive learning, ArXiv; 2020. https://www.semanticscholar.org/paper/Improved-Baselines-with-Momentum-Contrastive-Chen-Fan/a1b8a8df281bbaec148a897927a49ea47ea31515. Accessed 11 Mar 2024.
121. Philipsen RHHM, Maduskar P, Hogeweg L, Melendez J, Sánchez CI, van Ginneken B. Localized energy-based normalization of medical images: application to chest radiography. IEEE Trans Med Imaging. 2015;34:1965–75. https://doi.org/10.1109/TMI.2015.2418031.
122. Anand D, Annangi P, Sudhakar P. Benchmarking self-supervised representation learning from a million cardiac ultrasound images. In: 2022 44th annual international conference of the IEEE engineering in medicine & biology society EMBC, IEEE, Glasgow, Scotland, United Kingdom; 2022. p. 529–32. https://doi.org/10.1109/EMBC48229.2022.9871511.
123. Amunts K, LepageLePage C, Borgeat L, Mohlberg H, Dickscheid T, Rousseau M-É, Bludau S, Bazin P-L, Lewis LB, Oros-Peusquens A-M, Shah NJ, Lippert T, Zilles K, Evans AC. BigBrain: an ultrahigh-resolution 3D human brain model. Science. 2013;340:1472–5. https://doi.org/10.1126/science.1235381.
124. Kumar N, Verma R, Anand D, Zhou Y, Onder OF, Tsougenis E, Chen H, Heng P-A, Li J, Hu Z, Wang Y, Koohbanani NA, Jahanifar M, Tajeddin NZ, Gooya A, Rajpoot N, Ren X, Zhou S, Wang Q, Shen D, Yang C-K, Weng C-H, Yu W-H, Yeh C-Y, Yang S, Xu S, Yeung PH, Sun P, Mahbod A, Schaefer G, Ellinger I, Ecker R, Smedby O, Wang C, Chidester B, Ton T-V, Tran M-T, Ma J, Do MN, Graham S, Vu QD, Kwak JT, Gunda A, Chunduri R, Hu C, Zhou X, Lotfi D, Safdari R,

Kascenas A, O'Neil A, Eschweiler D, Stegmaier J, Cui Y, Yin B, Chen K, Tian X, Gruening P, Barth E, Arbel E, Remer I, Ben-Dor A, Sirazitdinova E, Kohl M, Braunewell S, Li Y, Xie X, Shen L, Ma J, Baksi KD, Khan MA, Choo J, Colomer A, Naranjo V, Pei L, Iftekharuddin KM, Roy K, Bhattacharjee D, Pedraza A, Bueno MG, Devanathan S, Radhakrishnan S, Koduganty P, Wu Z, Cai G, Liu X, Wang Y, Sethi A. A multi-organ nucleus segmentation challenge. IEEE Trans Med Imaging. 2020;39:1380–91. https://doi.org/10.1109/TMI.2019.2947628.

125. Vu QD, Graham S, Kurc T, To MNN, Shaban M, Qaiser T, Koohbanani NA, Khurram SA, Kalpathy-Cramer J, Zhao T, Gupta R, Kwak JT, Rajpoot N, Saltz J, Farahani K. Methods for segmentation and classification of digital microscopy tissue images. Front Bioeng Biotechnol. 2019;7:53. https://doi.org/10.3389/fbioe.2019.00053.
126. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, Ginneken B, Kopp-Schneider A, Landman B, Litjens G, Menze BH, Ronneberger O, Summers R, Bilic P, Christ P, Do R, Gollub M, Golia-Pernicka J, Heckers S, Jarnagin W, McHugo M, Napel S, Vorontsov E, Maier-Hein L, Cardoso MJ. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, ArXiv; 2019. https://www.semanticscholar.org/paper/A-large-annotated-medical-image-dataset-for-the-and-Simpson-Antonelli/4654aa505e5bcdb089d0df202cd7ceabc9d2d41f. Accessed 13 Mar 2024.
127. Kavur AE, Gezer NS, Barış M, Aslan S, Conze PH, Groza V, Pham DD, Chatterjee S, Ernst P, Özkan S, Baydar B, Lachinov D, Han S, Pauli J, Isensee F, Perkonigg M, Sathish R, Rajan R, Sheet D, Dovletov G, Speck O, Nürnberger A, Maier-Hein KH, Bozdağı Akar G, Ünal G, Dicle O, Selver MA. CHAOS Challenge—combined (CT-MR) healthy abdominal organ segmentation. Med Image Anal. 2021;69: 101950. https://doi.org/10.1016/j.media.2020.101950.
128. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 2017;4: 170117. https://doi.org/10.1038/sdata.2017.117.
129. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M, Alberts E, Lipkova J, Freymann J, Kirby J, Bilello M, Fathallah-Shaykh H, Wiest R, Kirschke J, Wiestler B, Colen R, Kotrotsou A, Lamontagne P, Marcus D, Milchenko M, Nazeri A, Weber MA, Mahajan A, Baid U, Gerstner E, Kwon D, Acharya G, Agarwal M,. Alam M, Albiol A, Albiol A, Albiol FJ, Alex V, Allinson N, Amorim PHA, Amrutkar A, Anand G, Andermatt S, Arbel T, Arbelaez P, Avery A, Azmat MPB, Bai W, Banerjee S, Barth B, Batchelder T, Batmanghelich K, Battistella E, Beers A, Belyaev M, Bendszus M, Benson E, Bernal J, Bharath HN, Biros G, Bisdas S, Brown J, Cabezas M, Cao S, x Cardoso S, Carver EN, Casamitjana A, Castillo LS, Catà M, Cattin P, Cerigues A, Chagas VS, Chandra S, Chang YJ, Chang S, Chang K, Chazalon J, Chen S, Chen W, Chen JW, Chen Z, Cheng K, Choudhury AR, Chylla R, Clérigues A, Colleman S, Colmeiro RGR, Combalia M, Costa A, Cui X, Dai Z, Dai L, Daza LA, Deutsch E, Ding C, Dong C, Dong S, Dudzik W, Eaton-Rosen Z, Egan G, Escudero G, Estienne T, Everson R, Fabrizio J, Fan Y, Fang L, Feng X, Ferrante E, Fidon L, Fischer M, French AP, Fridman N, Fu H, Fuentes D, Gao Y, Gates E, Gering D, Gholami A, Gierke W, Glocker B, Gong M, González-Villá S, Grosges T, Guan Y, Guo S, Gupta S, Han WS, Han IS, Harmuth K, He H, Hernández-Sabaté A, Herrmann E, Himthani N, Hsu W, Hsu C, Hu X, Hu X, Hu Y, Hu Y, Hua R, Huang TY, Huang W, Van Huffel S, Huo QVHV, Iftekharuddin KM, Isensee F, Islam M, Jackson AS, Jambawalikar SR, Jesson A, Jian W, Jin P, Jose VJM, Jungo A, Kainz B, Kamnitsas K, Kao PY, Karnawat A, Kellermeier T, Kermi A, Keutzer K, Khadir MT, Khened M, Kickingereder P, Kim G, King N, Knapp H, Knecht U, Kohli L, Kong D, Kong X, Koppers S, Kori A, Krishnamurthi G, Krivov E, Kumar P, Kushibar K, Lachinov D, Lambrou T, Lee J, Lee C, Lee Y, Lee M, Lefkovits S, Lefkovits L, Levitt J, Li T, Li H, Li W, Li H, Li X, Li Y, Li H, Li Z, Li X, Li Z, Li X, Li W, Lin ZS, Lin F, Lio P, Liu C, Liu B, Liu X, Liu M, Liu J, Liu L, Llado X, Lopez MM, Lorenzo PR, Lu Z, Luo L, Luo Z, Ma J, Ma K, Mackie T, Madabushi A, Mahmoudi I, Maier-Hein KH, Maji P, Mammen CP, Mang A, Manjunath BS, Marcinkiewicz M, McDonagh S, McKenna S, McKinley R, Mehl M, Mehta S, Mehta R, Meier R, Meinel C, Merhof D, Meyer C, Miller R, Mitra S, Moiyadi A, Molina-Garcia D, Monteiro MAB, Mrukwa G, Myronenko A, Nalepa J, Ngo T, Nie D, Ning H, Niu C, Nuechterlein NK, Oermann E, Oliveira A, Oliveira DDC, Oliver A, Osman AFI, Ou YN, Ourselin S, Paragios N, Park MS, Paschke B, Pauloski JG, Pawar K, Pawlowski N, Pei L, Peng S, Pereira SM, Perez-Beteta J, Perez-Garcia VM, Pezold S, Pham B, Phophalia A, Piella G, Pillai GN, Piraud M, Pisov M, Popli A, Pound MP, Pourreza R, Prasanna P, Prkovska V, Pridmore TP, Puch S, Puybareau É, Qian B, Qiao X, Rajchl M, Rane S, Rebsamen M, Ren H, Ren X, Revanuru K, Rezaei M, Rippel O, Rivera LC, Robert C, Rosen B, Rueckert D, Safwan M, Salem M, Salvi J, Sanchez I, Sánchez I, Santos HM, Sartor E, Schellingerhout D, Scheufele K, Scott MR, Scussel AA, Sedlar S, Serrano-Rubio JP, Shah NJ, Shah N, Shaikh M, Shankar BU, Shboul Z, Shen H, Shen D, Shen L, Shen H, Shenoy V, Shi F, Shin HE, Shu H, Sima D, Sinclair M, Smedby O, Snyder JM, Soltaninejad M, Song G, Soni M, Stawiaski J, Subramanian S, Sun L, Sun R, Sun J, Sun K, Sun Y, Sun G, Sun S, Suter YR, Szilagyi L, Talbar S, Tao D, Tao D, Teng Z, Thakur S, Thakur MH, Tharakan S, Tiwari P, Tochon G, Tran T, Tsai YM, Tseng KL, Tuan TA, Turlapov V, Tustison N, Vakalopoulou M, Valverde S, Vanguri R, Vasiliev E, Ventura J, Vera L, Vercauteren T, Verrastro CA, Vidyaratne L, Vilaplana V, Vivekanandan A, Wang G, Wang Q, Wang CJ, Wang W, Wang D, Wang R, Wang Y, Wang C, Wang G, Wen N, Wen X, Weninger L, Wick W, Wu S, Wu Q, Wu Y, Xia Y, Xu Y, Xu X, Xu P, Yang TL, Yang X, Yang HY, Yang J, Yang H, Yang G, Yao H, Ye X, Yin C, Young-Moxon B, Yu J, Yue X, Zhang S, Zhang A, Zhang K, Zhang X, Zhang L, Zhang X, Zhang Y, Zhang L, Zhang J, Zhang X, Zhang T, Zhao S, Zhao Y, Zhao X, Zhao L, Zheng Y, Zhong L, Zhou C, Zhou X, Zhou F, Zhu H, Zhu J, Zhuge Y, Zong W, Kalpathy-Cramer J, Farahani K, Davatzikos C, van Leemput K B. Menze, identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge; 2019. https://doi.org/10.48550/arXiv.1811.02629.
130. LIDC-IDRI; 2015. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.
131. Kermany D, Zhang K, Goldbaum M. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. Mendeley Data. 2018. https://doi.org/10.17632/rscbjbr9sj.2.
132. Armato SG, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J, Farahani K. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging Bellingham Wash. 2018;5: 044501. https://doi.org/10.1117/1.JMI.5.4.044501.
133. Setio AAA, Traverso A, de Bel T, Berens MSN, van den Bogaard C, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B, van der Gugten R, Heng PA, Jansen B, de Kaste MMJ, Kotov V, Lin JY-H, Manders JTMC, Sóñora-Mengana A, García-Naranjo JC, Papavasileiou E, Prokop M, Saletta M, Schaefer-Prokop CM, Scholten ET, Scholten L, Snoeren MM, Torres EL, Vandemeulebroucke J, Walasek N, Zuidhof GCA, van Ginneken B, Jacobs C. Validation, comparison, and

combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal. 2017;42:1–13. https://doi.org/10.1016/j.media.2017.06.015.
134. Roth HR, Lu L, Farag A, Shin HC, Liu J, Turkbey EB, Summers RM. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab N, Hornegger J, Wells WM, Frangi A, editors. Medical image computer and computing-assisted intervention—MICCAI 2015. Springer International Publishing, Cham; 2015. p. 556–64. https://doi.org/10.1007/978-3-319-24553-9_68.
135. Kuijf HJ, Bennink E, Vincken KL, Weaver N, Biessels GJ, Viergever MA. MR brain segmentation challenge 2018 Data; 2024. https://doi.org/10.34894/E0U32Q.
136. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P-A, Cetin I, Lekadir K, Camara O, Gonzalez Ballester MA, Sanroma G, Napel S, Petersen S, Tziritas G, Grinias E, Khened M, Kollerathu VA, Krishnamurthi G, Rohe M-M, Pennec X, Sermesant M, Isensee F, Jager P, Maier-Hein KH, Full PM, Wolf I, Engelhardt S, Baumgartner CF, Koch LM, Wolterink JM, Isgum I, Jang Y, Hong Y, Patravali J, Jain S, Humbert O, Jodoin P-M. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging. 2018;37:2514–25. https://doi.org/10.1109/TMI.2018.2837502.
137. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, Litjens G, Menze B, Ronneberger O, Summers RM, Van Ginneken B, Bilello M, Bilic P, Christ PF, Do RKG, Gollub MJ, Heckers SH, Huisman H, Jarnagin WR, McHugo MK, Napel S, Pernicka JSG, Rhode K, Tobon-Gomez C, Vorontsov E, Meakin JA, Ourselin S, Wiesenfarth M, Arbeláez P, Bae B, Chen S, Daza L, Feng J, He B, Isensee F, Ji Y, Jia F, Kim I, Maier-Hein K, Merhof D, Pai A, Park B, Perslev M, Rezaiifar R, Rippel O, Sarasua I, Shen W, Son J, Wachinger C, Wang L, Wang Y, Xia Y, Xu D, Xu Z, Zheng Y, Simpson AL, Maier-Hein L, Cardoso MJ. The medical segmentation decathlon. Nat Commun. 2022;13:4128. https://doi.org/10.1038/s41467-022-30695-9.
138. Metz CT, Schaap M, Weustink AC, Mollet NR, Van Walsum T, Niessen WJ. Coronary centerline extraction from CT coronary angiography images using a minimum cost path approach. Med Phys. 2009;36:5568–79. https://doi.org/10.1118/1.3254077.
139. Martel AL, Nofech-Mozes S, Salama S, Akbar S, Peikari M. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital. Pathology. 2019. https://doi.org/10.7937/TCIA.2019.4YIBTJNO.
140. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318:2199–210. https://doi.org/10.1001/jama.2017.14585.
141. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, Gaiser T, Marx A, Valous NA, Ferber D, Jansen L, Reyes-Aldasoro CC, Zörnig I, Jäger D, Brenner H, Chang-Claude J, Hoffmeister M, Halama N. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Med. 2019;16:e1002730. https://doi.org/10.1371/journal.pmed.1002730.
142. Jaeger S, Candemir S, Antani S, Wáng Y-XJ, Lu P-X, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg. 2014;4:475–7. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.
143. MICCAI 2019 Challenge, automatic structure segmentation for radiotherapy planning challenge 2019; 2019. https://structseg2019.grand-challenge.org/Dataset/.
144. Zhuang X, Shen J. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med Image Anal. 2016;31:77–87. https://doi.org/10.1016/j.media.2016.02.006.
145. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.
146. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M. COVID-19 image data collection: prospective predictions are the future. Mach Learn Biomed Imaging. 2020;1:1–38. https://doi.org/10.59275/j.melba.2020-48g7.
147. Tabik S, Gómez-Ríos A, Martín-Rodríguez JL, Sevillano-García I, Rey-Area M, Charte D, Guirado E, Suárez JL, Luengo J, Valero-González MA, García-Villanova P, Olmedo-Sánchez E, Herrera F. COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-Ray images. IEEE J Biomed Health Inform. 2020;24:3595–605. https://doi.org/10.1109/JBHI.2020.3037127.
148. Lung Adenocarcinoma Study; 2018. https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/lung-adenocarcinoma-study. Accessed 27 Mar 2024.
149. Lung Squamous Cell Carcinoma Study; 2018. https://www.cancer.gov/ccg/research/structural-genomics/tcga/studied-cancers/lung-squamous-cell-carcinoma-study. Accessed 27 Mar 2024.
150. Buda M, Saha A, Walsh R, Ghate S, Li N, Święcicki A, Lo JY, Mazurowski MA. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. JAMA Netw Open. 2021;4: e2119100. https://doi.org/10.1001/jamanetworkopen.2021.19100.
151. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010;7:32–43. https://doi.org/10.3109/15412550903499522.
152. Morozov SP, Andreychenko AE, Pavlov NA, Vladzymyrskyy AV, Ledikhova NV, Gombolevskiy VA, Blokhin IA, Gelezhe PB, Gonchar AV,Yu. V Chernina, MosMedData: chest CT scans with COVID-19 related findings dataset; 2020. https://doi.org/10.1101/2020.05.20.20100362.
153. D.J. Bell, R. Sharma, H. Knipe, COVID-19 CT Dataset, (2020). https://doi.org/10.53347/rID-73913.
154. Rapid A. Accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis. IEEE Trans Med Imaging. 2020;39:2638–52. https://doi.org/10.1109/TMI.2020.3001810.
155. Armato SG, Mclennan G, Bidaut L, Mcnitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, Macmahon H, van Beek EJR, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DPY, Roberts RY, Smith AR, Starkey A, Batra P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick

N, Freymann J, Kirby J, Hughes B, Vande A, Gupte S, Sallam M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY, Clarke LP. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38:915–31. https://doi.org/10.1118/1.3528204.

156. Ma J, Wang Y, An X, Ge C, Yu Z, Chen J, Zhu Q, Dong G, He J, He Z, Cao T, Zhu Y, Nie Z, Yang X. Toward data-efficient learning: a benchmark for COVID-19 CT lung and infection segmentation. Med Phys. 2021;48:1197–210. https://doi.org/10.1002/mp.14676.
157. 15K Chest X-Ray Images (COVID-19); n.d. https://www.kaggle.com/datasets/scipygaurav/15k-chest-xray-images-covid19. Accessed 13 Sept 2024.
158. Shih G, Wu CC, Halabi SS, Kohli MD, Prevedello LM, Cook TS, Sharma A, Amorosa JK, Arteaga V, Galperin-Aizenberg M, Gill RR, Godoy MCB, Hobbs S, Jeudy J, Laroia A, Shah PN, Vummidi D, Yaddanapudi K, Stein A. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell. 2019. https://doi.org/10.1148/ryai.2019180041.
159. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Abul Kashem SB, Islam MT, Al Maadeed S, Zughaier SM, Khan MS, Chowdhury MEH. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med. 2021;132: 104319. https://doi.org/10.1016/j.compbiomed.2021.104319.
160. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Emadi NA, Reaz MBI, Islam MT. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access. 2020;8:132665–76. https://doi.org/10.1109/ACCESS.2020.3010287.
161. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, Le DD, Pham CM, Tong HTT, Dinh DH, Do CD, Doan LT, Nguyen CN, Nguyen BT, Nguyen QV, Hoang AD, Phan HN, Nguyen AT, Ho PH, Ngo DT, Nguyen NT, Nguyen NT, Dao M, Vu V. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data. 2022;9:429. https://doi.org/10.1038/s41597-022-01498-w.
162. Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, Bell JD, Boultwood C, Collins R, Conroy MC, Crabtree N, Doherty N, Frangi AF, Harvey NC, Leeson P, Miller KL, Neubauer S, Petersen SE, Sellors J, Sheard S, Smith SM, Sudlow CLM, Matthews PM, Allen NE. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun. 2020;11:2624. https://doi.org/10.1038/s41467-020-15948-9.
163. Liew S-L, Lo BP, Donnelly MR, Zavaliangos-Petropulu A, Jeong JN, Barisano G, Hutton A, Simon JP, Juliano JM, Suri A, Wang Z, Abdullah A, Kim J, Ard T, Banaj N, Borich MR, Boyd LA, Brodtmann A, Buetefisch CM, Cao L, Cassidy JM, Ciullo V, Conforto AB, Cramer SC, Dacosta-Aguayo R, de la Rosa E, Domin M, Dula AN, Feng W, Franco AR, Geranmayeh F, Gramfort A, Gregory CM, Hanlon CA, Hordacre BG, Kautz SA, Khlif MS, Kim H, Kirschke JS, Liu J, Lotze M, MacIntosh BJ, Mataró M, Mohamed FB, Nordvik JE, Park G, Pienta A, Piras F, Redman SM, Revill KP, Reyes M, Robertson AD, Seo NJ, Soekadar SR, Spalletta G, Sweet A, Telenczuk M, Thielman G, Westlye LT, Winstein CJ, Wittenberg GF, Wong KA, Yu C. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Sci Data. 2022;9:320. https://doi.org/10.1038/s41597-022-01401-7.
164. Ji Y, Bai H, Yang J, Ge C, Zhu Y, Zhang R, Li Z, Zhang L, Ma W, Wan X, Luo P. AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Neural Inf Process Syst; 2022. https://www.semanticscholar.org/paper/AMOS%3A-A-Large-Scale-Abdominal-Multi-Organ-Benchmark-Ji-Bai/27d5abc68c2e5555bca47a3aff3f074b01f35b8c. Accessed 13 Sept 2024.

Publisher's Note