Machine Intelligence Research
www.mi-research.net

20(4), August 2023, 483-513
DOI: 10.1007/s11633-022-1406-4

A Review of Predictive and Contrastive Self-supervised Learning for Medical Images

Wei-Chien Wang[1] Euijoon Ahn[2] Dagan Feng[1] Jinman Kim[1]

[1] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Computer Science, The University of Sydney, Sydney NSW 2006, Australia

[2] College of Science and Engineering, James Cook University, Cairns QLD 4811, Australia

Abstract: Over the last decade, supervised deep learning on manually annotated big data has been progressing significantly on computer vision tasks. But, the application of deep learning in medical image analysis is limited by the scarcity of high-quality annotated medical imaging data. An emerging solution is self-supervised learning (SSL), among which contrastive SSL is the most successful approach to rivalling or outperforming supervised learning. This review investigates several state-of-the-art contrastive SSL algorithms originally on natural images as well as their adaptations for medical images, and concludes by discussing recent advances, current limitations, and future directions in applying contrastive SSL in the medical domain.

Keywords: Self-supervised learning (SSL), contrastive learning, deep learning, medical image analysis, computer vision.

Citation: W. C. Wang, E. Ahn, D. Feng, J. Kim. A review of predictive and contrastive self-supervised learning for medical images. *Machine Intelligence Research*, vol.20, no.4, pp.483–513, 2023. http://doi.org/10.1007/s11633-022-1406-4

1 Introduction

In recent years, deep learning networks, such as convolutional neural networks (CNNs), have seen massive progress in image analysis techniques. LeCun et al.[1] showed that CNNs achieved superior performance on diverse computer vision tasks, including semantic segmentation, image classification, object detection, and activity recognition. When a large amount of data and manually annotated labels are available, CNNs can automatically learn to approximate the relationship between the data and its labels. This type of deep learning algorithm is called supervised learning[2]. However, supervised learning can also be limited by large-scale labelled image data availability, where manual annotation is costly, labour-intensive, time-consuming, and prone to human subjectivity and error[3–5]. CNNs have also been broadly applied with medical imaging modalities and are considered state-of-the-art in many medical image analysis applications[6], such as with breast cancer classification[7], COVID-19 detection[8] and skin lesion analysis[9].

A variety of methods have been proposed to address the problem of limited training images and labels. Transfer learning has become the established method for this problem. With transfer learning, the model is pretrained on a larger image dataset, such as the ImageNet dataset of labelled natural images, and is then fine-tuned on a smaller dataset in the target domain that does not need to be from the same image domain, such as with a type of medical imaging modality[10]. Although transfer learning has demonstrated promising results in various medical imaging analysis applications[11, 12], there are known limitations[10, 13]. The primary limitation is that the image features extracted from the natural image dataset are not directly relevant to medical imaging datasets. Thus, supervised learning methods optimally designed using natural images do not necessarily translate well when applied to medical imaging analysis[10]. There are several key differences between medical images and natural images. As an example, medical images typically involve the identification of a small part of the images related to its pathologies or abnormalities, also known as regions of interest (ROIs), by utilizing variations in local textures from the whole image; examples of these are small red dots in retinal fundus images which are signs of microaneurysms and diabetic retinopathy[14], and white opaque local patches in chest X-ray images indicate consolidation and pneumonia. Natural image datasets, however, often have a large and salient object of interest in images. Another key difference is that, compared to natural images with diverse content and colours, a large variety of medical images, typically from X-ray, computer tomography (CT), and magnetic resonance imaging (MRI), are greyscale and have similar colours and content attributes

Review
Manuscript received on August 30, 2022; accepted on December 13, 2022; published online on June 3, 2023
Recommended by Associate Editor Jun-Zhou Huang
Colored figures are available in the online version at https://link.springer.com/journal/11633

across the image dataset, with fewer diversities and contrasts than natural images. Additionally, most medical image datasets have fewer image samples despite large variability in the image visual attributes between them, e.g., the number of images in the medical image datasets varies from one thousand[15] to one hundred thousand[16, 17]; in comparison, natural image datasets often have over 1 million images (e.g., ImageNet). Considering these differences between natural and medical images, transfer learning of a natural image pretrained model to medical image application is not always an effective solution. He et al.[18] demonstrated that pretraining on ImageNet merely accelerates the model convergence early during the training process.

To address the scarcity of medical image labels, researchers have been using other deep learning methods that do not entirely rely on labelled image data, and instead utilize abundant unlabelled image data[19, 20]. To address these issues, LeCun[21] presented the first concept of self-supervised learning (SSL) in 2017. His talk at the AAAI 2020 conference started to attract people's attention, and people gradually realized SSL had a potential future. He described, "In SSL, the system learns to predict part of its input from other parts of its input". SSL, as its name implies, creates supervisory information that is derived from the data itself. As represented in Fig. 1, there are some examples of SSL, such as predicting future data (yellow color) from past data (purple color) and predicting past data from present data (blue color). Taking sequential datasets, e.g., the target objects or images can be seen as anchors. The objects or images before these anchors can be seen as past data, while the objects or images after these anchors can be seen as future data. SSL has been widely employed in computer vision applications using natural images. For example, the Bootstrap your own latent (BYOL)[22] method obtained better image classification results than some supervised learning approaches on the ImageNet dataset. Other experiments[23, 24] further demonstrated how SSL could efficiently learn generalizable visual representations from the images. For example, Tendle and Hasan[25] analysed the SSL representations that were trained from the ImageNet source dataset and then fine-tuned on two different target datasets: One that was considerably different from the source dataset, and the other that was similar to the source dataset. By investigating the invariance property of learned representations, such as rotation, scale change, translation (vertical and horizontal) and background change, their experiments demonstrated that SSL representations produced better generalizability in contrast to supervised learning representations.

Among SSL methods, contrastive self-supervised learning, or contrastive SSL, is the most successful approach that achieved outstanding performance close to, or even surpassing, the supervised learning counterparts[26].

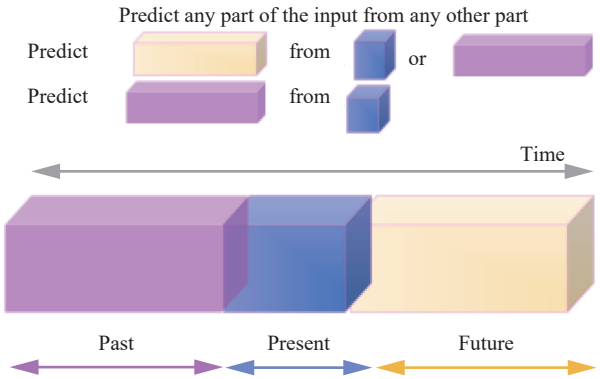

Fig. 1 Concept of self-supervised learning[1]

Contrastive learning encourages learning feature representation with interclass separability and intraclass compactness, which can assist in classifier learning[3, 27]. More specifically, intraclass compactness refers to how closely image representations from the same class are related to one another, and interclass separability refers to how widely apart image representations are from different classes; this is due to SSL capability to learn without labels and therefore being able to leverage large datasets. Contrastive SSL has already been widely studied among both natural and medical image domains. There have been several comprehensive reviews on natural images, such as contrastive learning of visual representations[28], generative learning and contrastive learning[3], pretrained language models[29], and self-supervised contrastive learning[30]. However, these reviews did not focus on medical images that are different from natural ones with inherent medical image specific challenges and requirements. In addition, there were some SSL reviews on medical images[31, 32]. Some of them discussed three categories, including predictive, generative, and contrastive learning, but in the contrastive learning category, the authors did not divide it into subsections and provide structured portioning of the work. However, our paper exclusively focused on predictive and contrastive learning and used subsections to describe more details of the related backgrounds. In this study, we provide a state-of-the-art review of SSL research, focusing on predictive learning and contrastive SSL learning, and their adaptation and optimization for the medical imaging domain. With the focus of our paper on medical images, where possible, we have used medical images in our example figures. Our contributions are as follows: Section 2 introduces a systematic categorization of the state-of-the-art predictive learning and contrastive SSL methods and discusses their methodology; these methods are based on natural images. Section 3 presents a review of predictive learning and contrastive SSL methods applied to medical images and their unique adaptations from the natural image method counterparts. Section 4 concludes the review and dis-

cusses the limitations of predictive learning and contrastive SSL on medical images and makes suggestions for future research and directions.

2 Predictive learning and contrastive self-supervised learning

2.1 Predictive learning

By predicting geometric transformations of images, predictive learning tasks learn the structural and contextual semantics. Three types of spatially relevant position pretext tasks, as shown in Fig. 2(a), were described in this section: relative position, solving jigsaw puzzle, and rotation.

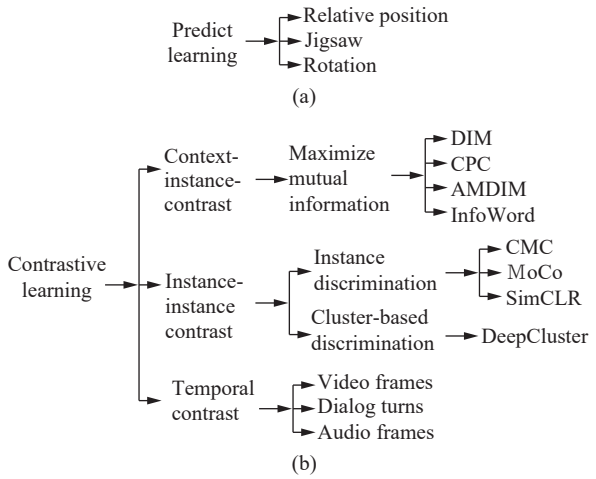

Fig. 2 Categorization of predictive learning and contrastive self-supervised learning: (a) Categorization of predictive learning; (b) Categorization of contrastive SSL.

2.1.1 Relative position

The relative position model[33] was trained to learn the relationships between a selected patch and the patches around it. The relative position model selected a particular size of the area of an image sample and divided this area into a certain number of disconnected patches. The number and the area in a patch, as shown in Fig. 3, were used to learn the relationship between the centre patch, called the anchor, and the neighbouring patches. As a result, the model learned the relationships between the patches. It is worth noting that the gaps between patches and the random displacement of patches prevent the model from learning the shortcut. Such a shortcut might be provided by low-level indications like boundary patterns or textures that continue between patches. There were three disadvantages with the relative position approach. First, multiple different objects could be included in two individual patches. For example, one patch contained the left atrium and another patch consisted of the right atrium. There was no relevance between these

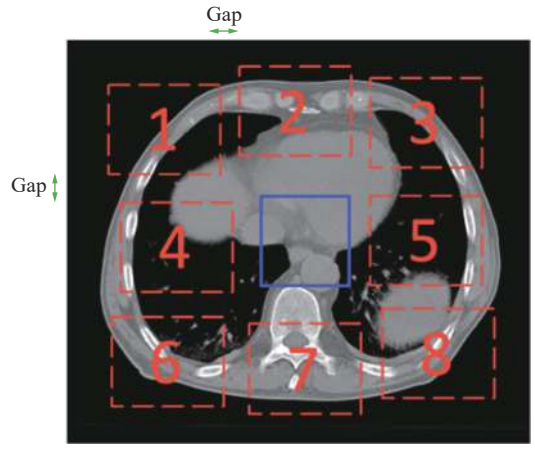

Fig. 3 An example of the predicting relative spatial position[33] pretext task on a CT lung image. The algorithm is trained to learn the relationships between a selected patch (blue centre) and the patches around it (red numbered patches).

two objects that are only located in the individual patches. As a result, no information could be learned about the relationship between those two objects. Second, in the relative position approach, CNNs could learn trivial features, such as the shared corners or edges of patches, instead of semantic feature representations that are beneficial to downstream discriminative tasks, including segmentation and classification tasks. Although some methods, such as randomly jittering patches, were designed to prevent the model from learning trivial features, there are possibilities that patch positions would be learned from other places, such as background patterns. Third, since the relative position approach only involves the patches, it did not include the global information of images. This led to limited performance on downstream tasks that rely on global information of images, such as in image classification tasks. However, some of these tasks counted on ad hoc heuristics that might restrict the transferability and generalization of learned feature representations for the following downstream tasks.

2.1.2 Solving the jigsaw puzzle

One additional type of relative position was termed as "solving the jigsaw puzzle"[34]. The principal idea of this pretext task was to learn positional relations among divided patches of an input sample. In this approach, by solving the jigsaw puzzles, the algorithm learned to recognize the elemental structure of the objects, including objects and their relative parts. As shown in Fig. 4, within an image sample, the jigsaw puzzle solution first selected a particular size of the area that was relevant to the topic of interest. Then, this area was divided into nine puzzle patches shuffled based on the nine predefined permutation sets as inputs. The model was trained to learn feature representation by correcting the order of those nine patches. The sequence of nine patches was used for the training model. The greatest challenge of the jigsaw puzzle was that the model required greater computation-

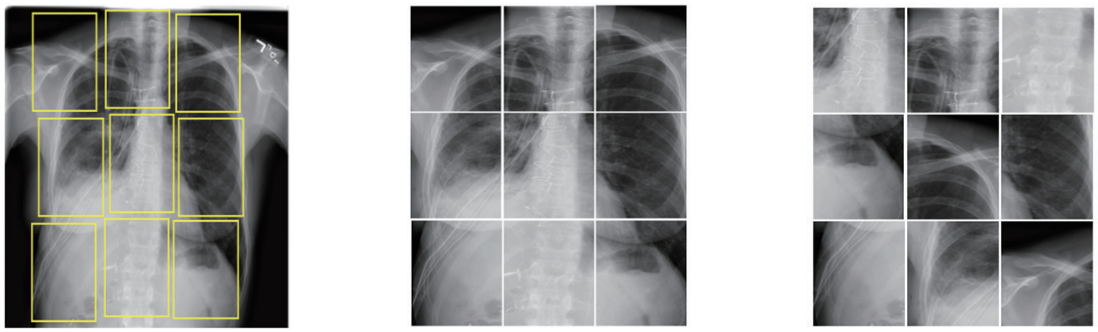

Fig. 4 An example of the "solving the jigsaw puzzle"[34] pretext task on an X-ray pneumothorax image. The algorithm is trained to learn the positional relations among nine divided patches (yellow-framed patches).

al complexity and memory consumption. Noroozi and Favaro[34] also extended this to more complicated pretext tasks, such as the setting of 64 predefined permutations, demonstrating that more information on relative position can be learned.

2.1.3 Rotation

Another context-based pretext task was designed for learning high-level semantic features by training the model to predict the degrees to which the input images were rotated. The rotation angle could be seen as a pseudo label for training the model. This is exemplified in Fig. 5. The results of [35] showed that CT lung images rotated by angles of 0, 90, 180 or 270 degrees learn better feature representations than those rotated by other degrees. Li et al.[36] also conducted research based on the rotation pretext task in which the angle was an expansion to 360 degrees. Lee et al.[37] trained the model with multiple pretext task learning strategies, including two types of transformations, rotations, and color permutation, as those various self-supervised data augmentations enabled the reduction of the effects from the transformation invariant.

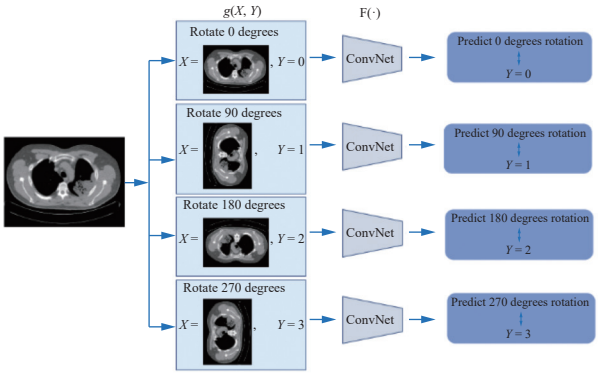

Fig. 5 An example of the predicting image rotations[35] pretext task on a CT lung image. The algorithm utilizes the rotation angle as a kind of supervision for training the model.

2.2 Contrastive self-supervised learning

Contrastive learning is a method to learn feature representations via contrastive loss functions to distinguish between negative and positive image samples. Positive image samples are an augmentation of a target image (also called an anchor), while negative image samples are from other nontarget samples within the training set. The contrastive learning approach encourages models to learn general-purpose feature representations that can be reused to enhance learning specifically in downstream tasks, e.g., segmentation and classification tasks, where the models are built using the learned features[38].

Contrastive learning methods typically vary in how they use unlabelled data to create or define negative and positive image pairs and in how they are sampled during training. Based on the idea of Liu et al.[3], contrastive learning categories are divided into two subcategories: context-instance contrast and instance-instance contrast. The context-instance contrast, also known as the global-local contrast, is concerned with modelling the relationship between a sample's local feature and its global context representation. Instance-instance contrast investigates the connections between the instance-level local representations of distinct samples. However, these two categories do not cater to the specific needs of sequential image or time series datasets. Any data that have elements that are arranged in sequences are referred to as sequential data[39]. Sequences of user actions, time series, and DNA sequences are a few examples. Yue et al.[40] mentioned that time-series medical images include rich spatial and temporal information. Therefore, we suggest a third category named temporal contrast, which is related to SSL designed for sequential datasets. The three categorizations of contrastive SSL are shown in Fig. 2(b).

To train on unlabelled data, SSL uses "pretext" tasks as an alternative way to extract useful latent representations. By solving the pretext tasks, pseudo labels, as supervisory signals, are generated automatically based on the dataset's properties. For example, with the rotation pretext task, the supervisory signals of "rotation angles" are derived from the unlabelled input samples. There are two different application paradigms for downstream tasks using the pretext task results. Fig. 6(a) shows that the first paradigm is learning transferable features. After solving the pretext tasks, the model will try to learn feature representation, which can then be further trained, e.g.,

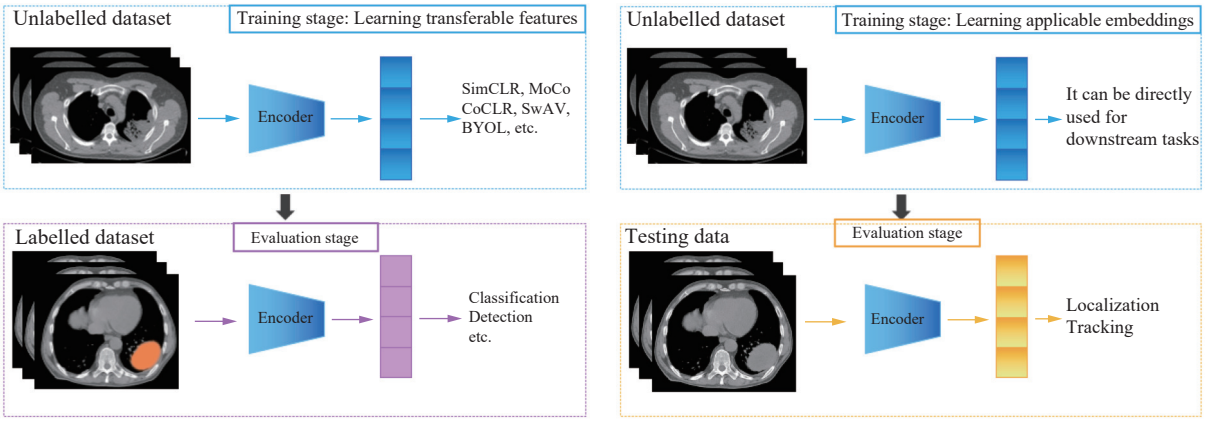

Fig. 6 Two different application paradigms for downstream tasks. In (a), further training such as fine-tuning is needed; while in (b), no annotation is needed for downstream tasks.

fine-tuning for different tasks such as classification and detection. In contrast, Fig. 6(b) illustrates an example of learning "applicable embeddings" that refers to the pretext tasks used to directly learn generalizable features for downstream tasks.

Various pretext tasks are designed with those different-augmentation transformations to capture the expected semantic or structural characteristics of images for downstream tasks. Before diving into subcategories, the contrastive learning loss function is defined in Section 2.2.1 for a fundamental understanding of SSL. Then, context-instance contrast learning and instance-instance contrast learning are described in Sections 2.2.2 and 2.2.3, respectively. Finally, temporal contrast is introduced in Section 2.4.

2.2.1 Contrastive learning

To learn meaningful features from the images, SSL uses "data augmentation" techniques to generate additional data by increasing the diversity of the data transformation. Data augmentation involved image manipulation techniques, i.e., image scaling, cropping, flipping, padding, rotation, translation, and color augmentation, such as brightness, contrast, saturation, and hue. The fundamental concept of contrastive learning was to group the images with their augmentations closer together and place the other images further away. This description can be expressed as

$$score\left(f\left(x\right), f\left(x^{+}\right)\right) > score(f\left(x\right), f(x^{-})) \qquad (1)$$

where $f(x)$ is an encoder. The target image (also called an anchor), x, and the anchor's augmented sample, x^{+}, can be grouped as a positive pair. However, the anchor and other sample from the training dataset, x^{-}, are grouped as a negative pair. As a result, the score of the similar sample, x and x^{+}, is higher than that of the dissimilar samples, x and x^{-}. This score is a metric that compares the similarity between the two samples. Based on this concept, the following subsections discuss several common loss functions used in SSL.

Triplet loss

Triplet loss[41] is a type of metric learning with a similar concept to (1), with changes in how it calculates the distance on the embedding space. In detail, minimizing the triplet loss, as in (2), encourages the distance between the anchor and the positive sample to 0; and the distance between the anchor and the negative sample to be greater than the distance between the anchor and the positive sample plus with margin. When the representations created for a negative pair are distant enough, the purpose of the margin is to prevent effort wasted on enlarging this distance.

$$\mathcal{L} = \max\left(d(x, x^{+}) - d(x, x^{-}) + margin, 0\right) \qquad (2)$$

where $d\left(x, x^{+}\right)$ denotes the distance between the anchor and the positive sample, and $d\left(x, x^{-}\right)$ represents the distance between the anchor and the negative sample. The margin parameter is set to represent the minimum offset between the distances of the pairs.

Noise-contrastive estimation (NCE) loss[42] and InfoNCE loss[43]

To decrease the complexity of optimization, NCE was introduced to transform the calculation from multiclass classification problems to a binary logistic regression to classify data from noises. Inspired by NCE loss, InfoNCE loss used categorical cross-entropy loss to find positive samples from a collection of unrelated noisy samples. InfoNCE used a similar data pattern for training, including one positive sample and many negative samples. However, InfoNCE loss often generated higher accuracies due to the selection of negative samples. This was explained by the grouping of the negative samples in the NCE algorithm as a unit for calculating an approximate value, while InfoNCE calculated the negative samples as an individual sample and hence can keep more information about each of the data points. InfoNCE is formulated as

$$L_N^{InfoNCE} = -\frac{E}{X}\left[\log\frac{f_k\left(x_{t+k}, C_t\right)}{\sum_{x_j \in X} f_k\left(x_j, C_t\right)}\right] \tag{3}$$

where $f_k\left(x_{t+k}, C_t\right)$ represents the density ratio, $t+k$ denotes the future time steps after t from the dataset, $\{x_{t-k}, \cdots, x_{t-1}, x_t, \cdots, x_{t+k}\} \in X$, where $f_k\left(x_{t+k}, C_t\right)$ and $f_k\left(x_j, C_t\right)$ can be seen as the positive sample pair and negative sample pair, respectively, in the collection of samples, C_t.

Mutual information (MI)

Mutual information[44] is a concept of reducing uncertainty about one random sample after observing another sample. Simply put, MI is a measure for assessing the relationship between arbitrary variables[45]. For example, Linsker[46] presented the InfoMax principle by using MI to calculate the relationship between the input and the output in the presence of processing noise. The relationship between InfoNCE and MI has been used in many state-of-art contrastive learning methods, and after optimizing (3), it can be expressed as

$$I\left(x_{t+k}, C_t\right) \geq \log\left(N\right) - \mathcal{L}_N^{opt} \tag{4}$$

where $I(x_{t+k}, C_t)$ is equal to or larger than $\log\left(N\right)$, and N is the number of samples, minus the optimized InfoNCE, $\mathcal{L}_N^{opt}$.

2.2.2 Context-instance contrast learning

Spatial context from images could be used to learn feature representations. It was originally from the concept of the skip-gram Word2Vec[47] algorithm used in natural language processing (NLP) and later implemented for images by Doersch et al.[33] With spatial context, feature representations were learned by predicting the position of an image patch relative to other patches. The context-instance contrast learned the relationship between local and global image features. The idea of context-instance contrast was to capture the local features that can adequately represent the global features. In this category, the most popular algorithm is maximizing MI.

Maximizing MI

Unsupervised learning of feature representations could be achieved by maximizing the MI between an input image and the output encoded by a deep neural network. The principle of high MI captures useful information rather than low-level noise. Tschannen et al.[44] conducted research on MI maximization for unsupervised or self-supervised representation learning, including deep InfoMax (DIM)[48], contrastive multiview coding (CMC)[49], and contrastive predictive coding (CPC)[43].

Deep InfoMax (DIM)[48] and augmented multiscale DIM (AMDIM).[50] Hjelm et al.[48] showed that, depending on the downstream task, it is often insufficient to learn effective representations by maximizing the MI between the encoder output (i.e., global MI) and the entire input. This was because global MI maximizes MI between global representation pairs, which included an entire image together with a single feature vector summarized from patches divided from the results of encoding input images. However, global Infomax has the problem that the model captured undesirable information such as trivial noise that was particular to local patches or pixels and that was useless for certain tasks such as image classification. This was because grabbing feature information particular to only belonging parts of the input through encoders did not enhance the MI with other patches that did not include those trivial noise. Hence, this issue arose the idea of local Infomax to encourage the encoders to learn feature representation that is shared across the patches of an input image. Hjelm et al.[48] showed that adding location information of the input into the object enables a considerable increase in a representation's fitness for subsequent tasks. Hence, they proposed the ideas of global DIM and local DIM to train the encoders by maximizing MI between global and local patch features. Local infomax maximizes MI between the summarized patch feature vector and each local patch feature, where both are extracted from different layers of the same structure of the convolutional network. Later, Bachman et al.[50] extended the idea of local DIM by maximizing MI between features generated through augmentation of each input image. The author improved the local DIM from the following three perspectives: data augmentation, multiscale mutual information, and encoder. For data augmentation, they first performed a random horizontal flip and then some common data augmentations, including random in the crop, jitter in color space, and grayscale transformation. This model learned features by maximizing MI between the global and augmented local features. To determine the similar part in augmented local features and global features. For multiscale mutual information, the model learned features by maximizing MI within features from different layers with different scales. The MI between multiscale features in the same images was higher than that in different images. For the encoder, AMDIM improved the encoder based on the ResNet-based framework to control receptive files. The result was worse when there was too much overlap within the features of positive sample pairs.

Contrastive predictive coding (CPC).[43, 51] Contrastive predictive coding[52, 53] focused on sequential data and utilizes useful information of previous sequential components of the data to predict the future sequential signal. During the predictive coding, the information of image content was embedded. Using autoregressive models, CPC encoded key shared information within different parts of the previous sequential signal to high-level latent space, and this was used to predict the future that conditionally relies on the same shared information. This resulted in keeping a similar representation from the same images encoding more global and common features, and

discarding low-level information and local variations, such as noise. Additionally, the use of probabilistic contrastive loss for learning high-dimensional representations in latent embedding space maximized useful information for predicting future samples. Based on the ideas of NCE, CPC proposed InfoNCE and its relationship with MI. That is, minimizing the InfoNCE loss enabled maximizing a lower bound on MI between representations that were encoded.

2.2.3 Instance-instance contrast learning

Under the instance contrast learning[54] category, instance comparisons were used from two points of view. The first was to design or modify contrastive loss functions and use specific structures for training SSL (see Section 2.3.1). The second was to directly compare instances to derive distinctive information within the instances (see Sections 2.3.2 and 2.3.3).

SSL design on contrastive loss function-based variation and specific structures

Within many strategies of designing SSL models, we discuss two ideas based on either the varied contrastive loss functions or the specified structure in the subcategories.

SSL design on contrastive loss function-based variation. When contrastive loss functions are designed or modified based on the principle of (1), they had been applied to many different tasks for specified learning approaches. The five learning approaches introduced in this section are 1) multimodal learning, 2) local representation learning, 3) multiscale learning, 4) texture representation learning, and 5) structural representation learning.

1) For multimodal learning, most papers conducted SSL research on only one modality dataset. Hence, some studies have started working on multimodal SSL training to learn more meaningful semantic information that might compensate for each other. For computer vision, multimodality could group different types of resources, such as text and images, or different types of data formats, such as CT, X-ray, and MRI. 2) For local representation learning, most of the common instance-instance contrast learning methods concentrated only on extracting image-level global consistency between instances but neglect explicitly learning the distinctive local consistency within the instances. Distinctive local representations played a vital role in obtaining structural information for dense or per-pixel prediction tasks, including segmentation. 3) For multiscale learning, some medical data were large, such as histology images. Such large images as input for training the network slowed down the calculation and increased the training time. Hence, for the domain of histopathology, some studies used relatively small areas or objects, such as nuclei, to predict whole histology images. However, some works utilized a variety of sizes of input for the training model and Yoo et al.[55] demonstrated how multiscale local activations could enhance visual representation based on CNN activations. Finally, some SSL works designed the contrastive loss for learning 4) texture representation and 5) structural representation, respectively.

SSL design on specific structures. Except for the design and modification of contrastive loss functions and the selected sample strategies, some works focused on the specific structures for training SSL, such as Siamese-based learning, and teacher-student-based learning. For Siamese network learning, a Siamese neural network included two or more identical subnetworks that were used to estimate the similarity between two samples by two feature extractors with shared weights, and were utilized in many applications, such as the prediction of camera poses[56] and lip poses[57]. A large number of batch sizes or negative pairs applied in common SSL methods made them more difficultly be implemented on 3D medical datasets. Chen and He[58] proved that the Siamese network could be used to avoid such problems on a 2D network. In addition, without relying on larger batch sizes or negative pairs, the Siamese network enabled model to keep the spatial relationship in the embedding space through contrastive loss. For the teacher-student-based learning, teacher-student learning was a transfer learning approach in which the student network was taught by the teacher′s network to predict the same result as the teacher′s. A small network, the student network, could be learned by the labels produced by a complex model, the teacher network. Moreover, the mean teacher model, an extended model based on the teacher-student, was implemented for the medical image analysis tasks to average model weights to aggregate information after every step instead of every epoch. The mean teacher model also provided more robust intermediate representations since the weight averages captured all layer outputs, not just the top output.

Instance-based discrimination

There were a variety of techniques designed for collecting negative samples to compare with a positive sample in the training process, such as memory bank, momentum encoder pretext-invariant representation learning (PIRL)[23], simple framework for contrastive learning of visual representations (SimCLR)[20], momentum encoder and momentum contrast (MoCo)[17, 59, 60], and bootstrap your own latent (BYOL)[22]. Although for different purposes, these methods could be considered to create dynamic dictionaries. In these dictionaries, the "queries" and "keys" were obtained from data, e.g., patches or images, which were embedding representations created through the query and key encoder networks, respectively. These encoders could be any CNNs[61]. SSL trained encoders to execute dictionary look-up: An encoded "query" should be comparable to its corresponding key while being distinct from others. The definition of query and key could be different. For example, Wu et al.[54] grouped a key and a query as a negative pair if they came

from a different image and otherwise as a positive sample pair. However, Ye et al.[62] selected two random "views" of the same image using random data augmentation to create a positive pair. It is worth noting that inconsistency was a major challenge in this method. Inconsistency existed between the query and key embedding representation. Specifically, inconsistency occurred when calculating the contrastive loss between the positive features from the query encoder that was updated each epoch and the negative features saved in the memory that was updated from several previous epochs. Hence, many approaches were proposed to solve this inconsistency. He et al.[19] hypothesized that it was possible to create consistent and large dictionaries during the training process and that in the dictionary, the keys should be represented through the similar or same encoder to provide consistency in comparisons to the query.

Memory bank and MoCo.[19] Based on the principle of contrastive loss, the number of negative samples significantly affected the accuracy, which was proven by Nozawa and Sato.[63] In one batch, it included an original image, its augmented example, and many negative samples. The number of negatives sampled depended on the batch size, and a large batch size means that we could contain more negative samples. However, the batch size was limited by the GPU memory size. The memory bank was designed to address this problem by accumulating and regularly updating many embeddings of negative samples that resulted from the key encoder without increasing the batch size but with less gradient calculation from the encoded key query during training. PIRL learned invariant representations by using a memory bank based on a pretext task related to solving the jigsaw puzzle.

Although memory banks could contain a larger number of negative samples, inconsistency existed between the query and key embedding representations that resulted from the query and key encoders, respectively. To address the inconsistency problem, MoCo decoupled the batch size from the negative samples by replacing the memory bank with a moving-averaged encoder called the momentum encoder. This momentum encoder was built as a dictionary-like queue that progressively replaced samples by enqueuing the current mini-batch and dequeuing the oldest mini-batch in this queue. The benefit of removing the oldest mini-batch that was outdated was to maintain consistency with the newest samples from the query encoder. By doing this, the number of negative samples could be increased without expanding the batch size. In brief, MoCo decreased the dependency on mini-batch size and utilized a momentum encoder to update the queue that involves previously processed samples to create contrastive pair encodings. This was defined as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{5}$$

where the momentum coefficient, m, made the key encoder, θ_k, slowly progress, driven by the query encoder, θ_q, $(1-m)$. He et al.[19] proved that the performance was the best when m was 0.99 because this setting updated the key encoder slowly through a large part of the previous key encoders and a small part of the newest query encoder. This could keep a large and consistent dictionary that facilitates contrastive learning to train a visual representation encoder. Based on MoCo, the same team further designed MoCo v2[17] by adding a multilayer perceptron (MLP) projection head, data augmentation, and a cosine learning rate schedule.

SimCLR.[20] SimCLR was an end-to-end learning architecture and learned feature representations by maximizing the agreement between dissimilar augmented views of the same input via a contrastive loss calculation[64]. Through experiments, the results of SimCLR showed four components that affect the quality of contrastive representation learning. The combination of data augmentation, random cropping, and color distortion was shown to be better than other combinations or single transformations. Moreover, compared to supervised contrastive learning, unsupervised contrastive learning obtained greater advantages from longer training, larger batch sizes, and stronger data augmentation. However, similar to supervised learning, contrastive learning obtained an advantage from a deeper and wider framework. It is worth noting that the introduction of the nonlinear projection head significantly improved the learning representations during training. Based on SimCLR, the same team further improved three steps for designing a semisupervised learning framework called SimCLR v2[65].

Contrastive multiview coding (CMC).[49] Unlike DIM, CPC, and AMDIM, which used one view of the image, CMC worked on images that were acquired in more than one view. The goal of CMC was to learn feature representations with information shared between various sensory channels obtained from the same image. Specifically, CMC used NCE-based softmax cross-entropy loss to learn feature embeddings by maximizing MI between various views from the same scene. A 4-view dataset, RGB and depth (RGBD)[66], from the same scene, was brought together in embedding space as positive samples, but the views from different scenes were pushed apart as the negative sample. CMC also proposed "core view" and "full graph" paradigms. The full graph outperforms not only because more cross-view learning could obtain better representation but also because the full graph can deal with missing information of views.

Bootstrap your own latent (BYOL).[22] Some contrastive learning methods in Section 2.3.2, such as SimCLR and MoCo, relied heavily on many negative samples for learning the discriminative features. Hence, those methods were sensitive to selecting data augmentation policies and require many trials to determine good data augment-

ation[67, 68]. Moreover, SimCLR required a long training time on large datasets, out of 3 200 epochs on the 1.2 million ImageNet images[69], to obtain improved performance. Unlike SimCLR, BYOL used mean squared error (MSE) rather than a contrastive loss, so as to rely less on the availability of large-scale negative samples.

Cluster-based discrimination

In computer vision, clustering algorithms were a class of unsupervised learning techniques that have been largely researched and applied. Although clustering techniques were the first stage of success in classifying images, relatively few papers introduced applying them to CNNs end-to-end training on large scale datasets[70, 71]. A problem is that clustering techniques were primarily built on linear models for calculating the top of fixed features, and they seldom function when the features must be simultaneously learned.

Based on the clustering technique, DeepCluster was designed to simultaneously learn the features′ cluster assignments and the neural network′s parameters. More specifically, they iteratively clustered the features with a normal clustering algorithm, *k*-means, and utilized the cluster assignments as supervision signals to learn the parameters of the network. Unlike context instance contrast, clustering had the benefits of needing little domain knowledge and no particular signal from the inputs. In addition, some contrastive learning methods highly depended on the online calculation of many pairwise feature comparisons. Hence, the authors of swapping assignments between multiple views (SwAV)[72] designed an online algorithm with a cluster-based idea to reduce the amount of computation. SwAV employed a "swapped" prediction technique in which the cluster assignments of one view were predicted based on the representation of another view. This method could work in large and small batch sizes without needing a momentum encoder or a large memory bank. A multicrop technique was designed by making use of smaller-sized images to boost views without raising a training′s memory or processing demands.

2.2.4 Temporal contrast

Medical imaging datasets, of CT or MRI, sometimes have follow-up scans with spatial or structural information. A sequence of CT or MRI, such as from left to right or from top to bottom of the patient′s body, assists in learning more semantic representations. Compared to 2D data, videos or image sequences have richer information that allows better feature representation to be learned through SSL. There are three common types of 3D SSL, including finding the similarity of adjacent frames, tracking the objects, and correcting the temporal order.

Finding similarities of adjacent frames

First, adjacent frames should share similar features[73]. By training CNNs to learn the similarities within neighborhood frames, contextual semantic representations could be learned. Moreover, temporal continuity[74] in sports activities, such as playing table tennis, and the characteristic of frames expressing a swing action should also be smooth. In this case, in the same sequence, the adjacent frames selected within a small design range were closer in embedding space than, frames selected from distant timesteps, as shown in Fig. 7. In addition to learning from the same video, Sermanet et al.[75] also learned from multiview (multiple modalities) videos to obtain viewpoint and agent invariant feature representations. In this case, positive paired images obtained simultaneously with different viewpoints were closer in the embedding space than negative paired images obtained from a dissimilar time in the same sequence.

Tracking the objects

Second, based on visual tracking-provided supervision for training models, Wang and Gupta.[76] learned visual representations by unsupervised tracking within thousands of unlabelled moving videos. More specifically, two frames connected by a track should share a visual representation in feature space, such as cycling, because they probably corresponded to the same target of the moving object or were part of the object. Based on this idea, Walker et al.[77] utilized CNNs to learn similar objects that shared similar visual representations, and Purushwalkam et al.[78, 79] researched human poses. In this case[76], designed a ranking loss function to encourage, in feature space, the first two frames connected through a track to be much

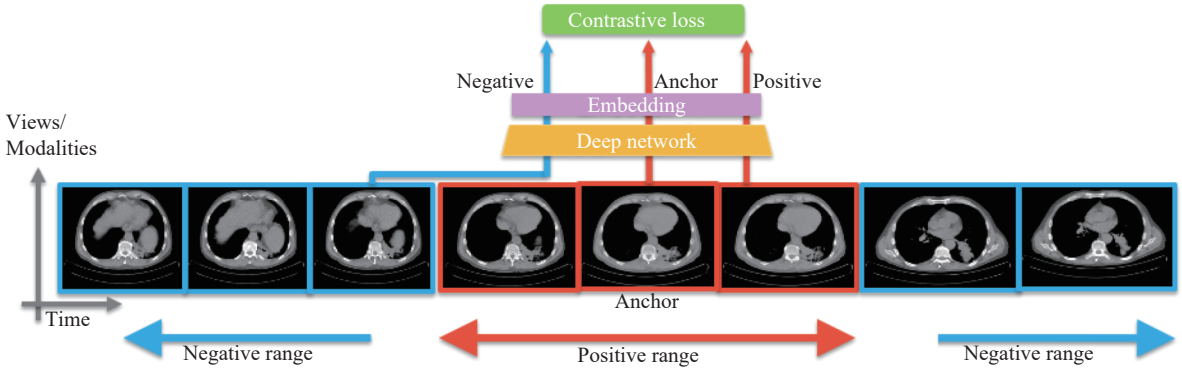

Fig. 7 Selection of positive samples and negative samples from a set of adjacent frames

closer than the first frame and a random frame.

Correcting for the temporal order

Third, it was a method to learn visual representation through an unsupervised sequential verification task, which corrected frame order from a sequence of video frames[80–82]. In this case, the correct order was a positive sample, and the wrong order was a negative sample, as shown in Fig. 8.

3 Predictive contrastive SSL applied to medical images

Contrastive SSL has been broadly applied and optimized for medical images. Four forms of contrastive SSL were commonly applied to medical images: contrastive learning estimation, context-instance contrast learning, instance-instance contrast learning and temporal contrast SSL.

3.1 Predicting learning for medical image analysis

3.1.1 Relative position

SSL based on the relative position approach was also used in the medical area[83] for learning useful semantic features by utilizing image context restoration. Architecture with the combination of multiple SSL methods was used, including relative position prediction[33], colorization[84], exemplar CNNs[85], and inpainting[86]. In particular, the relative position was used to find the relationship between the central patch and eight nearby patches with in a selected 3×3 patch grid. Inspired by the work of context prediction of adjacent patches[33], Blendowski et al.[87] proposed self-supervised 3D context feature learning, which included a new idea of image-intrinsic spatial offset relations with a heatmap regression loss. Jana et al.[88] used image context restoration[83] as the pretext task for checking nonalcoholic fatty liver disease that led to granular textural changes in the liver and could progress to liver cancer. Since one of the signs of nonalcoholic fatty liver disease was texture change in the liver, Jana et al.[88] encouraged the network to learn neighboring pixel information for downstream tasks, including fibrosis and NAS score prediction. Based on [33], Li et al.[89] analysed the issue of COVID-19 severity assessment by training the SSL model to predict the relative location between two patches of the same CT slice. Fashi[90] utilized the primary site information as pseudo-labels and modified the histopathology patch order for the training feature extractor. The added supervised contrastive learning loss boosted more robust feature representations for WSI classification.

3.1.2 Solving jigsaw puzzles

Based on solving jigsaw puzzles, SSL was applied to learn useful semantic features by blending patches from various medical imaging modalities[91]. This multimodal jigsaw puzzle task first drew random puzzle patches from dissimilar medical imaging modalities and combined them into the same puzzle. Combining these medical imaging modalities at the data level encouraged the model to derive modality-agnostic representations of the images and derive modality-invariant views of the objects, including tissues and organ structures. The learned feature representations from many medical imaging modalities could contain cross-modal information, which combined complementary information across the modalities. Taleb et al.[91] augmented multimodal data using cross-modal generation techniques to address modality imbalance problems in real-world clinical situations. In addition, their two modality experiments showed that the proposed multimodal puzzles learn powerful representations, even when the modalities were nonregistered. One was on prostate segmentation of two MRI modalities, and the other was on liver segmentation of both CT and MRI modalities.

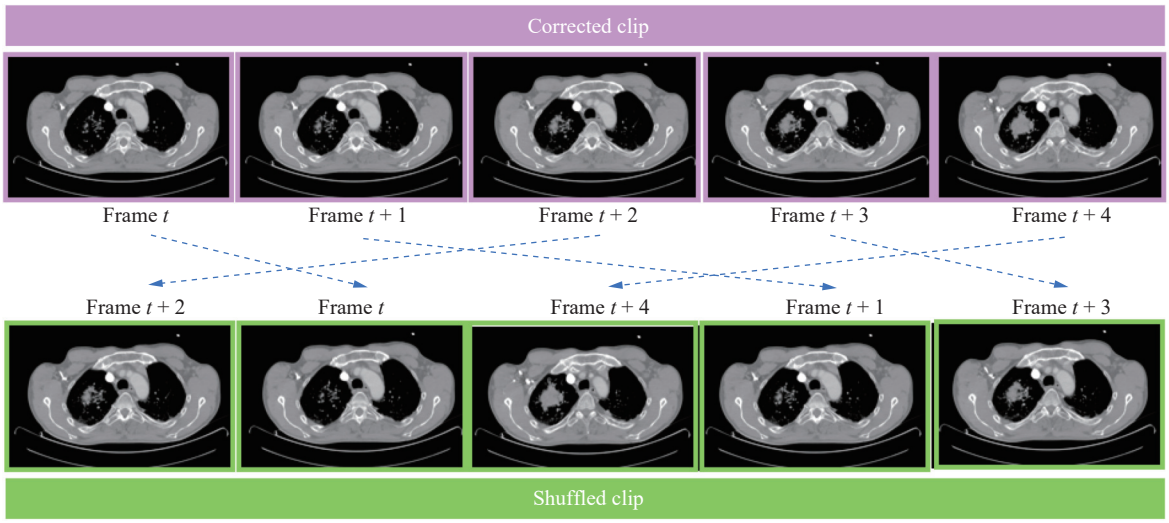

Fig. 8 Positive slice examples (correct order) and the negative slice examples (incorrect order) from a video are trained to learn the semantic representations.

By increasing performance on downstream tasks and data efficiency, it summed up that the multimodal jigsaw puzzle created better semantic representations when comparing the performance on each modality independently. Later, the same team proposed multimodal puzzle solving as a proxy task to assist feature representation learning from multiple image modalities[92]. Navarro et al.[93] compared and assessed the robustness and generalizability of both SSL and fully supervised learning networks on downstream tasks, including pneumonia detection in X-ray images and segmentation of various organs in CT images. By solving jigsaw puzzles on those medical datasets, they summarized that they efficiently learned feature mapping of object parts and their spatial arrangement through SSL. Based on the idea of a jigsaw puzzle-solving strategy, Manna et al.[94] learned spatial context-invariant features from magnetic resonance video clips to check knee medical conditions. They mentioned that the first work applied SSL to class imbalanced multilabel magnetic resonance (MR) video datasets. Based on the jigsaw puzzles transformation[34], Li et al.[95] designed a self-supervised network by modifying two processes. The first was to increase the variety of permutations, and the second was to merge the jigsaw puzzles pretext task into the end-to-end semisupervised framework. They applied the proposed semisupervised learning method to two medical image segmentation tasks, including nuclei[96–98] segmentation and skin lesion[99–101] segmentation. To classify cervix images as normal against cancerous, Chae et al.[102] presented a new SSL patch based on puzzle pretext tasks to predict the relative position. They found that the pivotal area of the image to search for cervix cancer was highly potential around the centre and the irrelevant parts were near the periphery. In the domain of histopathology, based on the relative patch algorithm, Santilli et al.[103] implemented domain adaptation from the skin to breast spectra because of the low-level resemblance in the outline between skin tissue and breast cancer. They applied a relative patch pretext task for training on skin data to learn positional relations among divided patches of an input sample and then transferred the learned weights to the following downstream task, breast cancer classification. Zhuang et al.[104, 105] and Tao et al.[106], inspired by the jigsaw puzzle, proposed a novel 3D proxy task by playing a Rubik′s cube, called Rubik′s cube recovery. Since the jigsaw puzzle was designed for 2D data, Rubik′s cube recovery was introduced for 3D volumetric data. During Rubik′s cube recovery process, rich feature information from 3D medical images was obtained, including cube rearrangement and cube rotation. This enforced the model to learn the features invariant from both translational and rotational perspectives. It is worth noting that the difficulty increased when the cube rotation operation was added to Rubik′s cube recovery, as it encouraged networks to exploit more spatial information. Li et al.[107] extended Rubik′s cube by adding a random masking operation to obtain feature representations from COVID-19 and negative CT volumes.

3.1.3 Rotation

Li et al.[108] observed that each fundus image included obvious structures, such as the optic disc and blood vessels, that were sensitive to orientations. Hence, they proposed a rotation-oriented collaborative approach to learning complementary information, including rotation-related and rotation-invariance features. With these two pretext tasks, vessel structures in fundus images and discriminative features for retinal disease diagnosis were learned. In addition to the rotation pretext task, Yang et al.[109] applied elastic transformation prediction[110], to cross-modality liver segmentation from CT to MR. Inspired by [35, 111, 112], Liu et al.[113] presented SSL based on a 3D feature pyramid network for assisting multiscale pulmonary nodule detection. Dong et al.[114] classified focal liver lesions by utilizing several relative position pretext tasks, such as predicting the relative position between patches of an input, predicting the rotation, or solving a jigsaw puzzle. Imran et al.[115] presented a new semisupervised multiple-task model utilizing self-supervision and adversarial training to classify and segment anatomical structures on spine X-ray images. Several pretext tasks were used several SSL simultaneously for medical imaging analysis, such as the studies that worked on the combination of rotation prediction[35] and jigsaw puzzle assembly[34]. However, Tajbakhsh et al.[116] combined two different types of SSL, such as rotation (contrastive SSL) and reconstruction[117] and colorization[118] (generative SSL), on retinal images for diabetic retinopathy classification. In histopathology, Koohbanani et al.[119] utilized and combined various self-supervised tasks for domain-specific and domain agnostic purposes to obtain contextual, multiresolution, and semantic features in pathology images. Vats et al.[120] adopted those two pretext tasks for wireless capsule endoscopy diagnosis.

3.2 Contrastive learning estimation for medical image analysis

To focus on abnormalities, Liu et al.[121] introduced a learnable alignment module into contrastive learning to alter all input samples to be geometrically canonical. More specifically, after extracting high-level feature representations of the image pair, the highly structured character of inputs was used to calculate the L1 distance between corresponding pixels on the positive and negative images. The result could be seen as an indication of possible lesion location on the latter. Their model could alleviate the difference in scales, angles, and displacements of X-ray samples created under bad scan conditions. They demonstrated that the learned features represent localization information that enabled better identification and localization of downstream tasks, including

infiltration, mass and pneumothorax diagnosis.

3.2.1 Contrastive learning

Triplet loss for medical application

Xie et al.[122] proposed a novel SSL framework with scalewise triplet loss and count ranking loss to encourage a neural network to automatically learn the information of nuclei quantity and size from the raw data for nuclei segmentation.

Noise-contrastive estimation loss[42] and InfoNCE[43] for medical image analysis

Sun et al.[123] presented a context-aware self-supervised representation learning approach for learning anatomy-specific and subject-specific representations at the patch and graph levels, respectively. Interestingly, they utilized InfoNCE loss to learn patch-level textural features and contrastive learning objectives for learning graph-level representation. They also took advantage of MoCo, including a queue of data samples and a momentum update scheme to enhance the number of negative samples during training. The features learned through the proposed method demonstrated better performance in staging lung tissue abnormalities associated with COVID-19 than those learned by other unsupervised baselines, such as MedicalNet, Models Genesis, and MoCo. Most existing methods that used the maximization of MI as contrastive loss utilized image pairs for training; however, Zhang et al.[124] made use of image-text pairs. Their work enhanced the visual representation learning of medical images by taking advantage of the combined information from textual data and image pairs. Through a bidirectional contrastive objective loss between those two different modalities, this approach dependeds on maximizing the agreement between real medical representation image-text pairs and randomly chosen pairs. More specifically, bidirectional contrastive objective losses were utilized similarly to the InfoNCE loss. Minimizing this loss encourages encoders to reserve the MI between real representation image-text pairs. Punn and Agarwal[125] utilized the Barlow twins (BT) framework to pretrain an encoder through redundancy reduction, similar to the InfoNCE objective, to learn feature representation over four biomedical imaging segmentation tasks, including cell nuclei, breast tumour, skin lesion, and brain tumour. Except for InfoNCE-based contrastive loss based on the MoCo framework, Kaku et al.[126] added two additional losses, mean squared error (MSE) and BT. By minimizing the MSE of feature representations between the intermediate layer or using BT to make their cross-correlation matrix closer to an identity matrix, the model was encouraged to learn augmentation-invariant feature representations that were not only focused on the final layer of the encoder but also extracting the intermediate layers. Their results showed that performance was better than MoCo on three medical datasets, including breast cancer histopathology, Clinical Center of the National Institute of Health (NIH) chest X-ray and diabetic retinopathy. Taher et al.[127] found that instance-based objectives learned the most discriminative global feature representations, which might not be sufficient to discriminate medical images. Hence, inspired by the integration of generative and discriminative approaches, preservational contrastive representation learning (PCRL)[128], Taher et al.[127] developed an SSL framework, context-aware instance discrimination, to encourage instance discrimination learning with context-aware feature representations.

3.2.2 Context-instance contrast learning for medical image analysis

Maximizing MI for medical image analysis

Deep infomax (DIM)[48] and augmented multiscale DIM (AMDIM).[50] Chen et al.[129] combined two different types of self-supervised methods, one from the context-instance category, DIM, and the other from the instance-instance category, SimCLR[20], for learning disease concept embedding. They utilized the proposed model to extract medical information from electronic health records and disease retrieval.

Contrastive predictive coding (CPC).[43] Stacke et al.[130] implemented and evaluated CPC on histopathology. After experimenting with some model and data-specific parameters on CPC models on histopathology images, those models were estimated for linear tumour classification on three tissue types. This work summarized the restriction of the learned representation for linear tumour classification on histopathology images because only low-level features in the first CPC layers were used. The diversity of distribution of the histology dataset makes little difference for linear tumour classification on histopathology images. Taleb et al.[131] extended this idea to a 3D CPC version. Instead of the time sequence dataset used in CPC, 3D CPC utilized a feature representation set obtained from patches cropped from the upper or left part of the 2D image sample to predict the encoded feature representations of the remaining part, lower or right part. In addition, they also developed a 3D version for rotation prediction, relative patch location, jigsaw puzzles, and exemplar networks. They demonstrated that the feature representations learned from 3D models were more accurate and efficient for solving downstream tasks than training the models from scratch and pretraining them on 2D slices. Zhu et al.[132] investigated the feature complementarity within multiple SSL approaches and presented a greedy algorithm to add multiple proxy tasks. More specifically, based on the assumption that a weaker correlation indicated a higher complementarity between two features, they calculated the correlation measure between the features created by different proxy tasks and then utilized the greedy algorithm to iteratively include a proxy task in the current task pool to form a multitask SSL framework. They applied it to the 3D medical volume brain haemorrhage dataset by adding multiple proxy tasks, including 3D rotation, Models Genesis[133],

3D CPC, and Rubik's cube. After locating the potential lesions through supervoxel estimation utilizing simple linear iterative clustering, Zhu et al.[134] calibrated CPC to learn a 3D visual representation. More specifically, calibrating the CPC scheme on the subvolumes cropped from supervoxels embedded the rich contextual lesion information into 3D neural networks. Cerebral haemorrhage classification and benign and malignant nodule classification were implemented using the proposed method on the brain haemorrhage and lung cancer datasets, respectively.

3.2.3 Instance-instance contrastive learning for medical image analysis

SSL design on contrastive loss function-based variation and specific structures for medical image analysis

SSL design on contrastive loss function-based variation. Based on the principle of the contrastive learning loss function, some papers worked on selecting positive and negative samples. For example, Jian et al.[135] combined a multilayer network and VGG-16 to discriminate images with helicobacter pylori infection from images without helicobacter pylori infection well. However, some papers modified the principle of the contrastive learning loss function for particular applications, such as the following five applications. 1) Learning multimodality for medical applications – Holmberg et al.[136] proposed a new large-scale and cross-modality SSL in the field of ophthalmology. This SSL pretext task encoded shared information between two high-dimensional modalities, including infrared fundus photography and optical coherence tomography. The fundus representation learned from the SSL pretext task contains disease-relevant features that were efficient for downstream diabetic retinopathy classification and retinal thickness measurement. However, the audio and video data used for training SSL could be seen, e.g., in [137]. In detail, by assuming that there was dense correspondence between the ultrasound video and the relevant narrative diagnosis/interpretation speech audio of the sonographer, Jiao et al.[137] proposed SSL with multimodal input, including ultrasound video-speech raw data. Interestingly, to learn domain-agnostic feature representation, Tamkin et al.[138] designed the model architecture and objective to pretrain on six unlabelled datasets. Those datasets from various domains include text, natural images, medical imaging, multichannel sensor data, speech recordings and paired text and images. 2) Learning local representation for medical applications – Xie et al.[139] also focused on local regions by utilizing spatial transformation to create dissimilar augmented views of the same input. This encouraged consistent latent feature representations of the same region from different views of the same input image and assured such consistency by minimizing a local consistency loss. The proposed algorithm was used for pretraining to initialize a downstream network and improve four publicly available CT datasets, including two tumours and 11 different types of primary human organs. Chaitanya et al.[140, 141] not only used global contrastive learning but also proposed a local version of contrastive learning. In particular, the local version of contrastive learning loss encouraged feature representations of local areas in an image to be similar with different transformations but dissimilar to different local areas in the same image. The combination of global and local contrastive learning benefited the downstream MRI segmentation task. One similar work proposed by Ouyang et al.[142, 143] employed super pixel pseudo labels and was devised for the tuning-free few-shot segmentation task, including cardiac segmentation of the MRI dataset, and organ segmentation of the abdominal MRI and CT datasets. Furthermore, the same team[144] designed a local pixelwise contrastive loss to learn discriminative pixel-level feature representations. This enabled the model to learn better interclass separability and intraclass compactness for the segmented classes on three public medical datasets with two anatomies, including cardiac and prostate. Yan et al.[145] proposed a pixel-level contrastive learning framework with a coarse-to-fine architecture to learn both local and global information and designed customized negative sampling strategies. More specifically, the global embedding was trained to discriminate various body parts on a coarse scale, assisting the local embedding to concentrate on a smaller region to distinguish finer features. The learned embeddings were applied in different downstream areas, such as landmark detection and lesion matching, on various radiological image modalities, including 3D CT and 2D X-ray of varying body parts, such as the chest, hand, and pelvis. 3) Learning multiscale information for medical applications – in histopathology, Sahasrabudhe et al.[146] proposed a self-supervised method for nuclei segmentation on whole-slide histopathology images. They utilized scale classification as a self-supervision signal under the hypothesis that the texture and size of nuclei could be seen as the level of magnification at which a patch was obtained. Sun et al.[147] introduced a multiscale SSL framework to precisely segment tissues for a multisite paediatric brain MRI dataset with motion/Gibbs artifacts. 4) Learning texture representation for medical applications – Chen et al.[148] proposed a new computer-aided diagnosis approach with contrastive texture learning loss to learn cervical optical coherence tomography image texture features. 5) Learning structural representation for medical applications – Tang et al.[149] estimated the similarity between original and augmented images through the designed structural similarity loss for enhancing medical image classification.

SSL design on specific structures. Recently, Siamese networks and teacher-student networks have become popular structures applied in medical areas. Siamese network learning for medical applications – Spitzer et al.[150] utilized a Siamese network to calculate spatial distances between image patches sampled randomly from the cor-

tex in random sections of the same brain. Learning to discriminate several cortical brain areas through their model implicitly indicated that the designed pretext task was suitable for high-resolution cytoarchitectonic mapping. Due to the benefits of decreasing the calculational expense of 3D medical imaging, Li et al.[151] extended a 2D Siamese network to a 3D Siamese network to avoid using negative pairs or large batch sizes. Their proposed SSL coped with an imbalance problem that assisted the learned radiomics features for two downstream classification tasks, including discrimination of the level of brain tumours on the MRI dataset and the stage of lung cancer on the CT dataset. Ye et al.[152] applied a Siamese network to stereo images to access depth in robotic surgery. For kidney segmentation from abdominal CT volumes, Dhere and Sivaswamy[153] used a Siamese CNN to classify whether a given pair of kidneys belonged to the same side. They designed a proxy task by utilizing the anatomical asymmetry of kidneys, and the slight variation in shape, size, and spatial location between the left and right kidneys varied slightly. Moreover, some patients were scanned many times in a so-called longitudinal manner to track therapy or to estimate changes in the disease state. Hence, some studies on longitudinal information of the scans were used for training a Siamese network to compare the embeddings of scans from the same person or different persons. To pretrain on the example of T2-weighted sagittal lumbar MRIs, Jamaludin et al.[154] utilized SSL with a Siamese CNN trained through the two losses described as follows: 1) a contrastive loss on the pairs of images scanned from the same patient (i.e., longitudinal information) at different points in time and on the pairs of images of different patients, and 2) a classification loss was used to predict vertebral body level and disc degeneration radiological grading. Rivail et al.[155] presented a self-supervised method based on a Siamese network for modelling disease progression from longitudinal data, such as longitudinal retinal optical coherence tomography. Taking advantage of a generic time-specific task, this self-supervised model learned to evaluate the time interval between pairs of scans obtained from the same patient. Teacher-student learning for medical applications – Li et al.[156] designed a new SSL approach based on the teacher-student architecture to learn distinguishing representations from gastric X-ray images for a downstream task, gastritis detection. One of the student-teacher frameworks, Mean Teacher in [157], was integrated by Liu et al.[158] in the pretraining process for semisupervised fine-tuning for thorax disease multilabel classification. Park et al.[159] used information distillation between teacher and student frameworks and the vision transformer model for chest X-ray diagnosis, including tuberculosis, pneumothorax, and COVID-19. You et al.[160, 161] also demonstrated that the distillation framework improved medical image synthesis, registration and enhancement on the left atrial (LA) segmentation challenge and the NIH pancreas CT dataset. Later, they also proposed another semi-supervised approach that used stronger data augmentation and understood the nearest neighbours whose anatomical characteristics were homogeneous from the same class but distinct for other classes in unlabelled and clinically unbalanced circumstances[162].

Instance-based discrimination for medical image analysis

Memory bank momentum MoCo.[19] The model[163] that incorporated PIRL and transfer learning could learn the invariance property for skin lesion analysis, and the results outperformed those obtained using only transfer learning or only SSL. Taking advantage of MoCo while reducing dependency on batch size, Sowrirajan et al.[164] utilized it as a fundamental framework for reducing two constraints caused during training on the X-ray image. These two constraints were large X-ray image sizes and high computational requirements. The proposed MoCo-chest X-ray (CXR) model that adjusted the data augmentation strategy used in MoCo obtained high-quality feature representations and transferable initializations for the following detection of pathologies on chest X-ray images and across different chest X-ray datasets.

Several works used MoCo for COVID-19 diagnosis. Sriram et al.[165] applied MoCo to the COVID-19 adverse event prediction task from both single and multiple images and oxygen requirement prediction. To learn meaningful and unbiased visual representations for decreeing the risk of overfitting, He et al.[166] integrated contrastive SSL training on a similar dataset into transfer learning. Zhu[167] utilized the combination of rotation and division as the supervisory signal on the SSL framework for COVID-19 classification in the small shot scenario. Based on the MoCo v2 algorithm, hierarchical pretraining, applied by Reed et al.[168], consistently converged to learn representations for experimenting on 15 of the 16 diverse datasets, spanning visual domains, including medical, driving, aerial, and simulated images. For medical datasets, they checked whether any of the five conditions were in each image of the CheXpert dataset[169] and classified 4-way pneumonia on the chest-X-ray-kids dataset[170]. Hierarchical retraining was a way to train models on datasets that were gradually more similar to the target dataset. Liang et al.[171] also employed MoCo v2 as the base for conducting a neural architecture search to search for an optimal local architecture from its data. They applied it to CheXpert-14[169] and ModelNet40[172] for five classification tasks, including pleural effusion, atelectasis, consolidation, edema, and cardiomegaly. Interestingly, to train the encoder that could extract feature representation from the panoramic radiograph of the jaw, Hu et al.[173] utilized MoCo v2 to train the feature extractor on massive healthy samples. The joint with localization consistency loss and patch-covering data augmentation strategy could improve the model′s reliability. Wu et al.[174, 175] integrated contrastive learning with federated learning[176–178]

to collaboratively learn a shared image-level representation. Federated learning trained an algorithm within different decentralized edge devices to learn a shared model and each device kept local data samples without exchanging them. They experimented on 3D cardiac MRI images using MoCo architecture for local contrastive learning. Dong and Voiculescu[179] also federated SSL based on MoCo for COVID-19 detection. He et al.[180] combined a new surrogate loss proposed by Yuan et al.[181] with MoCo-based SSL for computer-aided screening of COVID-19 infected patients utilizing radiography images. This novel surrogate loss maximized the area under the receiver operating characteristic curve (AUC), and this combination facilitated vital metrics while also maintaining model trust. Saillard et al.[182] implemented MoCo v2 on histology images from the cancer genome Atlas dataset for microsatellite instability prediction in gastric and colorectal cancers. Tomar et al.[183] applied a Style encoder to the SSL framework utilizing volumetric contrastive loss through momentum contrast[19]. The Style encoder was designed to encourage content-invariant image-level feature representation that gathered similar styled images and dispersed dissimilar styled images.

SimCLR.[20] Azizi et al.[184] proposed a new method, multi-instance contrastive learning (MICLe), to classify two kinds of medical images, dermatology on camera images and multilabel on chest X-ray images. Unlike the traditional pretrained model, this work pretrained the model on unlabelled ImageNet using SimCLR. Then, this work used MICLe to perform self-supervised pretraining on unlabelled medical images to create moderate positive pairs. Finally, supervised fine-tuning was performed on labelled medical images. Gazda et al.[185] proposed a self-supervised deep neural network that combined SimCLR and MoCo to first pretrain on an unlabelled CheXpert dataset of chest X-ray images and then transferred the pretrained representations to downstream tasks, including COVID-19 and pneumonia detection tasks, that is, the classification of respiratory diseases. In the histopathology domain, based on SimCLR, Ciga et al.[186] discovered that the combination of multiple multiorgan datasets with several types of staining and resolution properties enhanced the quality of the learned features. Li et al.[187] addressed whole-slide image classification by training the feature extractor SimCLR. Interestingly, for SimCLR training, they used patches as inputs extracted from the whole slide image and were densely cropped without overlap, which could be seen as an individual input. Ciga[188] also implemented SimCLR for breast cancer detection in histopathology. Mojab et al.[189] verified the proposed model, a SimCLR-based framework with transfer learning, on real-world ophthalmic imaging datasets for glaucoma detection. Schirris et al.[190] utilized a SimCLR-based feature extractor pretrained on histopathology tiles and extended the deep multiple instance learning (DeepMIL)[191] classification framework for homologous recombination deficiency (HRD) and microsatellite instability (MSI) classification on colorectal and cancer datasets. Zhao and Zhou[192] added the fast mixed hard negative sample strategy to rapidly synthesize more hard negative samples[193] through a convex combination for training. The proposed model was pretrained in a self-supervised way on the Chest X-ray of pneumonia dataset and fine-tuned in a supervised way on the COVID-CT dataset. Wicaksono et al.[194] combined two types of contrasting learning, rotation, and jigsaw puzzle from the context contrastive instance category and SimCLR v1 from instance contrastive learning, for the human embryo image classification task. Based on SimCLR, Manna et al.[195] also proposed an asymptotic study of the lower bound of the designed novel loss function to test the MR-Net dataset, which was composed of magnetic resonance videos of the human knee. You et al.[196] presented two learning strategies for the volumetric medical image segmentation task. One used a voxel-to-volume contrastive algorithm to obtain global information from 3D images, and the other used local voxel-to-voxel distillation to better utilize local signals in the embedding space. Yao et al.[197] were motivated by contrastive learning[20, 198], which localized the object landmark with only one labelled image available in a coarse-to-fine fashion to create pseudo-annotation for training a terminal landmark detector. The proposed model demonstrated high-performance cephalometric landmark detection, comparable to popular fully supervised approaches utilizing more than one training image. Ali et al.[199] used 3D SimCLR during pretraining and Monte Carlo dropout during prediction on two tasks, including 3D CT pancreas tumour and 3D MRI brain tumour segmentation. Inglese et al.[200] followed a similar optimization method of SimCLR to train an SSL network for distinguishing between two diagnostically different systemic lupus erythematosus patient groups. To learn task-agnostic properties, such as texture and intensity distribution, from heterogeneous data, Zheng et al.[201] first aggregated a dataset from various medical challenges. Then, they presented hierarchical SSL based on SimCLR with contrasting and classification strategies to provide supervision signals for image-level, task-level, and group-level pretext tasks. On the downstream tasks, they segmented the heart, prostate, and knee on the MRI dataset and the liver, pancreas, and spleen on the CT dataset.

Cluster-based discrimination for medical application

Abbas et al.[202] proposed a new SSL mechanism, 4S-DT, that assisted coarse-to-fine transfer learning according to a self-supervised sample decomposition of unannotated chest X-ray input. Super sample decomposition[203] was a pretext task that trained networks using cluster assignments as pseudo labels. The coarse transfer learning utilized an ImageNet pretrained CNN model for

classifying pseudo labelled chest X-ray images, creating chest X-ray related convolutional features. Fine transfer learning was used in downstream training tasks from chest X-ray recognition tasks to COVID-19 detection tasks. In histopathology, Abbet et al.[204] conducted research on learning cancerous tissue areas that could be utilized to enhance prognostic stratification for colorectal cancer. They presented an SSL method that combined the learning of tissue region representations and a clustering metric to extract their underlying patterns. Mahapatra et al.[205] utilized one of the deep clustering methods[206], named SwAV, without using class attribute vectors commonly used for natural images. They proved the effectiveness of the proposed model across different datasets with at least three disease classes. Chaves et al.[207] evaluated five SSL methods, including InfoMin, MoCo, SimCLR, BYOL, and SwAV, for diagnosing skin lesions. They compared those SSL methods and three self-supervised pipelines on five test datasets with in-distribution and out-distribution scenarios. They summarized that self-supervision is competitive both in increasing accuracy and decreasing outcome variability. Chen et al.[208] developed an SSL strategy to perform joint deep embedding and cluster assignment for dMRI tractography white matter fiber clustering. Ciga et al.[209] utilized a two-step pretraining on three popular contrastive techniques, SimCLR, BYOL and SwAV, to validate better performance on two natural and three medical images, including ChestX-ray8, breast ultrasound, and brain tumour MRI. Islam et al.[210] pretrained and compared models within fourteen different SSL approaches for pulmonary embolism classification on CT pulmonary angiography scans.

3.2.4 Temporal contrastive SSL for medical image analysis

Temporal contrastive SSL learned feature representation by grabbing the spatial or structural information between adjacent frames. Sequential images were utilized in two kinds of ways as self-supervision for the training model, such as the objects shown in the adjacent frames or the process of correcting frame order.

Finding similarities of adjacent frames for medical image analysis

One of the most common applications of temporal contrastive SSL was to find the similarity in adjacent frames. This enabled the mode to learn contextual semantic representations. In histopathology, Gildenblat and Klaiman[211] utilized the image characteristic that spatially adjacent histopathological tissue image slices were more similar to one another than distance slices, which was used to train on a Siamese network for learning image similarity. In another application, due to the cardiac MR scans composed of different angulated planes relative to the heart, Bai et al.[212] learned feature representation, through the proposed model, from information automatically defined by the heart chamber view planes. That information included anatomical positions, and the relative orientation of long-axis and short-axis views could be used to create a pretext task for SSL training. Kragh et al.[213] implemented a self-supervised video alignment method, temporal cycle consistency[214], to obtain temporal similarities between embryo videos, and this information to predict pregnancy possibility. By utilizing the position information in volumetric medical slices, Zeng et al.[181] provided a new position contrastive learning framework to produce contrastive data pairs. The framework can successfully eliminate false negative pairings in the currently common contrastive learning techniques for medical segmentation.

Tracking objects for medical image analysis

Lu et al.[215, 216] designed a pretext task to predict the density map of fiber streamlines that were the representations of generic white matter pathways for white matter tracts. They took advantage of two characteristics of the fiber streamlines. These fiber streamlines could be calculated with fiber tracking obtained automatically with tractography, and the density map of fiber streamlines was acquired as the number of streamlines accross each voxel. In short, fiber streamlines were jointed line segments with directions and could be seen as white matter pathways that provide supervision. To segment white matter tracts on diffusion magnetic resonance imaging scans, learned features of white matter tracts through the designed pretext task could predict the density map of fiber streamlines from the training data obtained through tractography.

Correcting frame orders from 3D medical images

The process of correcting frame orders from shuffled frames assisted the model in learning feature representation. Zhang et al.[217] utilized spatial context information in 3D CT and MR volumes as a source of supervision created by solving the tasks of transversal 2D slice ordering for fine-grained body part recognition. Nguyen et al.[218] also demonstrated that predicting the 2D slice order in a sequence could obtain both spatial and semantic features for downstream tasks, the detection of organ segmentation, and intracranial hemorrhage. Jiao et al.[219] corrected the order of a reshuffled fetal ultrasound video. By utilizing the characteristics of the tube-like structure of axons, Klinghoffer et al.[220] learned feature representation by training the model to predict the permutation that was utilized to reformulate the slices of each input 3D microscopy subvolume for axon segmentation. The design of the pretext task, resolution sequence prediction[221], was inspired by the approach in which a pathologist looked for cancerous regions in whole-slide images. More specifically, a pathologist zoomed in and out several times to inspect the tissue at high to low resolution to acquire the details of individual cells and the surrounding area. Srinidhi et al.[221] utilized multiresolution contextual information as a supervisory signal to train a de-

signed SSL network. This network learned visual representations by predicting the order of sequences of resolution that could be generated from the multiresolution histology whole slide image patches.

4 Conclusions and future directions

This study reviews the state-of-the-art contrastive SSL algorithms on natural images, along with their novel adaptations for medical imaging data. We cover fundamental problems when implementing SSL in medical areas and its future directions.

1) The pretext tasks of SSL can create implicit supervisory signals from unlabelled datasets to perform unsupervised learning close to, or even equal to, that of human labelling. Most of the pretext tasks we survey are all manually created by experts, and require both domain and machine learning skills, together with a comprehensive set of experiments. We believe there is an opportunity to frame the pretext task creation as an optimization problem, which is conceptually comparable to the pursuit of the best architecture for a deep learning challenge. Furthermore, learning a reliable representation from medical images will not be optimal by simply adopting pretext tasks that have been developed on natural images. Hence, such methods require to be further modified and improved to suit the nature of medical images and enable the extraction of robust representations.

2) Similar to pretext tasks, augmentation techniques used in contrastive SSL methods that are designed and optimized for natural images may not be suitable for medical images. As an example, medical images that are already grayscale would not be transformed in a meaningful way by color jittering or random grayscale, which are common techniques applied to natural images. The effects of various additional augmentations and their combinations should be studied in further research.

3) Sampling strategies are one of the reasons for the success of mutual information-based systems, as noted by Tschannen et al.[44] Sampling strategies may affect contrastive SSL methods, such as MoCo and SimCLR, which need huge amounts of negative samples. Hence, how to decrease the reliance on sampling strategies is still an appealing and unsolved problem. A suitable negative sample can be built based on the properties of medical images, and from there, more valuable data features can be extracted[222, 223]. There needs to be a furthur investigation on how to create negative samples and how to better adapt SSL to downstream tasks to enhance the performance of SSL approaches in the medical imaging domain. Moreover, along with data augmentation, the redesign of the contrastive loss function plays a crucial role in the performance. Some researchers have worked on designing contrastive loss functions for their particular purposes in medical areas and related to e.g., multimodal learning[136, 137, 224], local representation learning[139], multiscale learning, and texture[148] or structural[149] representation learning.

Appendix

In Tables A1–A3, the contents of column dataset used

Table A1 Self-supervision: Predict learning

Pretext task	Author(s)	Dataset(s) used (in pretraining, testing, and downstream tasks)	Application(s)
Relative position	Chen et al.[83], 2019	2D fetal ultrasound image D abdominal CT image Brain TMR image (BraTS challenge)	Fetal standard scan plane classification Abdominal multiorgan localization Brain tumor segmentation
	Blendowski et al.[87], 2019	VISCERAL Anatomy CT dataset	Multiorgan segmentation (liver, spleen, left kidney, right kidney, left psoas major muscle, and right psoas major muscle)
	Jana et al.[88], 2021	MICCAI 20 7 LiTS challenge dataset CT images	Fibrosis classification NAS score classification (nonalcoholic fatty liver disease (NAFLD), activity scores (NAS))
	Li et al.[89], 2021	Chest CT images	COVID-19 severity level prediction
Jigsaw puzzle	Taleb et al.[91], 2020	BraTS dataset Prostate dataset CHAOS multimodal dataset	Survival days prediction, and multimodal brain tumor segmentation Prostate segmentation Liver segmentation
	Taleb et al.[92], 2021	BraTS challenge	Brain tumor segmentation Survival prediction regression
	Navarro et al.[93], 2021	X-ray images (RSNA) VISCERAL CT dataset Grand challenges CT dataset	Pneumonia classification Multiorgan segmentation
	Manna et al.[94], 2021	MRNet dataset	Three knee conditions classification (abnormality, ACL tear, and meniscus tear)

Table A1 (continued) Self-supervision: Predict learning

Pretext task	Author(s)	Dataset(s) used (in pretraining, testing, and downstream tasks)	Application(s)
	Li et al.[95], 2020	MoNuSeg dataset ISIC dataset	[Histopathological images] Nuclei segmentation Skin lesion segmentation
	Chae et al.[102], 2021	Cervix image dataset	Cervical cancer classification
	Santilli et al.[103], 2021	REIMS data	Breast cancer classification
Rubik's cube	Zhuang et al.[104], 2019	Brain hemorrhage CT dataset (private dataset) BraTS-20 8	Brain hemorrhage classification Brain tumor segmentation
	Zhu et al.[105], 2020	Cerebral hemorrhage dataset BraTS-20 8	Cerebral hemorrhage classification Brain tumor segmentation
	Tao et al.[106], 2020	NIH Pancreas CT dataset MRBrainS 8 dataset	Pancreas segmentation Brain tissue segmentation
	Li et al.[107], 2020	COVID- 9 CT dataset	Distinguishing COVID- 9 from other two cases: Nonpneumonia and community acquired pneumonia (CAP) on chest CT exams
Rotation	Tajbakhsh et al.[116], 2019	LIDC-IDRI chest CTs Diabetic retinopathy (DR) fundus image dataset Private dataset (color, telemedicine)	False-positive reduction (FPR) for nodule detection Lung lobe segmentation DR classification in fundus images Skin segmentation
	Li et al.[108], 2021	iChallenge-AMD dataset iChallenge-PM dataset EyePACS dataset/Kaggle DR	Retinal disease classification
	Yang et al.[109], 2020	LiTS 2017 MICCAI	Cross-modality liver segmentation
	Liu et al.[113], 2019	NLST dataset LUNA 6 dataset SPIE-AAPM dataset Lung TIME dataset HMS Lung cancer dataset	Pulmonary nodule classification
	Dong et al.[114], 2021	CT images dataset	Focal liver lesions classification
	Koohbanani et al.[119], 2020	Camelyon 6 LNM-OSCC Kather	[histopathology image] Histology image classification

Table A2 Self-supervision: Context-instance contrast/Maximizing mutual information

Pretext task	Author(s)	Dataset(s) used (in pretraining, testing, and downstream tasks)	Application(s)
Contrastive predictive coding (CPC)	Stacke et al.[130], 2020	STL-10 CAMELYON17 AIDA-LNCO AIDA-SKIN	[Histopathological images] Tumor classification
	Taleb et al.[131], 2020	Multimodal brain tumor segmentation (BraTS) 2018 Pancreas dataset Diabetic Retinopathy 2019 Kaggle challenge UK Biobank (UKB)	Brain tumor segmentation Pancreas tumor segmentation Diabetic retinopathy detection
	Zhu et al.[132], 2021	3D brain hemorrhage dataset (private dataset)	Brain hemorrhage classification
	Zhu et al.[134], 2020	Brain hemorrhage dataset (private dataset) LUNA16 dataset	Brain hemorrhage classification Lung nodule classification

and column applications are from the reference mentioned in column authors.

Acknowledgements

Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Open Access

Table A3 Self-supervision: Instance-instance contrast/Predicting spatial relative position

Pretext task	Author(s)	Dataset(s) used (in pretraining, testing, and downstream tasks)	Application(s)
PIRL	Kwasigroch et al.[163], 2020	ISIC2017 challenge dataset	Skin lesion classification
MoCo	Sowrirajan et al.[164], 2021	CheXpert dataset Shenzhen hospital X-ray dataset	Chest X-ray interpretation
	Sriram et al.[165], 2021	MIMIC-CXR dataset CheXpert NYU COVID dataset	Adverse event prediction from single images (SIP) Adverse event prediction from multiple images (MIP) Oxygen requirements prediction from single images (ORP)
	He et al.[166], 2020	They built the COVID19-CT dataset through collecting medical images from COVID-19 relative bioRxiv and medRxiv papers	Diagnosing COVID-19 from CT scans
	Reed et al.[168], 2022	Chexpert chest-X-ray-kids	Five classification on Chexpert dataset Singular classification on Chest-X-ray-kids dataset
	Liang et al.[171], 2018	FedCheXpert	Multiclass classification
MoCo + SimCLR	Gazda et al.[185], 2021	CheXpert dataset Cell dataset ChestX-ray14 C19-Cohen dataset COVIDGR dataset	Pneumonia classification COVID-19 classification
SimCLR	Azizi et al.[184], 2021	Dermatology dataset CheXpert dataset	Dermatology skin condition classification Five pathologies chest X-ray classification
	Ciga et al.[186], 2022	BACH challenge dataset Patch Camelyon BreakHis NCT-CRC-HE-100K/Kather PANDA BACH challenge dataset Gleason2019 DigestPath2019 BreastPathQ dataset	[histopathology images] Breast cancer classification Lymph node classification Breast tumor classification Colorectal cancer classification Prostate cancer grading Breast cancer segmentation Prostate cancer grading Colon tumor segmentation. Percentage of cancer cellularity of each patch
	Li et al.[187], 2021	Camelyon16 The cancer genome atlas (TCGA) lung cancer dataset	[histopathology images] Breast cancer classification and localization lung cancer classification
	Inglese et al.[200], 2022	Private dataset for diagnosing NPSLE	NPSLE/non-NPSLE classification
	Zheng et al.[201], 2021	LASC LiTS MSD Knee ACDC M&Ms	Eight medical image segmentation: cardiovascular structures, liver & tumours, spleen, knee bones & cartilages, and prostate
Cluster discrimination	Abbas et al.[202], 2021	Collected from three different dataset COVID-19 dataset-A COVID-19 dataset-B	Detection of COVID-19 cases
	Abbet et al.[204], 2020	Kather dataset	[WSIs histopathological images] Colorectal cancer classification
	Chaves et al.[207], 2022	Isic19 Isic20 Derm7pt–derm and derm7pt–clinic Pad-ufes-20	Skin lesions classification
SSL design on contrastive loss function-based variation	Holmberg et al.[136], 2020	Kaggle diabetic retinopathy dataset Tissue segmentation infrared (IR) fundus image dataset	Diabetic retinopathy classification OCT retinal thickness measurements
	Xie et al.[139], 2020	Collected from public datasets (RibFrac dataset and medical segmentation decathlon (MSD) challenge) Liver dataset Spleen dataset KiTS dataset BCV dataset	Human organs and two tumor, such as liver and kidney segmentation

Table A3 (continued) Self-supervision: Instance-instance contrast/Predicting spatial relative position

Pretext task	Author(s)	Dataset(s) used (in pretraining, testing, and downstream tasks)	Application(s)
	Chaitanya et al.[140,141], 2020	ACDC dataset Prostate dataset MMWHS dataset	Cardiac multistructures segmentation Prostate structures segmentation Heart multistructures segmentation
	Yan et al.[145], 2022	DeepLesion CT datasets NIH-LN ChestCT dataset	3D universal lesion matching on CT 2D landmark detection on hand and pelvic X-rays 3D 19 landmark detection on chest CT
	Sahasrabudhe et al.[146], 2020	MoNuSeg TNBC CoNSeP	[Histopathological images] Nuclei segmentation
	Xie et al.[122], 2020	MoNuSeg 2018 Dataset	[Histopathological images] Nuclei segmentation
SSL design on specific structures	Spitze et al.[150], 2018	Generated a dataset based on BigBrain	[Histological images] Cytoarchitectonic segmentation of human brain areas
	Li et al.[151], 2021	BraTS NSCLC-radiomics	Brain tumor classification Lung cancer staging
	Dhere and Sivaswamy[153], 2021	KiTS 2019 challenge	Kidney segmentation
	Jamaludin et al.[154], 2017	In-house dataset (TwinsUK registry)	Radiological grading classification
	Rivail et al.[155], 2019	Longitudinal dataset	Conversion to advanced AMD classification
	Li et al.[156], 2021	Gastric X-ray image dataset	Gastritis detection
	Liu et al.[158], 2021	Chest X-ray14	Thorax disease multilabel classification
Temporal contrast	Gildenblat and Klaiman[211], 2020	Camelyon16	[Histopathological images] Image retrieval for tumor areas
	Bai et al.[212], 2019	UK Biobank	Cardiac MR image segmentation
	Lu et al.[215,216], 2021	HCP dMRI scan dataset	White matter tract segmentation
	Islam et al.[210, 218], 2021	StructSeg dataset RSNA Intracranial hemorrhage is a CT scan dataset	Organ segmentation Intracranial hemorrhage detection
	Jiao et al.[219], 2020	Clinical fetal US dataset	Standard plane detection and saliency prediction
	Klinghoffer et al.[220], 2020	SHIELD PVGPe dataset Single neuron Janelia dataset	Axon segmentation
	Srinidhi et al.[221], 2022	BreastPathQ dataset Camelyon16 dataset Kather multiclass dataset	Tumor metastasis detection Tissue type classification Tumor cellularity quantification

References

[1] Y. LeCun, Y. Bengio, G. Hinton. Deep learning. *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[2] S. J. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach*, 3rd ed., Upper Saddle River, USA: Prentice Hall, 2010.

[3] X. Liu, F. J. Zhang, Z. Y. Hou, L. Mian, Z. Y. Wang, J. Zhang, J. Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2023. DOI: 10.1109/TKDE.2021.3090866.

[4] X. Yan, S. Z. Gilani, M. T. Feng, L. Zhang, H. L. Qin, A. Mian. Self-supervised learning to detect key frames in videos. *Sensors*, vol. 20, no. 23, Article number 6941, 2020. DOI: 10.3390/s20236941.

[5] S. L. Happy, A. Dantcheva, F. Bremond. A Weakly Su-

pervised learning technique for classifying facial expressions. *Pattern Recognition Letters*, vol. 128, pp. 162–168, 2019. DOI: 10.1016/j.patrec.2019.08.025.

[6] I. B. Senkyire, Z. Liu. Supervised and semi-supervised methods for abdominal organ segmentation: A review. *International Journal of Automation and Computing*, vol. 18, no. 6, pp. 887–914, 2021. DOI: 10.1007/s11633-021-1313-0.

[7] K. Kumar, A. C. S. Rao. Breast cancer classification of image using convolutional neural network. In *Proceedings of the 4th International Conference on Recent Advances in Information Technology*, IEEE, Dhanbad, India, pp. 1–6, 2018. DOI: 10.1109/RAIT.2018.8389034.

[8] R. Sarki, K. Ahmed, H. Wang, Y. C. Zhang, K. T Wang. Automated detection of COVID-19 through convolutional neural network using chest X-ray images. *PLoS One*, vol. 17, no. 1, Article number e0262052, 2022. DOI: 10.1371/journal.pone.0262052.

[9] M. Cullell-Dalmau, S. Noé, M. Otero-Viñas, I. Meić, M. Manzo. Convolutional neural network for skin lesion classification: Understanding the fundamentals through hands-on learning. *Frontiers in Medicine*, vol. 8, Article number 644327, 2021. DOI: 10.3389/fmed.2021.644327.

[10] M. Raghu, C. Y. Zhang, J. M. Kleinberg, S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 301, 2019. DOI: 10.5555/3454287.3454588.

[11] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. M. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016. DOI: 10.1109/TMI.2016.2535302.

[12] H. C. Shin, H. R. Roth, M. C. Gao, L. Lu, Z. Y. Xu, I. Nogues, J. H. Yao, D. Mollura, R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016. DOI: 10.1109/TMI.2016.2528162.

[13] V. Cheplygina, M. de Bruijne, J. P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, vol. 54, pp. 280–296, 2019. DOI: 10.1016/j.media.2019.03.009.

[14] C. P. Wilkinson, F. L. Ferris III, R. E. Klei, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer. Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003. DOI: 10.1016/S0161-6420(03)00475-5.

[15] P. Khosravi, E. Kazemi, Q. S. Zhan, M. Toschi, J. Malmsten, C. Hickman, M. Meseguer, Z. Rosenwaks, O. Elemento, N. Zaninovic, I. Hajirasouliha. Robust automated assessment of human blastocyst quality using deep learning. bioRxiv, 2018.

[16] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016. DOI: 10.1001/jama.2016.17216.

[17] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. [Online], Available: https://arxiv.org/abs/1711.05225, 2017.

[18] K. M. He, R. Girshick, P. Dollar. Rethinking ImageNet pre-training. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 4917–4926, 2019. DOI: 10.1109/ICCV.2019.00502.

[19] K. M. He, H. Q. Fan, Y. X. Wu, S. N. Xie, R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9726–9735, 2020. DOI: 10.1109/CVPR42600.2020.00975.

[20] T. Chen, S. Kornblith, M. Norouzi, G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Article number 149, 2020. DOI: 10.5555/3524938.3525087.

[21] Y. LeCun. Self-supervised learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence Invited Speaker Program*, New York, USA, 2020. [Online], Available: https://aaai.org/conference/aaai/aaai-20/invited-speakers/.

[22] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020.

[23] I. Misra, L. van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 6706–6716, 2020. DOI: 10.1109/CVPR42600.2020.00674.

[24] A. Newell, J. Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 7343–7352, 2020. DOI: 10.1109/CVPR42600.2020.00737.

[25] A. Tendle, M. R. Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, vol. 6, Article number 100124, 2021. DOI: 10.1016/j.mlwa.2021.100124.

[26] K. Kotar, G. Ilharco, L. Schmidt, K. Ehsani, R. Mottaghi. Contrasting contrastive self-supervised representation learning pipelines. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 9929–9939, 2021. DOI: 10.1109/ICCV48922.2021.00980.

[27] P. Wang, K. Han, X. S. Wei, L. Zhang, L. Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 943–952, 2021. DOI: 10.1109/CVPR46437.2021.00100.

[28] L. L. Jing, Y. L. Tian. Self-supervised visual feature

learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2021. DOI: 10.1109/tpami.2020.2992393.

[29] X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, X. J. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020. DOI: 10.1007/s11431-020-1647-3.

[30] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, vol. 9, no. 1, Article number 2, 2021. DOI: 10.3390/technologies9010002.

[31] S. Shurrab, R. Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, vol. 8, Article number e1045, 2022. DOI: 10.7717/peerj-cs.1045.

[32] C. Y. Zhang, Y. Gu. Dive into self-supervised learning for medical image analysis: Data, models and tasks. [Online], Available: https://arxiv.org/abs/2209.12157, 2022.

[33] C. Doersch, A. Gupta, A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1422–1430, 2015. DOI: 10.1109/ICCV.2015.167.

[34] M. Noroozi, P. Favaro. Unsupervised learning of visual representations by solving Jigsaw puzzles. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 69–84, 2016. DOI: 10.1007/978-3-319-46466-4_5.

[35] S. Gidaris, P. Singh, N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[36] J. N. Li, J. Q. Liu, Y. K. Wong, S. Nishimura, M. S. Kankanhalli. Self-supervised representation learning using 360° data. In *Proceedings of the 27th ACM International Conference on Multimedia*, ACM, Nice, France, pp. 998–1006, 2019. DOI: 10.1145/3343031.3351019.

[37] H. Lee, S. J. Hwang, J. Shin. Rethinking data augmentation: Self-supervision and self-distillation. [Online], Available: https://arxiv.org/abs/1910.05872, 2019.

[38] L. Ericsson, H. Gouk, C. C. Loy, T. M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–46, 2022. DOI: 10.1109/MSP.2021.3134634.

[39] Sequential data. In *Encyclopedia of Machine Learning*, C. Sammut, G. I. Webb, Eds., Boston, USA: Springer, 2011. [Online], Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_754.

[40] Z. J. Yue, S. Ding, L. Zhao, Y. T. Zhang, Z. H. Cao, M. Tanveer, A. Jolfaei, X. Zheng. Privacy-preserving time-series medical images analysis using a hybrid deep learning framework. *ACM Transactions on Internet Technology*, vol. 21, no. 3, Article number 57, 2021. DOI: 10.1145/3383779.

[41] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[42] M. Gutmann, A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, pp. 297–304, 2010.

[43] A. van den Oord, Y. Z. Li, O. Vinyals. Representation learning with contrastive predictive coding. [Online], Available: https://arxiv.org/abs/1807.03748, 2018.

[44] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, M Lucic. On mutual information maximization for representation learning. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[45] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, D. Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 531–540, 2018.

[46] R. Linsker. Self-organization in a perceptual network. *Computer*, vol. 21, no. 3, pp. 105–117, 1988. DOI: 10.1109/2.36.

[47] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, Scottsdale, USA, 2013.

[48] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[49] Y. L. Tian, D. Krishnan, P. Isola. Contrastive multiview coding. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 776–794, 2020. DOI: 10.1007/978-3-030-58621-8_45.

[50] P. Bachman, R. D. Hjelm, W. Buchwalter. Learning representations by maximizing mutual information across views. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1392, 2019.

[51] O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, A. van den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, pp. 4182–4192, 2020.

[52] P. Elias. Predictive coding-I. *IRE Transactions on Information Theory*, vol. 1, no. 1, pp. 16–24, 1955. DOI: 10.1109/TIT.1955.1055126.

[53] B. S. Atal, M. R. Schroeder. Adaptive predictive coding of speech signals. *The Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, 1970. DOI: 10.1002/j.1538-7305.1970.tb04297.x.

[54] Z. R. Wu, Y. J. Xiong, S. X. Yu, D. H. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3733–3742, 2018. DOI: 10.1109/CVPR.2018.00393.

[55] D. Yoo, S. Park, J. Y. Lee, I. S. Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Boston, USA, pp. 71–80, 2015. DOI: 10.1109/CVPRW.2015.7301274.

[56] P. Agrawal, J. Carreira, J. Malik. Learning to see by moving. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 37–45, 2015. DOI: 10.1109/ICCV.2015.13.

[57] J. S. Chung, A. Zisserman. Lip reading in profile. In *Proceedings of British Machine Vision Conference*, BMVA Press, London, UK, pp. 155, 2017. DOI: 10.5244/C.31.155.

[58] X. L. Chen, K. M. He, Exploring simple Siamese representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 15745–15753, 2021. DOI: 10.1109/CVPR46437.2021.01549.

[59] X. L. Chen, H. Q. Fan, R. Girshick, K. M. He. Improved baselines with momentum contrastive learning. [Online], Available: https://arxiv.org/abs/2003.04297, 2020.

[60] X. L. Chen, S. N. Xie, K. M. He. An empirical study of training self-supervised vision transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 9620–9629, 2021. DOI: 10.1109/ICCV48922.2021.00950.

[61] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. DOI: 10.1162/neco.1989.1.4.541.

[62] M. Ye, X. Zhang, P. C. Yuen, and S. F. Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 6203–6212, 2019. DOI: 10.1109/CVPR.2019.00637.

[63] K. Nozawa, I. Sato. Understanding negative samples in instance discriminative self-supervised representation learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 5784–5797, 2021.

[64] S. Becke, G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, vol. 355, no. 6356, pp. 161–163, 1992. DOI: 10.1038/355161a0.

[65] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th Conference on Neural Information Processing System*, Vancouver, Canada, 2020.

[66] N. Silberman, D. Hoiem, P. Kohli, R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 746–760, 2012. DOI: 10.1007/978-3-642-33715-4_54.

[67] C. J. Reed, S. Metzger, A. Srinivas, T. Darrell, K. Keutzer. SelfAugment: Automatic augmentation policies for self-supervised learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 2673–2682, 2021. DOI: 10.1109/CVPR46437.2021.00270.

[68] T. T. Xiao, X. L. Wang, A. A. Efros, T. Darrell. What should not be contrastive in contrastive learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[69] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. A. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[70] J. W. Yang, D. Parikh, D. Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 5147–5156, 2016. DOI: 10.1109/CVPR.2016.556.

[71] J. Y. Xie, R. Girshick, A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, New York, USA, pp. 478–487, 2016.

[72] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th Conference on Neural Information Processing System*, Vancouver, Canada, 2020.

[73] H. Mobahi, R. Collobert, J. Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, Montreal, Canada, pp. 737–744, 2009. DOI: 10.1145/1553374.1553469.

[74] D. Jayaraman, K. Grauman, K. Slow and steady feature analysis: Higher order temporal coherence in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 3852–3861, 2016. DOI: 10.1109/CVPR.2016.418.

[75] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, G. Brain. Time-contrastive networks: Self-supervised learning from video. In *Proceedings of IEEE International Conference on Robotics and Automation*, Brisbane, Australia, pp. 1134–1141, 2018. DOI: 10.1109/ICRA.2018.8462891.

[76] X. L. Wang, A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2794–2802, 2015. DOI: 10.1109/ICCV.2015.320.

[77] J. Walker, A. Gupta, M. Hebert. Dense optical flow prediction from a static image. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2443–2451, 2015. DOI: 10.1109/ICCV.2015.281.

[78] S. Purushwalkam, A. Gupta. Pose from action: Unsupervised learning of pose features based on motion. [Online], Available: https://arxiv.org/abs/1609.05420, 2016.

[79] P. Sermanet, C. Lynch, J. Hsu, S. Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, USA, pp. 486–487, 2017. DOI: 10.1109/CVPRW.2017.69.

[80] I. Misra, C. L. Zitnick, M. Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 527–544, 2016. DOI: 10.1007/978-3-319-46448-0_32.

[81] B. Fernando, H. Bilen, E. Gavves, S. Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 5729–5738, 2017. DOI: 10.1109/CVPR.2017.607.

[82] H. Y. Lee, J. B. Huang, M. Singh, M. H. Yang. Unsuper-

vised representation learning by sorting sequences. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 667–676, 2017. DOI: 10.1109/ICCV.2017.79.

[83] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, vol. 58, Article number 101539, 2019. DOI: 10.1016/j.media.2019.101539.

[84] R. Zhang, P. Isola, A. A. Efros. Colorful image colorization. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 649–666, 2016. DOI: 10.1007/978-3-319-46487-9_40.

[85] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, T. Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016. DOI: 10.1109/TPAMI.2015.2496141.

[86] D. Pathak, R. Girshick, P. Dollár, T. Darrell, B. Hariharan. Learning features by watching objects move. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6024–6033, 2017. DOI: 10.1109/CVPR.2017.638.

[87] M. Blendowski, H. Nickisch, M. P. Heinrich. How to learn from unlabeled volume data: Self-supervised 3D context feature learning. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Shenzhen, China, pp. 649–657, 2019. DOI: 10.1007/978-3-030-32226-7_72.

[88] A. Jana, H. Qu, C. D. Minacapelli, C. Catalano, V. Rustgi, D. Metaxas. Liver fibrosis and NAS scoring from CT images using self-supervised learning and texture encoding. In*Proceedings of the 18th IEEE International Symposium on Biomedical Imaging*, Nice, France, pp. 1553–1557, 2021. DOI: 10.1109/ISBI48211.2021.9433920.

[89] Z. K. Li, W. Zhao, F. Shi, L. Qi, X. Z. Xie, Y. Wei, Z. X. Ding, Y. Gao, S. J. Wu, J. Liu, Y. H. Shi, D. G. Shen. A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. *Medical Image Analysis*, vol. 69, Article number 101978, 2021. DOI: 10.1016/j.media.2021.101978.

[90] P. A. Fashi. A Self-supervised Contrastive Learning Approach for Whole Slide Image Representation in Digital Pathology, Master dissertation, University of Waterloo, Canada, 2022.

[91] A. Taleb, C. Lippert, T. Klein, M. Nabi. Self-supervised learning for medical images by solving multimodal jigsaw puzzles. [Online], Available: https://arxiv.org/abs/1912.05396, 2020.

[92] A. Taleb, C. Lippert, T. Klein, M. Nabi. Multimodal self-supervised learning for medical image analysis. In *Proceedings of the 27th International Conference on Information Processing in Medical Imaging*, Springer, Cham, Switzerland, pp. 661–673, 2021. DOI: 10.1007/978-3-030-78191-0_51.

[93] F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, B. H. Menze. Evaluating the robustness of self-supervised learning in medical imaging. [Online], Available: https://arxiv.org/abs/2105.06986, 2021.

[94] S. Manna, S. Bhattacharya, U. Pal. SSLM: Self-supervised learning for medical diagnosis from MR video. [Online], Available: https://arxiv.org/abs/2104.10481, 2021.

[95] Y. X. Li, J. W. Chen, X. P. Xie, K. Ma, Y. F. Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 614–623, 2020. DOI: 10.1007/978-3-030-59710-8_60.

[96] M. Luna, M. Kwon, S. H. Park. Precise separation of adjacent nuclei using a Siamese neural network. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Shenzhen, China, pp. 577–585, 2019. DOI: 10.1007/978-3-030-32239-7_64.

[97] H. Oda, H. R. Roth, K. Chiba, J. Sokolić, T. Kitasaka, M. Oda, A. Hinoki, H. Uchida, J. A. Schnabel, K. Mori. BESNet: Boundary-enhanced segmentation of cells in histopathological images. In *Proceedings of the 21st International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Granada, Spain, pp. 228–236, 2018. DOI: 10.1007/978-3-030-00934-2_26.

[98] Y. N. Zhou, O. F. Onder, Q. Dou, E. Tsougenis, H. Chen, P. A. Heng. CIA-Net: Robust nuclei instance segmentation with contour-aware information aggregation. In *Proceedings of the 26th International Conference on Information Processing in Medical Imaging*, Springer, Hong Kong, China, pp. 682–693, 2019. DOI: 10.1007/978-3-030-20351-1_53.

[99] Y. X. Li, L. L. Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, vol. 18, no. 2, Article number 556, 2018. DOI: 10.3390/s18020556.

[100] E. Nasr-Esfahani, S. Rafiei, M. H. Jafari, N. Karimi, J. S. Wrobel, S. Samavi, S. M. R. Soroushmehr. Dense pooling layers in fully convolutional network for skin lesion segmentation. *Computerized Medical Imaging and Graphics*, vol. 78, Article number 101658, 2019. DOI: 10.1016/j.compmedimag.2019.101658.

[101] Y. J. Tang, F. Yang, S. F. Yuan, C. A. Zhan. A multi-stage framework with context information fusion structure for skin lesion segmentation. In *Proceedings of the 16th IEEE International Symposium on Biomedical Imaging*, Venice, Italy, pp. 1407–1410, 2019. DOI: 10.1109/ISBI.2019.8759535.

[102] J. Chae, R. Zimmermann, D. Kim, J. Kim. Attentive transfer learning via self-supervised learning for cervical dysplasia diagnosis. *Journal of Information Processing Systems*, vol. 17, no. 3, pp. 453–461, 2021. DOI: 10.3745/JIPS.04.0214.

[103] A. M. L. Santilli, A. Jamzad, A. Sedghi, M. Kaufmann, K. Logan, J. Wallis, K. Y. M. Ren, N. Janssen, S. Merchant, J. Engel, D. McKay, S. Varma, A. M. Wang, G. Fichtinger, J. F. Rudan, P. Mousavi. Domain adaptation and self-supervised learning for surgical margin detection. *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 5, pp. 861–869, 2021. DOI: 10.1007/s11548-021-02381-6.

[104] X. R. Zhuang, Y. X. Li, Y. F. Hu, K. Ma, Y. J. Yang, Y. F. Zheng. Self-supervised feature learning for 3D medical images by playing a Rubik's cube. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer,

Shenzhen, China, pp. 420–428, 2019. DOI: 10.1007/978-3-030-32251-9_46.

[105] J. W. Zhu, Y. X. Li, Y. F. Hu, K. Ma, S. K. Zhou, Y. Zheng. Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis. *Medical Image Analysis*, vol. 64, Article number 101746, 2020. DOI: 10.1016/j.media.2020.101746.

[106] X. Tao, Y. X. Li, W. H. Zhou, K. Ma, Y. F. Zheng. Revisiting Rubik's cube: Self-supervised learning with volume-wise transformation for 3D medical image segmentation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 238–248, 2020. DOI: 10.1007/978-3-030-59719-1_24.

[107] Y. X. Li, D. Wei, J. W. Chen, S. L. Cao, H. Y. Zhou, Y. C. Zhu, J. R. Wu, L. Lan, W. B. Sun, T. Y. Qian, K. Ma, H. B. Xu, Y. F. Zheng. Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank. *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2787–2797, 2020. DOI: 10.1109/JBHI.2020.3018181.

[108] X. M. Li, X. W. Hu, X. J. Qi, L. Q. Yu, W. Zhao, P. A. Heng, L. Xing. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2284–2294, 2021. DOI: 10.1109/TMI.2021.3075244.

[109] J. L. Yang, X. X. Li, D. Pak, N. C. Dvornek, J. Chapiro, M. D. Lin, J. S. Duncan. Cross-modality segmentation by self-supervised semantic alignment in disentangled content space. In *Proceedings of the 2nd MICCAI Workshop on Domain Adaptation and Representation Transfer, DCL: MICCAI Workshop on Distributed and Collaborative Learning*, Springer, Lima, Peru, pp. 52–61, 2020. DOI: 10.1007/978-3-030-60548-3_6.

[110] P. Y. Simard, D. Steinkraus, J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the 7th International Conference on Document Analysis and Recognition*, IEEE, Edinburgh, UK, pp. 958–963, 2003. DOI: 10.1109/ICDAR.2003.1227801.

[111] L. L. Jing, Y. L. Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. [Online], Available: https://arxiv.org/abs/1811.11387, 2018.

[112] T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 936-944, 2017. DOI: 10.1109/CVPR.2017.106.

[113] J. Y. Liu, L. L. Cao, O. Akin, Y. L. Tian. Accurate and robust pulmonary nodule detection by 3D feature pyramid network with self-supervised feature learning. [Online], Available: https://arxiv.org/abs/1907.11704, 2019.

[114] H. H. Dong, Y. Iwamoto, X. H. Han, L. F. Lin, H. J. Hu, X. J. Cai, Y. W. Chen. Case discrimination: Self-supervised feature learning for the classification of focal liver lesions. In *Innovation in Medicine and Healthcare*, Y. W. Chen, S. Tanaka, R. J. Howlett, L. C. Jain, Eds., Singapore: Springer, pp. 241–249, 2021. DOI: 10.1007/978-981-16-3013-2_20.

[115] A. A. Z. Imran, C. Huang, H. Tang, W. Fan, Y. Xiao, D. Hao, Z. Qian, D. Terzopoulos. Self-supervised, semi-supervised, multi-context learning for the combined classification and segmentation of medical images (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 10, pp. 13815–13816, 2020. DOI: 10.1609/aaai.v34i10.7179.

[116] N. Tajbakhsh, Y. F. Hu, J. L. Cao, X. J. Yan, Y. Xiao, Y. Lu, J. M. Liang, D. Terzopoulos, X. W. Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In *Proceedings of the 16th IEEE International Symposium on Biomedical Imaging*, Venice, Italy, pp. 1251–1255, 2019. DOI: 10.1109/ISBI.2019.8759553.

[117] M. Arjovsky, S. Chintala, L. Bottou. Wasserstein GAN. [Online], Available: https://arxiv.org/abs/1701.07875, 2017.

[118] G. Larsson, M. Maire, G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 840–849, 2017. DOI: 10.1109/CVPR.2017.96.

[119] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, N. Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2845–2856, 2021. DOI: 10.1109/TMI.2021.3056023.

[120] A. Vats, M. Pedersen, A. Mohammed. A preliminary analysis of self-supervision for wireless capsule endoscopy. In *Proceedings of the 9th European Workshop on Visual Information Processing*, IEEE, Paris, France, 2021. DOI: 10.1109/EUVIP50544.2021.9484012.

[121] J. Y. Liu, G. M. Zhao, Y. Fei, M. Zhang, Y. Z. Wang, Y. Z. Yu. Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision. In *Proceeding of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 10631–10640, 2019. DOI: 10.1109/ICCV.2019.01073.

[122] X. P. Xie, J. W. Chen, Y. X. Li, L. L. Shen, K. Ma, Y. F. Zheng. Instance-aware self-supervised learning for nuclei segmentation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Lima, Peru, pp. 341–350, 2020. DOI: 10.1007/978-3-030-59722-1_33.

[123] L. Sun, K. Yu, K. Batmanghelich. Context matters: Graph-based self-supervised representation learning for medical images. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 6, pp. 4874–4882, 2021. DOI: 10.1609/aaai.v35i6.16620.

[124] Y. H. Zhang, H. Jiang, Y. Miura, C. D. Manning, C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. [Online], Available: https://arxiv.org/abs/2010.00747, 2020.

[125] N. S. Punn, S. Agarwal. BT-Unet: A self-supervised learning framework for biomedical image segmentation using Barlow Twins with U-Net models. *Machine Learning*, vol. 111, no. 12, pp. 4585–4600, 2022. DOI: 10.1007/s10994-022-06219-3.

[126] A. Kaku, S. Upadhya, N. Razavian. Intermediate layers matter in momentum contrastive self supervised learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 24063–24074, 2021.

[127] M. R. H. Taher, F. Haghighi, M. B. Gotway, J. M. Liang. CAiD: Context-aware instance discrimination for self-supervised learning in medical imaging. [Online], Available: https://arxiv.org/abs/2204.07344, 2022.

[128] H. Y. Zhou, C. X. Lu, S. B. Yang, X. G. Han, Y. Z. Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 3479–3489, 2021. DOI: 10.1109/ICCV48922.2021.00348.

[129] Y. P. Chen, Y. H. Lo, F. P. Lai, C. H. Huang. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: Algorithm development and validation study. *Journal of Medical Internet Research*, vol. 23, no. 1, Article number e25113, 2021. DOI: 10.2196/25113.

[130] K, Stacke, C. Lundström, J. Unger, G. Eilertsen. Evaluation of contrastive predictive coding for histopathology applications. In *Proceedings of Machine Learning for Health NeurIPS Workshop*, pp. 328–340, 2020.

[131] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, C. Lippert. 3D self-supervised methods for medical imaging. In *Proceedings of the 34th Conference on Neural Information Processing System*, Vancouver, Canada, pp. 1524, 2020.

[132] J. W. Zhu, Y. X. Li, S. K. Zhou. Aggregative self-supervised feature learning from a limited sample. [Online], Available: https://arxiv.org/abs/2012.07477, 2021.

[133] Z. W. Zhou, V. Sodha, M. M. R. Siddiquee, R. B. Feng, N. Tajbakhsh, M. B. Gotway, J. M. Liang. Models genesis: Generic autodidactic models for 3D medical image analysis. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Shenzhen, China, pp. 384–393, 2019. DOI: 10.1007/978-3-030-32251-9_42.

[134] J. W. Zhu, Y. X. Li, Y. F. Hu, S. K. Zhou. Embedding task knowledge into 3D neural networks via self-supervised learning. [Online], Available: https://arxiv.org/abs/2006.05798, 2020.

[135] G. Z. Jian, G. S. Lin, C. M. Wang, S. L. Yan. Helicobacter pylori infection classification based on convolutional neural network and self-supervised learning. In *Proceedings of the 5th International Conference on Graphics and Signal Processing*, ACM, Nagoya, Japan, pp. 60–64, 2021. DOI: 10.1145/3474906.3474912.

[136] O. G. Holmberg, N. D. Köhler, T. Martins, J. Siedlecki, T. Herold, L. Keidel, B. Asani, J. Schiefelbein, S. Priglinger, K. U. Kortuem, F. J. Theis. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, vol. 2, no. 11, pp. 719–726, 2020. DOI: 10.1038/s42256-020-00247-1.

[137] J. B. Jiao, Y. F. Cai, M. Alsharid, L. Drukker, A. T. Papageorghiou, J. A. Noble. Self-supervised contrastive video-speech representation learning for ultrasound. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 534–543, 2020. DOI: 10.1007/978-3-030-59716-0_51.

[138] A. Tamkin, V. Liu, R. F. Lu, D. Fein, C. Schultz, N. Goodman. DABS: A domain-agnostic benchmark for self-supervised learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, 2021.

[139] Y. T. Xie, J. P. Zhang, Z. H. Liao, Y. Xia, C. H. Shen. PGL: Prior-guided local self-supervised learning for 3D medical image segmentation. [Online], Available: https://arxiv.org/abs/2011.12640, 2020.

[140] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Proceedings of the 34th Conference on Neural Information Processing System*, Vancouver, Canada, 2020.

[141] K. Chaitanya. Accurate medical image segmentation with limited annotations. Ph. D. dissertation, ETH Zurich, Switzerland, 2022. DOI: 10.3929/ethz-b-000533117.

[142] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Q. Qiu, D. Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1837–1848, 2022. DOI: 10.1109/TMI.2022.3150682.

[143] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Q. Qiu, D. Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 762–780, 2020. DOI: 10.1007/978-3-030-58526-6_45.

[144] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. [Online], Available: https://arxiv.org/abs/2112.09645, 2021.

[145] K. Yan, J. Z. Cai, D. K. Jin, S. Miao, D. Z. Guo, A. P. Harrison, Y. B. Tang, J. Xiao, J. J. Lu, L. Lu. SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2658–2669, 2022. DOI: 10.1109/TMI.2022.3169003.

[146] M. Sahasrabudhe, S. Christodoulidis, R. Salgado, S. Michiels, S. Loi, F. André, N. Paragios, M. Vakalopoulou. Self-supervised nuclei segmentation in histopathological images using attention. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 393–402, 2020. DOI: 10.1007/978-3-030-59722-1_38.

[147] Y. Sun, K. Gao, W. L. Lin, G. Li, S. J. Niu, L. Wang. Multi-scale self-supervised learning for multi-site pediatric brain MR image segmentation with motion/Gibbs artifacts. In *Proceedings of the 12th International Workshop on Machine Learning in Medical Imaging*, Springer, Strasbourg, France, pp. 171–179. 2021. DOI: 10.1007/978-3-030-87589-3_18.

[148] K. Y. Chen, Q. B. Wang, Y. T. Ma. Cervical optical coherence tomography image classification based on contrastive self-supervised texture learning. *Medical Physics*, vol. 49, no. 6, pp. 3638–3653, 2022. DOI: 10.1002/mp.15630.

[149] X. M. Tang, C. Zhou, L. T. Chen, Y. Wen. Enhancing medical image classification via augmentation-based pre-training. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, Houston, USA, pp. 1538–1541, 2021. DOI: 10.1109/BIBM52615.2021.9669817.

[150] H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, T. Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised Siamese networks. In *Proceedings of the 21st International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Granada, Spain, pp. 663–671, 2018. DOI: 10.1007/978-3-030-00931-1_76.

[151] H. W. Li, F. F. Xue, K. Chaitanya, S. D. Liu, I. Ezhov, B. Wiestler, J. G. Zhang. Imbalance-aware self-supervised

learning for 3D radiomic representations. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 36–46, 2021. DOI: 10.1007/978-3-030-87196-3_4.

[152] M. L. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, G. Z. Yang. Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery. In *Proceedings of the 10th Hamlyn Symposium on Medical Robotics*, Imperial College London, London, UK, pp. 27–28, 2017. DOI: 10.31256/hsmr2017.14.

[153] A. Dhere, J. Sivaswamy. Self-supervised learning for segmentation. [Online], Available: https://arxiv.org/abs/2101.05456, 2021.

[154] A. Jamaludin, T. Kadir, A. Zisserman. Self-supervised learning for spinal MRIs. In *Proceedings of the Third International Workshop on Deep Learning in Medical Image Analysis, ML-CDS: International Workshop on Multimodal Learning for Clinical Decision Support*, Springer, Québec City, Canada, pp. 294–302, 2017. DOI: 10.1007/978-3-319-67558-9_34.

[155] A. Rivail, U. Schmidt-Erfurth, W. D. Vogl, S. M. Waldstein, S. Riedl, C. Grechenig, Z. C. Wu, H. Bogunovic. Modeling disease progression in retinal OCTs with longitudinal self-supervised learning. In *Proceedings of the Second International Workshop on PRedictive Intelligence In Medicine*, Springer, Shenzhen, China, pp. 44–52, 2019. DOI: 10.1007/978-3-030-32281-6_5.

[156] G. Li, R. Togo, T. Ogawa, M. Haseyama. Self-supervised learning for gastritis detection with gastric X-ray images. [Online], Available: https://arxiv.org/abs/2104.02864, 2021.

[157] A. Tarvainen, H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceeding of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[158] F. B. Liu, Y. Tian, F. R. Cordeiro, V. Belagiannis, I. Reid, G. Carneiro. Self-supervised Mean Teacher for semi-supervised chest X-ray classification. In *Proceedings of the 12th International Workshop on Machine Learning in Medical Imaging*, Springer, Strasbourg, France, pp. 426–436, 2021. DOI: 10.1007/978-3-030-87589-3_44.

[159] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J. K. Lim, C. M. Park, J. C. Ye. Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nature Communications*, vol. 13, no. 13, Article number 3848, 2022. DOI: 10.1038/s41467-022-31514-x.

[160] C. Y. You, Y. Zhou, R. H. Zhao, L. Staib, J. S. Duncan. SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2228–2237, 2022. DOI: 10.1109/TMI.2022.3161829.

[161] C. Y. You, W. C. Dai, L. Staib, J. S. Duncan. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. [Online], Available: https://arxiv.org/abs/2206.02307, 2022.

[162] C. Y. You, W. C. Dai, F. L. Liu, H. R. Su, X. R. Zhang, X. X. Li, D. A. Clifton, L. Staib, J. S. Duncan. Mine yOur owN anatomy: Revisiting medical image segmentation with extremely limited labels. [Online], Available: https://arxiv.org/abs/2209.13476, 2022.

[163] A. Kwasigroch, M. Grochowski, A. Mikołajczyk. Self-supervised learning to increase the performance of skin lesion classification. *Electronics*, vol. 9, no. 11, Article number 1930, 2020. DOI: 10.3390/electronics9111930.

[164] H. Sowrirajan, J. B. Yang, A. Y. Ng, P. Rajpurkar. MoCo pretraining improves representation and transferability of chest X-ray models. In *Proceedings of the 4th Conference on Medical Imaging with Deep Learning*, Lübeck, Germany, pp. 728–744, 2021.

[165] A. Sriram, M. Muckley, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, W. Moore. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. [Online], Available: https://arxiv.org/abs/2101.04909, 2021.

[166] X. H. He, X. Y. Yang, S. H. Zhang, J. Y. Zhao, Y. C. Zhang, E. Xing, P. T. Xie. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. [Online], Available: https://www.medrxiv.org/content/10.1101/2020.04.13.20063941v1, 2020.

[167] Y. J. Zhu. Self-supervised learning for small shot COVID-19 classification. In *Proceedings of the 3rd International Conference on Information Technology and Computer Communications*, ACM, Guangzhou, China, pp. 36–40, 2021. DOI: 10.1145/3473465.3473472.

[168] C. J. Reed, X. Y. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. H. Zhang, D. Guillory, S. Metzger, K. Keutzer, T. Darrell. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, IEEE, Waikoloa, USA, pp. 1050–1060, 2022. DOI: 10.1109/WACV51458.2022.00112.

[169] J. Irvin, P. Rajpurkar, M. Ko, Y. F. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 590–597, 2019. DOI: 10.1609/aaai.v33i01.3301590.

[170] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, L. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, K. Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018. DOI: 10.1016/j.cell.2018.02.010.

[171] X. L. Liang, Y. Liu, J. H. Luo, Y. J. He, T. J. Chen, Q. Yang. Self-supervised cross-silo federated neural architecture search. [Online], Available: https://arxiv.org/abs/2101.11896, 2021.

[172] Z. R. Wu, S. R. Song, A. Khosla, F. Yu, L. G. Zhang, X. O. Tang, J. X. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1912-1920, 2015. DOI: 10.1109/CVPR.2015.7298801.

[173] J. C. Hu, Z. L. Feng, Y. N. Mao, J. Lei, D. Yu, M. L. Song. A location constrained dual-branch network for reliable diagnosis of jaw tumors and cysts. In *Proceedings of*

the 24th International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, Strasbourg, France, pp. 723–732, 2021. DOI: 10.1007/978-3-030-87234-2_68.

[174] Y. W. Wu, D. W. Zeng, Z. P. Wang, Y. Y. Shi, J. T. Hu. Federated contrastive learning for volumetric medical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 367–377, 2021. DOI: 10.1007/978-3-030-87199-4_35.

[175] Y. W. Wu, D. W. Zeng, Z. P. Wang, Y. Y. Shi, J. T. Hu. Distributed contrastive learning for medical image segmentation. *Medical Image Analysis*, vol. 81, Article number 102564, 2022. DOI: 10.1016/j.media.2022.102564.

[176] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, Austin, USA, pp. 429–450, 2020.

[177] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. Y. Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, USA, pp. 1273–1282, 2017.

[178] Y. Zhao, M. Li, L. Z. Lai, N. Suda, D. Civin, V. Chandra. Federated learning with non-IID data. [Online], Available: https://arxiv.org/abs/1806.00582, 2018.

[179] N. Dong, I. Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, 2021.

[180] S. Y. He, P. C. Xi, A. Ebadi, S. Tremblay, A. Wong. Performance or trust? Why not both. Deep AUC Maximization with self-supervised learning for COVID-19 chest X-ray classifications. [Online], Available: https://arxiv.org/abs/2112.08363, 2021.

[181] D. W. Zeng, Y. W. Wu, X. R. Hu, X. W. Xu, H. Y. Yuan, M. P. Huang, J. Zhuang, J. T. Hu, Y. Y. Shi. Positional contrastive learning for volumetric medical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 221–230, 2021. DOI: 10.1007/978-3-030-87196-3_21.

[182] C. Saillard, O. Dehaene, T. Marchand, O. Moindrot, A. Kamoun, B. Schmauch, S. Jegou. Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. In *Proceedings of MICCAI Workshop on Computational Pathology*, Strasbourg, France, pp. 191–205, 2021.

[183] D. Tomar, B. Bozorgtabar, M. Lortkipanidze, G. Vray, M. S. Rad, J. P. Thiran. Self-supervised generative style transfer for one-shot medical image segmentation. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, IEEE, Waikoloa, USA, pp. 1737–1747, 2022. DOI: 10.1109/WACV51458.2022.00180.

[184] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 3458–3468, 2021. DOI: 10.1109/ICCV48922.2021.00346.

[185] M. Gazda, J. Plavka, J. Gazda, P. Drotár. Self-supervised deep convolutional neural network for chest X-ray classification. *IEEE Access*, vol. 9, pp. 151972–151982, 2021. DOI: 10.1109/ACCESS.2021.3125324.

[186] O. Ciga, T. Xu, A. L. Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, vol. 7, Article number 100198, 2022. DOI: 10.1016/j.mlwa.2021.100198.

[187] B. Li, Y. Li, K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 14313–14323, 2021. DOI: 10.1109/cvpr46437.2021.01409.

[188] O. Ciga. Addressing the Data Annotation Bottleneck in Breast Digital Pathology, Ph. D. dissertation, University of Toronto, Canada, 2021.

[189] N. Mojab, V. Noroozi, D. Yi, M. P. Nallabothula, A. Aleem, P. S. Yu, J. A. Hallak. Real-world multi-domain data applications for generalizations to clinical settings. In *Proceedings of the 19th IEEE International Conference on Machine Learning and Applications*, Miami, USA, pp. 677–684, 2020. DOI: 10.1109/ICMLA51294.2020.00112.

[190] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, J. Teuwen. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images. [Online], Available: https://arxiv.org/abs/ 2107.09405, 2021.

[191] M. Ilse, J. M. Tomczak, M. Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 2127–2136, 2018.

[192] X. Zhao, S. S. Zhou. Fast mixing of hard negative samples for contrastive learning and use for COVID-19. In *Proceedings of the 4th International Conference on Big Data Technologies*, ACM, Zibo, China, pp. 6–12, 2021. DOI: 10.1145/3490322.3490324.

[193] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus. Hard negative mixing for contrastive learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1829, 2020.

[194] R. S. H. Wicaksono, A. A. Septiandri, A. Jamal. Human embryo classification using self-supervised learning. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Data Sciences*, IEEE, Ipoh, Malaysia, 2021. DOI: 10.1109/AiDAS53897.2021.9574328.

[195] S. Manna, S. Bhattacharya, U. Pal. Interpretive self-supervised pre-training: Boosting performance on visual medical data. In *Proceedings of the 12th Indian Conference on Computer Vision, Graphics and Image Processing*, ACM, Jodhpur, India, Article number 15, 2021. DOI: 10.1145/3490035.3490273.

[196] C. Y. You, R. H. Zhao, L. H. Staib, J. S. Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Singapore, pp. 639–652. 2022. DOI: 10.1007/978-3-031-16440-8_61.

[197] Q. S. Yao, Q. Quan, L. Xiao, S. K. Zhou. One-shot medical landmark detection. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 177–188, 2021. DOI: 10.1007/978-3-030-87196-3_17.

[198] H. Y. Zhou, S. Yu, C. Bian, Y. F. Hu, K. Ma, Y. F. Zheng. Comparing to learn: Surpassing ImageNet pretraining on radiographs by comparing image representations. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 398–407, 2020. DOI: 10.1007/978-3-030-59710-8_39.

[199] Y. Ali, A. Taleb, M. M. C. Höhne, C. Lippert. Self-supervised learning for 3D medical image analysis using 3D SimCLR and Monte Carlo Dropout. [Online], Available: https://arxiv.org/abs/2109.14288, 2021.

[200] F. Inglese, M. Kim, G. M. Steup-Beekman, T. W. J. Huizinga, M. A. van Buchem, J. de Bresser, D. S. KIM, I. Ronen. MRI-based classification of neuropsychiatric systemic lupus erythematosus patients with self-supervised contrastive learning. *Frontiers in Neuroscience*, vol. 16, Article number 695888, 2022. DOI: 10.3389/fnins.2022.695888.

[201] H. Zheng, J. Han, H. X. Wang, L. Yang, Z. Zhao, C. L. Wang, D. Z. Chen. Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 622–632, 2021. DOI: 10.1007/978-3-030-87193-2_59.

[202] A. Abbas, M. M. Abdelsamea, M. M. Gaber. 4S-DT: Self-supervised super sample decomposition for transfer learning with application to COVID-19 detection. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 2798–2808, 2021. DOI: 10.1109/TNNLS.2021.3082015.

[203] L. Rokach, O. Maimon, O. Arad. Improving supervised learning by sample decomposition. *International Journal of Computational Intelligence and Applications*, vol. 5, no. 1, pp. 37–53, 2005. DOI: 10.1142/S146902680500143X.

[204] C. Abbet, I. Zlobec, B. Bozorgtabar, J. P. Thiran. Divide-and-rule: Self-supervised learning for survival analysis in colorectal cancer. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Lima, Peru, pp. 480–489, 2020. DOI: 10.1007/978-3-030-59722-1_46.

[205] D. Mahapatra, B. Bozorgtabar, Z. Y. Ge. Medical image classification using generalized zero shot learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops*, IEEE, Montreal, Canada, pp. 3337–3346, 2021. DOI: 10.1109/ICCVW54120.2021.00373.

[206] M. Caron, P. Bojanowski, A. Joulin, M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 139–156, 2018. DOI: 10.1007/978-3-030-01264-9_9.

[207] L. Chaves, A. Bissoto, E. Valle, S. Avila. An evaluation of self-supervised pre-training for skin-lesion analysis.[Online], Available: https://arxiv.org/abs/2106.09229, 2022.

[208] Y. Q. Chen, C. Y. Zhang, Y. Song, N. Makris, Y. Rathi, W. D. Cai, F. Zhang, L. J. O'Donnell. Deep fiber clustering: Anatomically informed unsupervised deep learning for fast and effective white matter parcellation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Strasbourg, France, pp. 497–507, 2021. DOI: 10.1007/978-3-030-87234-2_47.

[209] O. Ciga, T. Xu, A. L. Martel. Resource and data efficient self supervised learning. [Online], Available: https://arxiv.org/abs/2109.01721, 2021.

[210] N. U. Islam, S. Gehlot, Z. W. Zhou, M. B. Gotway, J. M. Liang. Seeking an optimal approach for computer-aided pulmonary embolism detection. In *Proceedings of the 12th International Workshop on Machine Learning in Medical Imaging*, Springer, Strasbourg, France, pp. 692–702, 2021. DOI: 10.1007/978-3-030-87589-3_71.

[211] J. Gildenblat, E. Klaiman. Self-supervised similarity learning for digital pathology. [Online], Available: https://arxiv.org/abs/1905.08139, 2020.

[212] W. J. Bai, C. Chen, G. Tarroni, J. M. Duan, F. Guitton, S. E. Petersen, Y. K. Guo, P. M. Matthews, D. Rueckert. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Shenzhen, China, pp. 541–549, 2019. DOI: 10.1007/978-3-030-32245-8_60.

[213] M. F. Kragh, J. Rimestad, J. T. Lassen, J. Berntsen, H. Karstoft. Predicting embryo viability based on self-supervised alignment of time-lapse videos. *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 465–475, 2022. DOI: 10.1109/TMI.2021.3116986.

[214] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, A. Zisserman. Temporal cycle-consistency learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 1801–1810, 2019. DOI: 10.1109/CVPR.2019.00190.

[215] Q. Lu, Y. X. Li, C. Y. Ye. Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. *Medical Image Analysis*, vol. 72, Article number 102094, 2021. DOI: 10.1016/j.media.2021.102094.

[216] Q. Lu, Y. X. Li, C. Y. Ye. White matter tract segmentation with self-supervised learning. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Lima, Peru, pp. 270–279, 2020. DOI: 10.1007/978-3-030-59728-3_27.

[217] P. Y. Zhang, F. S. Wang, Y. F. Zheng. Self supervised deep representation learning for fine-grained body part recognition. In *Proceedings of the 14th IEEE International Symposium on Biomedical Imaging*, IEEE, Melbourne, Australia, pp. 578–582, 2017. DOI: 10.1109/ISBI.2017.7950587.

[218] X. B. Nguyen, G. S. Lee, S. H. Kim, H. J. Yang. Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access*, vol. 8, pp. 162973–162981, 2020. DOI: 10.1109/ACCESS.2020.3021469.

[219] J. B. Jiao, R. Droste, L. Drukker, A. T. Papageorghiou, J. A. Noble. Self-supervised representation learning for ultrasound video. In *Proceedings of the 17th IEEE International Symposium on Biomedical Imaging*, Iowa City, USA, pp. 1847–1850, 2020. DOI: 10.1109/ISBI45749.2020.9098666.

[220] T. Klinghoffer, P. Morales, Y. G. Park, N. Evans, K.

Chung, L. J. Brattain. Self-supervised feature extraction for 3D axon segmentation. *In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Seattle, USA, pp. 4213–4219, 2020. DOI: 10.1109/CVPRW50498.2020.00497.

[221] C. L. Srinidhi, S. W. Kim, F. D. Chen, A. L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, vol. 75, Article number 102256, 2022. DOI: 10.1016/j.media.2021.102256.

[222] M. Rahat Khan, A. S. M. Shafi. Statistical texture features based automatic detection and classification of diabetic retinopathy. *International Journal of Image, Graphics and Signal Processing*, vol. 13, no. 2, pp. 53–61, 2021. DOI: 10.5815/ijigsp.2021.02.05.

[223] R. Kama, K. Chinegaram, R. B. Tummala, R. R. Ganta. Segmentation of soft tissues and tumors from biomedical images using Optimized K-means Clustering via level set formulation. *International Journal of Intelligent Systems and Applications*, vol. 11, no. 9, pp. 18–28, 2019. DOI: 10.5815/ijisa.2019.09.03.

[224] X. M. Li, M. Y. Jia, M. T. Islam, L. Q. Yu, L. Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4023–4033, 2020. DOI: 10.1109/TMI.2020.3008871.

Wei-Chien Wang received the B. Sc. degree in mechatronic engineering from Huafan University, China in 2006, and the M. Sc. degree in manufacturing information and systems from the Cheng Kung University, China in 2008. She is currently a Ph. D. degree candidate in computer science at Biomedical Data Analysis and Visualisation (BDAV), School of Computer Science, The University of Sydney, Australia. She was a full-time research assistant at the Taiwan Normal University, China from 2009 to 2010. She worked as a software engineer at Hi-Lo System Research Co., Ltd., China from 2010 to 2011, and at the Software Design Center, Foxconn International Holdings, Ltd., Foxconn Technology Group, and the FIH Taiwan Design Center, Hon Hai Precision Industry Co., Ltd., China between 2011 and 2012. Since 2013, she has been a research student in Australia, working on various projects in deep learning and computer vision. She has also been a visiting researcher and lecturer with the Penghu University of Science and Technology, China since 2021.

Her research interests include visual deep learning and artificial intelligence of things (AIoT), she now focuses on self-supervised learning for medical image analysis.

E-mail: wwan7784@uni.sydney.edu.au (Corresponding author)

ORCID iD: 0000-0002-3255-7212

Euijoon Ahn received the B. Eng. degree in information technology from the University of Newcastle, Australia in 2009, and the M. Eng. degree in information technology and the M. Phil. degree in computer science from University of Sydney, Australia in 2014 and 2016, respectively. He received the Ph. D. degree in computer science from The University of Sydney, Australia in 2020. He is a lecturer at the College of Science and Engineering, James Cook University, Australia. Prior to this, he was a postdoctoral research fellow at the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Computer Science, The University of Sydney, Australia.

Dr. Ahn is a member of SICE, IEE, and IEEE. He produced top-tier publications in computer vision and medical image computing, including papers in IEEE T-MI, T-BME, JBHI, MedIA, PR, CVPR, AAAI and MICCAI. He is a regular reviewer for IEEE T-PAMI, T-MI, Nature Communications, CVPR, MICCAI and ISBI. He also works in translational health technology research, especially on health data analytics and telehealth.

His research in the development of machine learning and computer vision focuses on unsupervised and self-supervised deep learning models for biomedical image analysis, to improve image segmentation, retrieval, quantification, and classification without relying on labelled data.

E-mail: euijoon.ahn@jcu.edu.au

ORCID iD: 0000-0001-7027-067X

Dagan Feng received the M. Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University (SJTU), China in 1982, and the M. Sc. degree in biocybernetics and the Ph. D. degree in computer science from the University of California, Los Angeles (UCLA), USA in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. After briefly working as assistant professor at the University of California, USA, he joined the University of Sydney, Australia at the end of 1988, as lecturer, progressing onto professor in Department of Computer Science and head of School of Information Technologies. He is a professor emeritus at School of Computer Science, The University of Sydney, Australia, and the founding director of the Biomedical and Multimedia Information Technology (BMIT) Research Group. Prof. Feng has led more than 50 key research projects, published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He is a fellow of ACS, HKIE, IET, IEEE, and the Australian Academy of Technological Sciences and Engineering (ATSE). He has served as Chair of the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, Special Area Editor/Associate Editor/Editorial Board Member for a dozen of core journals in his area, and Scientific Advisor for a number of prestigious organizations. He has been invited to give over 100 keynote presentations in 23 countries and regions, and has organized/chaired over 100 major international conferences and symposia. He has also been appointed as Honorary Research Consultant, Royal Prince Alfred Hospital in Sydney, Australia; Chair Professor of Information Technology, Hong Kong Polytechnic University, China; Advisory Professor, Shanghai Jiao Tong University, China; Guest Professor, Northwestern Polytechnic University, China, Northeastern University, China, and Tsinghua University, China.

His research interests include biomedical systems modelling, functional imaging, biomedical information technology, and multimedia computing seeks to address the major challenges in "big data science" and provide innovative solutions for stochastic data acquisition, compression, storage, management, modelling, fusion, visualization, and communication. Currently, Prof. Feng and his research collaborators are working on new ways of improving the early detection of diseases such as cancer and de-

mentia.
E-mail: dagan.feng@sydney.edu.au
ORCID iD: 0000-0002-3381-214X

Jinman Kim received the B.Sc. (Hons.) and Ph.D. degrees in computer science from The University of Sydney, Australia in 2001 and 2006, respectively. He is an associate professor of computer science and the founding director of the Biomedical Data Analysis and Visualization (BDAV) Laboratory at The University of Sydney, Australia. He also serves as an associate director of School of Computer Science′s Biomedical and Multimedia Information Technology (BMIT) Research Group. He co-leads the "digital health imaging", as part of the Faculty of Engineering′s Digital Science Initiative, with the vision and strategy to improve the use and accessibility of medical imaging via AI innovations. Since 2006, he has been a research associate with the university′s leading teaching hospital, the Royal Prince Alfred Hospital. From 2008 to 2012, he was an ARC postdoctoral research fellow, with one year leave from 2009 to 2010 to join the MIRA Lab Research Group, Switzerland, as a marie curie senior research fellow. Since 2013, he has been with School of Computer Science, The University of Sydney where he was a senior lecturer, and was promoted to associate professor in 2016. He continuously publishes in top venues in his field and has received multiple competitive grants and scientific recognition. He is actively involved in his research communities where he is the vice president of the Computer Graphics Society (CGS), A/Editor of *Computer Methods and Program in Biomedicine (CMPB)*, A/Editor of *The Visual Computer* (TVCJ), and Reviewer for all major journals and conferences in his field. He has actively focused on research translation where he has worked closely with clinical partners to take his research into clinical practice. He is the research director of the Nepean Telehealth and Technology Centre (NTTC) at Nepean Hospital, NSW Health, responsible for translational telehealth and digital hospital research. Some of his research has been developed into clinical software that is being used at multiple hospitals. His work on telehealth has been recognized with multiple awards, including the 2016 Health Secretary Innovation Award from the NSW Ministry of Health.

His research interests include machine learning, biomedical image analysis and visualization, especially multimodal data processing, image-omics, and image data correlation to other health data.

E-mail: jinman.kim@sydney.edu.au
ORCID iD: 0000-0001-5960-1060