Self-supervised learning methods and applications in medical imaging analysis: a survey

Saeed Shurrab and Rehab Duwairi

Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan

ABSTRACT

The scarcity of high-quality annotated medical imaging datasets is a major problem that collides with machine learning applications in the field of medical imaging analysis and impedes its advancement. Self-supervised learning is a recent training paradigm that enables learning robust representations without the need for human annotation which can be considered an effective solution for the scarcity of annotated medical data. This article reviews the state-of-the-art research directions in self-supervised learning approaches for image data with a concentration on their applications in the field of medical imaging analysis. The article covers a set of the most recent self-supervised learning methods from the computer vision field as they are applicable to the medical imaging analysis and categorize them as predictive, generative, and contrastive approaches. Moreover, the article covers 40 of the most recent research papers in the field of self-supervised learning in medical imaging analysis aiming at shedding the light on the recent innovation in the field. Finally, the article concludes with possible future research directions in the field.

INTRODUCTION

Medical image analysis is mainly concerned with processing and analyzing medical images from different modalities to extract useful information that help in making precise diagnoses (*Anwar et al., 2018*). Medical image analysis falls into four main tasks which emerged from computer vision tasks and were tailored for the medical filed. These four tasks are classification, detection and localization, segmentation and registration (*Altaf et al., 2019*). Each of the mentioned tasks has its own methods and algorithms that help in understating and extracting useful information from the medical images.

The recent advancements in the artificial intelligence (AI) field brought significant improvements into the medical image analysis field by transforming it from a heuristic-based into a learning-based approach (*Ker et al., 2017*). To elaborate, learning-based analysis approaches aim at extracting useful information (features) that represent the input images in a way that fits the target medical image analysis task. In addition, features extraction can be accomplished manually (engineered) or automatically (learned) from

Corresponding author
Rehab Duwairi, rehab@just.edu.jo

the data (*Sarhan et al., 2020*). While manual feature extraction is the main concern of the Statistical Machine Learning field, the Deep Learning field is mainly concerned with the automatic extraction of features and it is highly preferred.

A convolutional neural network (CNN) is an example of deep learning models which deals with grid-based data such as images to learn the latent features in a hierarchical fashion from the fine level (lines and edges) to the complex level(objects). Mainly, seven types of layers constitute the structure of CNN, namely, convolutional layer, pooling layer, activation layer, fully connected layer, upsampling layer, dropout layer and batch normalization layer (*Yamashita et al., 2018*). While both convolutional and pooling layers are responsible for features' extraction and aggregation, activation layer is responsible for non-linear transformation. The fully connected layer is responsible for mapping the learned features into an output vector of a certain dimension in case of classification tasks. In other cases, such as dense prediction, a transposed convolutional block is employed by the CNN which acts as upsampling layer, which is responsible for mapping the learned features into an output array of certain dimension (*Zeiler et al., 2010*). Lastly, both dropout and batch normalization layers are responsible for regularization. The process of optimizing the learnable layers in CNNs is accomplished through the gradient descent algorithm and its variants which aim at minimizing the difference between the network's output and the ground truth labels (*i.e.,* minimize a loss function).

CNNs have become a popular choice in the field of medical image analysis and provided a tremendous progression into the various medical image analysis tasks due to their ability to deal with images in their raw formats; and the performance they provide which can be compared to the human performance at faster rates. Yet, CNNs are known to have an enormous number of trainable parameters to be estimated, usually in millions, to capture the underlying distribution in the input data. As a result, a relatively large amount of annotated data is required to achieve a better estimation of these parameters and enable performing supervised training (*Mitchell, 2021*).

Despite the remarkable success that CNNs have achieved in the medical image analysis field, there are some obstacles that hamper their advancement. Initially, building a large enough annotated medical dataset of high quality is expensive and time-consuming. In addition, unlike the natural scene image data which may be annotated by less skilled personnel, medical datasets require expert personnel with domain knowledge to accomplish the annotation process. Moreover, the annotation process is prone to patients' privacy issues especially when working with specific disorders (*Taleb et al., 2020*). Collectively, these factors render annotated data scarcity in terms of annotation and volume a major obstacle for machine learning applications in the medical field.

As an alternative solution, the concept of transfer learning came to the top of the table for situations where the amount of annotated data is relatively small. Transfer learning is the process of employing the knowledge that has been learned in a source task to another target task to improve the generalization and the performance (*Goodfellow et al., 2016*; *Torrey & Shavlik, 2010*). The most common form of transfer learning, in the machine learning community, is built upon pre-trained state of the art models such as VGG (*Simonyan & Zisserman, 2015*), GoogleNet (*Szegedy et al., 2015*), ResNet

(*He et al., 2016a*) and DenseNet (*Huang et al., 2017*) which are trained on the giant labeled image datasets such as ImageNet (*Deng et al., 2009*). ImageNet includes approximately 14 million natural images that belong to 22,000 visual categories and 1,000 labels (*Krizhevsky, Sutskever & Hinton, 2012*).

The employment of pre-trained models on ImageNet for medical applications is a controversial issue for three reasons. Firstly, the extracted features from the natural images domain may not be a good representation in the medical field due to the remarkable difference in features' distribution, resolution, and number of output labels between both domains. Secondly, ImageNet pre-trained models are over-parameterized models in terms of number of parameters when utilized for medical images analysis tasks. More clearly, ImageNet pre-trained models are designed to predict 1,000 labels which makes them require a larger number of parameters, especially in the last layers to fit the 1,000 classes. On the other hand, in the case of medical images, the number of classes may not exceed 10 classes, and hence, smaller models can be sufficient (*Holmberg et al., 2020*; *Raghu et al., 2019*). Thirdly, ImageNet pre-trained models are primarily trained on 2D images while the vast majority of medical imaging modalities are 3D such as CT, MRI, and OCT. This renders the pre-trained models on the ImageNet dataset an infeasible solution. Despite that, a set of guidelines exists that mainly depends on the target dataset size and domain similarity when dealing with ImageNet pre-trained models for different domains (*Karpathy et al., 2016*). Other approaches have been proposed to overcome such problems where Self-Supervised Learning is one of them.

Self-supervised learning is a recent learning paradigm that enables learning semantic features by generating supervisory signals from a pool of unlabeled data without the need for human annotation (*Chen et al., 2019*). The learned features from self-supervised learning are used for subsequent tasks where the amount of the annotated data is limited. From the unsupervised learning perspective, the self-supervised learning approach omits the need for manually annotated data, while the supervised perspective in the self-supervised learning approach is represented in model training with labels generated from the data itself (*Liu et al., 2021*).

Two tasks characterize the learning pipeline in the self-supervised learning approach which are the pretext task and downstream task. In the pretext task where the self-supervised learning actually occurs, a model is learned in a supervised fashion using the unlabeled data by creating labels from the data in a way that enables the model to learn the useful representation from the data. In the downstream task, the learned representations from the pretext task is transferred as initial weights to the downstream task to accomplish its intended goal (fine-tuning) (*Holmberg et al., 2020*). Figure 1 depicts the main workflow of the self-supervised learning approach.

Self-supervised learning became a popular choice in the field of medical image analysis where the amount of the available annotated data is relatively small, while the unlabeled data is comparatively large. Several researches have demonstrated the effectiveness of the self-supervised learning approach throughout various medical image analysis tasks such as detection and classification (*Lu, Chen & Mahmood, 2020*; *Li et al., 2021*; *Sriram et al., 2021*), detection and localization (*Chen et al., 2019*; *Sowrirajan et al., 2021*;
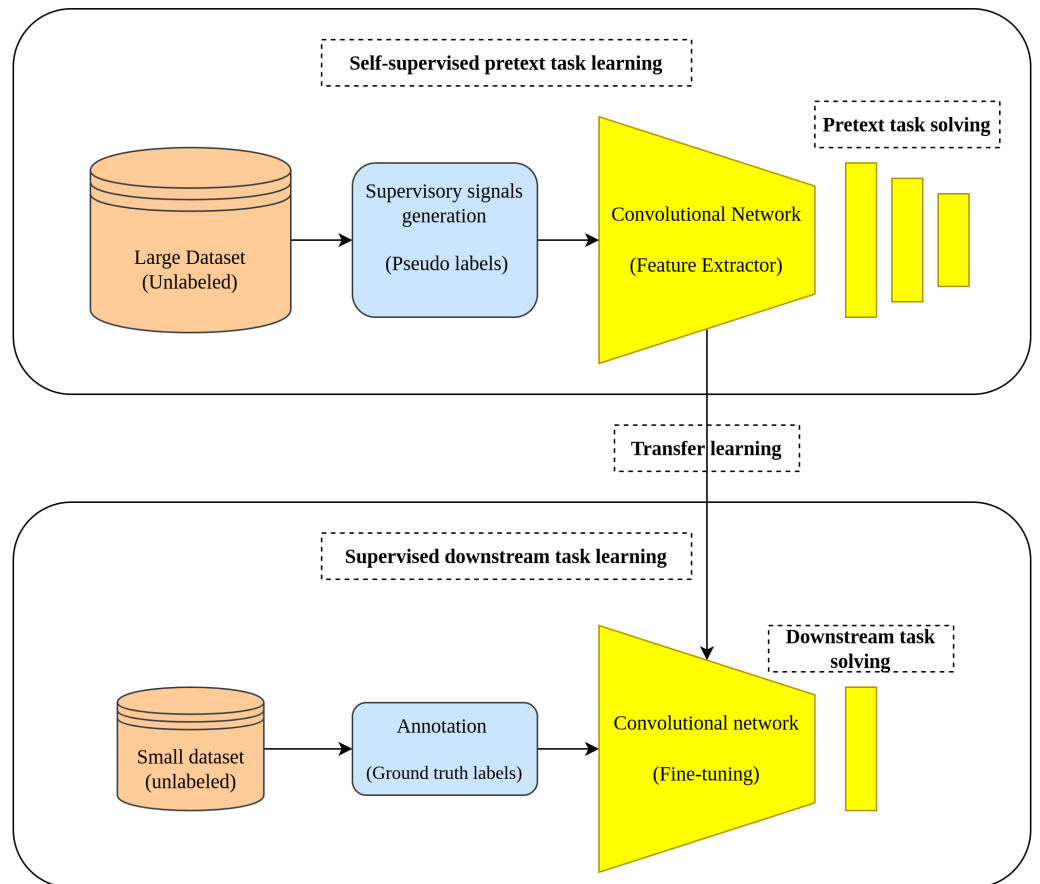
Figure 1 Self-supervised learning main workflow. (top): Self-supervised learning scheme is applied by training an auxiliary task using pseudo labels generated from a large unlabeled dataset. (bottom): The learned representations are transferred from the pretext task to the down-stream task to accomplish the training on small amount of data with ground truth labels.

Nguyen et al., 2020), and segmentation tasks (*Taleb et al., 2020*; *Xie et al., 2020*; *Chaitanya et al., 2020*).

This paper aims at reviewing the state-of-the-art research directions in self-supervised learning approaches for image data with a concentration on their applications in medical images analysis. Annotated data scarcity is a major problem that hampers the advancement of machine learning applications in the medical field, and self-supervised learning can act as an effective solution for such a problem. Our main goal, in this paper, is to shed the light on the recent innovations in the field of self-supervised learning in medical imaging analysis by providing a high-quality overview of the recently developed methods in the field to enable the reader to be familiar with such approach which is interesting and quickly becoming the choice for several researchers in the machine/deep learning field.

The prospective audience of this article includes, in the first place, machine/deep learning researchers and practitioners in medical images analysis and computer vision fields. Further, researchers and practitioners from the medical field who are interested in

medical imaging analysis *via* machine learning approaches form a second group of the prospective audience. Lastly, any reader with an interest in machine learning applications, in general, is considered as the third group of the prospective audience. It is worthy to note that this survey was presented in a simplified manner to fit the various groups of the prospective audiences.

Various research works, in the literature, have concentrated on self-supervised learning in computer vision *per se* such as *Jing & Tian, (2020)*, *Liu et al. (2021)*, *Ohri & Kumar (2021)*, *Jaiswal et al., (2021)*, while other researches briefly reviewed the role of self-supervised learning in the analysis of medical images as part of deep learning applications in medical image analysis such as *Tajbakhsh et al. (2020)*, *Chen et al. (2021a)*. To the best of our knowledge, this is the first survey on self-supervised learning applications in the field of medical images that aims at bridging the gap between computer vision and medical imaging fields. The key contributions of this paper can be summarized as follows:

- We provided a high-level overview of the state-of-the-art self-supervised learning methods in the computer vision field as they are general-purpose methods that can be used in the medical context. Further, we categorized these methods as predictive, generative, and contrastive self-supervised methods.
- We covered and provided a high-level overview for a list of the 40 most recent and impactful research works in the field of self-supervised learning in medical imaging analysis. In addition, we categorized these works in the same way we categorized the computer vision tasks. Further, we included an additional category called multiple-tasks/multi-tasking to fit those researches that utilized multiple tasks simultaneously.
- We developed a GitHub repository (https://github.com/SaeedShurrab/awesome-self-supervised-learning-in-medical-imaging) called Awesome Self-Supervised Learning in Medical Imaging that would serve as a resource for the literature in the field which will be updated continuously.

The rest of this survey is organized as follows: the 'Survey Methodology' summarizes the literature selection methodology. 'The Self-Supervised Learning Approaches' provides an in-depth overview of the self-supervised learning approach and its methods. The 'Self-Supervised Methods in Medical Imaging' reviews the recent self-supervised learning methods in medical imaging analysis. The 'Performance Comparison' compares the performance of the discussed self-supervised learning in medical imaging. The 'Discussion and Future Research Directions' highlights some open challenges and the possible future research directions in the field, while the last section concludes the paper. Lastly, Appendix A lists the available implementation codes of the discussed research throughout this paper.

SURVEY METHODOLOGY

This section summarizes the methodology followed, by the authors, to search for relevant literature on self-supervised learning applications in medical imaging analysis. This methodology includes the determination of literature sources, search keywords, inclusion/exclusion criteria, and papers selection criteria.

Sources and keywords

The first step in our methodology is to select the main sources of literature that will be used. As a result, we considered three bibliographic databases as primary sources of literature, namely:

- IEEE Explore (http://ieeexplore.ieee.org/)
- ScienceDirect (https://www.sciencedirect.com/)
- Springer Link (http://link.springer.com/)

We focused our literature search on these resources as they include reputable journals and conferences that are mainly concerned with machine learning applications in medical imaging. On the other hand, we considered two additional sources of literature as secondary sources which are:

- ArXiv Preprints (https://arxiv.org/).
- The related works sections in the selected papers.

For searching keywords, we opted the terms *self-supervised learning in medical imaging*, *pretext tasks in medical imaging*, *representation learning in medical imaging* and *contrastive learning in medical imaging* to investigate the selected resources.

Inclusion/exclusion criteria

Initially, we explored the literature in the field of self-supervised learning in medical image computing over the period 2017–2021, as this is the period where self-supervised learning started to creep into medical imaging analysis, with a high emphasis on the research works from the period 2019–2021 and excluded any other works outside this period. Further, we examined the titles and abstracts of the research articles resulting from querying the selected resources to judge the relevance of search results. As a result, we considered only research works that either have adopted a self-supervised learning approach directly to solve medical imaging tasks or presented a novel self-supervised learning approach in medical imaging that has not been seen before to our knowledge and excluded any other works of less relevance to our target. For self-supervised learning approaches from the computer vision field, we first explored the selected self-supervised learning research in medical imaging analysis literature and selected those methods that have been frequently used in the medical field even if they are not within the predefined period. We further added some additional state-of-the-art methods that have not been explored directly in the medical context and excluded any other methods. In addition, we kept refining our search results by selecting research articles that are published in journals or conferences with an impact factor of 3 or greater and excluded any other works published in venues with less impact factor than our threshold. For ArXiv preprints, we considered only those works cited in the selected published papers and excluded any other works. We further examined the affiliation and the research portfolio of the authors of these preprints before including their works. We also considered research works from outside the selected sources gathered by exploring the related works sections of the selected papers that are directly relevant to our target.
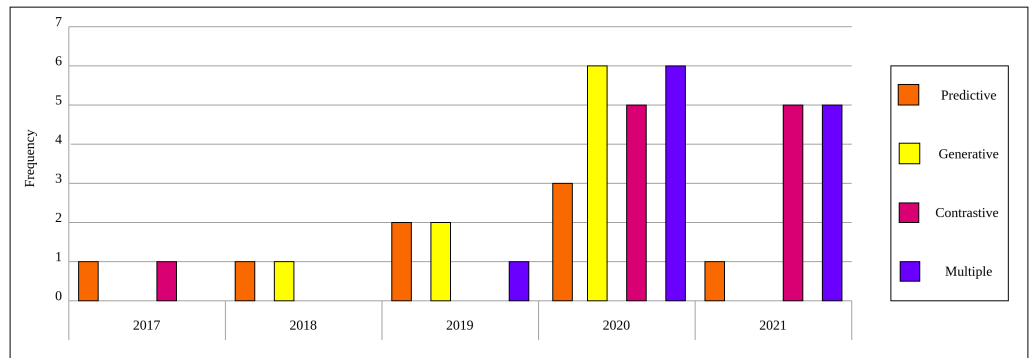
Figure 2 Distribution of selected publications by year and category for self-supervised learning in medical imaging.

Papers selection

As a result of the predefined inclusion/exclusion criteria, we settled on 15 self-supervised learning approaches that have been developed on natural images and exploited in the medical context. For self-supervised learning in medical imaging, we settled on 40 papers that relate directly to self-supervised learning applications in medical imaging analysis. Each of the selected papers has been reviewed thoroughly and a high-level overview is developed that focuses on the innovation in the self-supervised learning approach and presented throughout this survey. Figure 2 depicts the distribution by year and category for the 40 papers in the field of self-supervised learning in medical imaging.

SELF-SUPERVISED LEARNING APPROACHES

The formulation of early self-supervised learning concepts appear in the work of *Bengio et al. (2007)* by training deep neural networks in an unsupervised greedy layer-wise fashion. The authors trained a single-layer auto-encoder for each layer one at a time (self-supervised learning). After training each layer in the network separately, the resulting weights of each layer are used as initial weights to train the whole network on the target task (fine-tuning). One of the prominent downsides of the greedy layer-wise training approach is the inability to secure a complete optimal solution by grouping sub-optimal ones (*Goodfellow et al., 2016*). Further, the greedy layer-wise approach has been obsoleted by the emergence of end-to-end deep neural models that can be trained in a single run (*Mao, 2020*). Despite that, the greedy layer-wise methodology formed the nucleus for what so-called nowadays self-supervised learning approach and opened the door for its applications in computer vision, natural language processing, robotics, and other fields.

Pretext tasks play a central role in the self-supervised learning approach and act as its backbone. While the downstream task may differ according to the researchers' needs and targets, the pretext task can be common among different downstream tasks. For example, the same pretext task, *e.g.*, convolutional auto-encoder, could be used to learn visual features for two different downstream tasks with different data. This property makes it helpful to categorize self-supervised learning approaches according to the nature of the

pretext task. In this regard, we categorize self-supervised learning pretext tasks into three main categories including predictive, generative, and contrastive tasks. Such categorization aims at simplifying and grouping similar approaches together which in turn enables achieving a better understanding of the methods of each category. The upcoming sections introduce the reader to the most prominent methods for each category.

Predictive self-supervised learning

The predictive self-supervised learning approach aims at learning robust representations from unlabeled data by treating the pretext task as a classification or regression problem. More clearly, each unlabeled image is assigned a pseudo label, these labels are generated from the data itself, which can be either categorical or numerical depending on the pretext task design specifications. As a trivial example, applying a certain transformation to the input image can be considered a pseudo label. Consequently, the role of the pretext task is to predict this pseudo label correctly. It is worthy to note that pseudo labels must be carefully generated in order to enable learning robust representations from the data. Many predictive pretext tasks have been designed in the field of computer vision, the next sections illustrate in detail some of these methods.

Exemplar CNN

Exemplar CNN is one of the earliest predictive self-supervised pretext models which was proposed by *Dosovitskiy et al. (2015)*. Learning a good representation of the input data throughout the exemplar CNN method is hypothesized by the model's robustness to the applied transformations. To achieve this, a synthesized training dataset is created. This dataset consists of patches of objects or parts of the object with a size of 32×32 pixels which are cropped from the original images and they are called the exemplary patches. Following that, a set of predefined transformations including translation, scaling, rotation, contrast, and color adjustment are applied randomly to each generated patch as shown in Fig. 3. Consequently, each seed patch along with its applied transformations forms a surrogate class in the training dataset. Following that, a convolutional neural network is trained to learn useful representations by learning to discriminate between the different surrogate classes in the synthesized dataset.

Relative position prediction

Relative position prediction is another predictive pretext task proposed by *Doersch, Gupta & Efros (2015)* that is inspired by the word embedding Skip-Gram model (*Mikolov et al., 2013*) in natural language processing field. The main hypothesis of learning representations by relative position prediction is to understand the spatial context of the objects in the input image. The implementation details include dividing the input image into a 3×3 grid of patches as shown in Fig. 4. The central patch is considered an anchor patch, while the remaining eight patches are considered query patches. To increase the complexity and reduce the chance of learning shortcuts such as texture continuity and boundary patterns, a set of solutions was introduced including the addition of gaps and jitters to the patches, color channel processing by shifting certain channels to the gray-scale or partial channel dropping to avoid the chromatic aberration effect. Consequently, a late-fusion
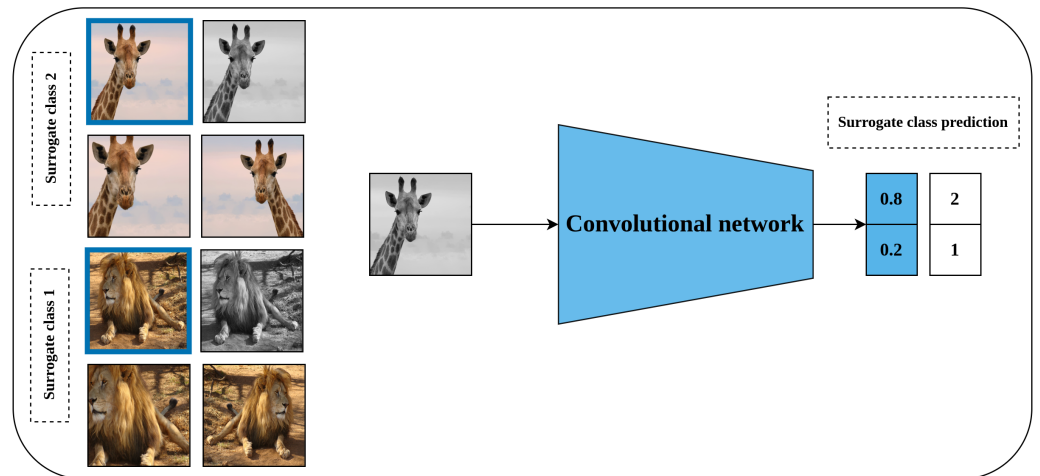
Figure 3 Illustration of the generation of surrogate classes for self-supervised features' learning with exemplar CNN. (left): The marked patch in blue represents exemplary patch cropped from a certain image in unlabeled dataset to serve as a seed for the surrogate class. The remaining patches are a set of random augmentation operations applied to the seed patch to generate multiple images for the same surrogate class. (right): A convolutional model is employed to learn representation by classifying the generated images into the specified surrogate classes. Image credit; upper: Frans Van Heerden, lower: Gary Whyte.

convolutional model is trained on a randomly sampled pair of patches of the central patch and query patch to predict the relative position of query patches with respect to the central patch.

Jigsaw puzzle

Solving a Jigsaw puzzle is another pretext task proposed by *Noroozi & Favaro (2016)* and inspired by the earlier work of *Doersch, Gupta & Efros (2015)* for relative position prediction. To solve a Jigsaw puzzle, a convolutional neural network is required to learn to restore a set of disordered patches, *e.g.*, nine patches, to their original spatial arrangement. For this purpose, a special convolutional network called Context-Free Networks (CFN) with Siamese architecture and shared weights was proposed by the authors as shown in Fig. 5. To train the network, a shuffled image with a random permutation of the nine patches is fed to the network. But, for nine patches there is $9! = 362, 880$ possible permutations. To avoid such a large solution space, the authors limit the number of permutations to a predefined set of permutations with a certain index for each permutation. Lastly, the defined architecture's role is to produce a likelihood vector over the set of predefined indices that maximize the probability of the input permutation.

Rotation prediction

Rotation prediction was first proposed by *Komodakis & Gidaris (2018)* to learn visual representations in a self-supervised fashion. The main idea behind the rotation prediction task is to learn a convolutional model that can recognize the applied geometric transformation on the input image as shown in Fig. 6 in a simple classification problem. Geometric transformations are represented by applying rotation angles by multiple of 90°
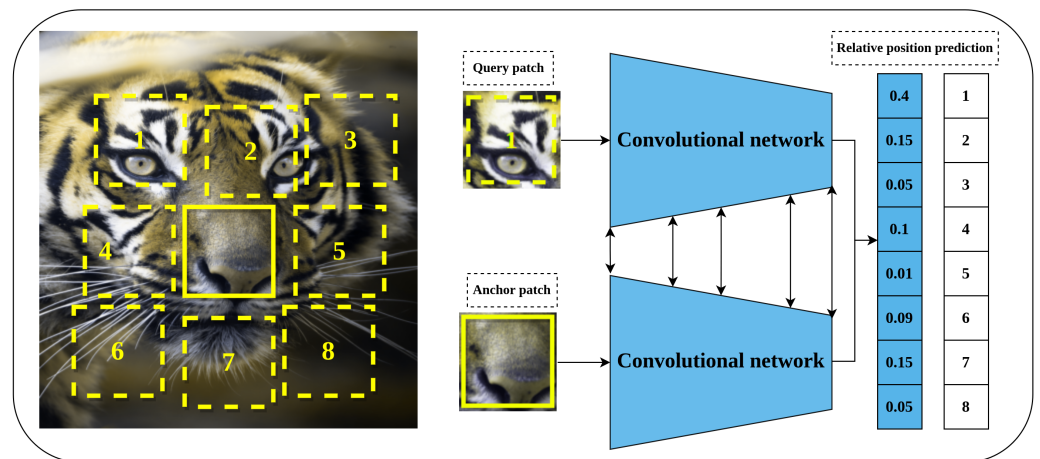
Figure 4 **Illustration of self-supervised learning by relative position prediction task.** (left): An image is divided into nine patches where the central patch (the one without number) represents the anchor patch and the remaining eight patches (delineated in dashed yellow lines) represent the query patches. (right): a training example that consists of an anchor patch and query patch is passed to a late-fusion convolutional model which share weights between the two branches to predict the position of the query patch with respect to anchor patch. Image credit: Gabriele Brancati.
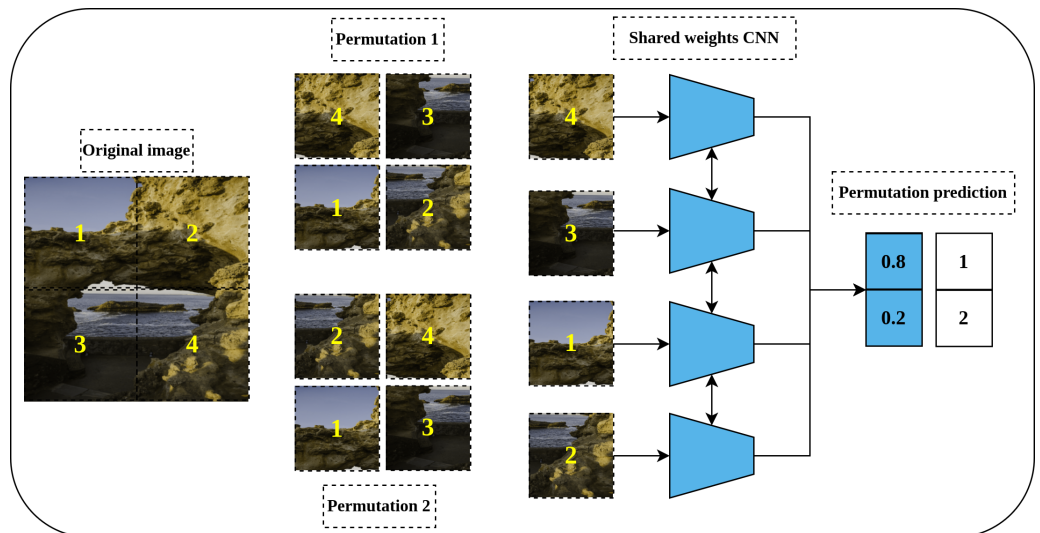
Figure 5 **Illustration of a Jigsaw puzzle pretext task.** (left): The Puzzle generation steps where an image is cropped into a set of patches that constitute the main blocks of the puzzle. The generated patches are shuffled according to a predefined set of permutations where each permutation has a specific index (permutation number). (right): a Siamese network, with shared weights, takes the shuffled patches as input according to certain permutation and classifies them to the respective permutation index. Image credit: Mathilde Langevin.
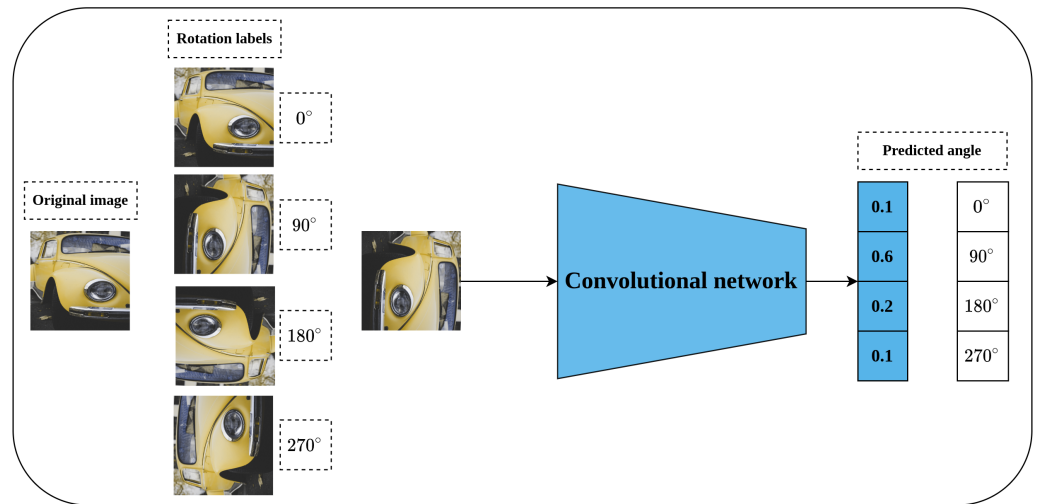
Figure 6 Illustration of rotation prediction pretext task. (left): Supervisory signals are generated from the data by applying a rotation angle in the range (0°, 270°) with multiple of 90° degrees to the input image. (right): The role of the network is to distinguish the applied rotation on the input image. Image credit: Lilartsy.

to the input image which may fall into one of four categories including [0°, 90°, 180°, 270°]. The main intuition behind the rotation prediction task is that enabling the convolutional network to learn to recognize the applied rotation to the input image is directly linked to the model's ability to learn the prominent objects in that image. To achieve this, the model needs to recognize the type and orientations of these objects in relation to the dominant geometric transformation to correctly learn the applied rotation. The same concepts hold for the human way of recognizing the rotation applied to a certain object in an image. For instance, to recognize a chair image which was rotated by 90°, a human needs to recognize the chair legs, base, back and their orientations. This way, rotation prediction enables learning semantic features by recognizing the orientations of images.

Generative self-supervised learning

Generative self-supervised learning approach aims at learning the latent features in the input data by treating pretext tasks as a generative problem. The intuition behind generative pretext tasks is that the model can learn useful representations from unlabeled data by learning to regenerate the same input data or by learning to generate new examples from the same distribution of the input data. Generally, auto-encoder-based architectures generative adversarial networks are utilized in this category. Several generative pretext tasks have been proposed in the literature, the next sections illustrate, in detail, a few of these methods.

Denoising auto-encoders

Auto-encoders are special neural models whose main task is to reconstruct their input (*Goodfellow et al., 2016*). The basic auto-encoder consists of two parts, namely, the encoder network and the decoder network. The encoder network plays the role of compressing the network's input into a latent dimensional space, while the decoder's role is to reconstruct
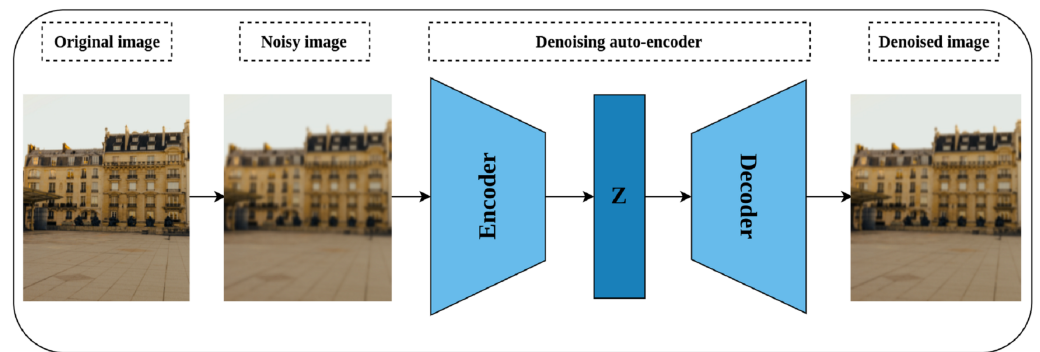
Figure 7 Illustration of self-supervised features' learning using image denoising. (left): A noisy image is created by injecting noise to the original image. (middle) an auto-encoder model learns representations by compressing the noisy image into a latent space (Z) *via* the encoder network, while the decoder tries to reconstruct the compressed image from the latent space. (right): A denoised image close to the original image. Image credit: Celine.

the compressed input from the latent space (*Tschannen, Bachem & Lucic, 2018*). After training the network, the decoder is discarded while the encoder is kept for further processing. Denoising auto-encoders are special models of auto-encoders proposed by *Vincent et al. (2008)* for representation learning through learning to reconstruct a noise-free output from noisy input. As shown in Fig. 7, a noisy version of the original image is created by introducing certain types of noise including but not limited to Gaussian noise, Poisson noise, Uniform noise, and Impulsive noise. The noisy image is then passed to the auto-encoder to reconstruct the original image from the noisy image by minimizing the reconstruction loss. The intuition behind denoising auto-encoder is related to the human ability to correctly recognize the object type, in an image, even if a certain part of it is partially corrupted. This situation is true as long as the partial corruption does not affect the global view of the object. For a convolutional model, learning robust representations is linked to the model's ability to learn the semantic features that will enable restoring the original image from a noisy version.

Image inpainting

Image inpainting or context encoder is a generative self-supervised pretext task proposed by *Pathak et al. (2016)* that aims to learn rich representations by fill-in-the-blank strategy. The intuition behind image inpainting is directly related to the human ability to complete the missing part of the image by observing the patterns in the surrounding pixels. Technically, part of the input image is cropped or masked, rather than introducing noise to it, and the role of the network is to complete the cropped part. Further, three forms of masking are proposed including central block, random blocks, and random region. An auto-encoder network and channel-wise fully connected latent space is employed for this task as shown in Fig. 8. In addition, a combined loss function that integrates both reconstruction loss and adversarial loss (*Goodfellow et al., 2014*) is optimized throughout the training. The reconstruction loss, L2, is meant to hold the overall structure of the input image and the
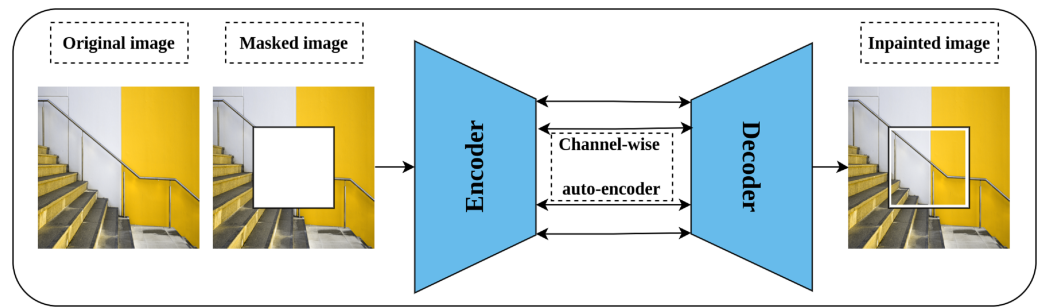
Figure 8 Illustration of context encoder model for Self-supervised features' learning. (left): An input image is modified by masking part of the image. (right): The context encoder learns useful representations by reconstructing the missing part in the masked image by minimizing the reconstruction and adversarial loss. Image credit: Sam.

masked part, while the adversarial loss aims to improve the appearance of the predicted masked part.

Image colorization

Generation of a colorized image from a gray-scale one was proposed by *Zhang, Isola & Efros (2016)* as a solution for automatic image colorization problem and self-supervised pretext task simultaneously. Lab color space is employed in this task rather than the RGB color space as it reflects the human color perception where the L channel represents the grayscale, while the a and b channels represent the color channels. Consequently, a convolution network is trained by taking the L channel as an input, and the channels a and b as supervisory signals—where the role of the network is to produce the input image in Lab color space as shown in Fig. 9. Nonetheless, image colorization is multi-modal in nature which means that the same object may have different valid colors *e.g.*, apple may be yellow, red, or green but not other colors. To compensate for this issue, the network is designed to predict the probability distribution of the possible colors for each pixel. In addition, a weighted cross-entropy Loss function is utilized to compensate for rare colors. Then, the annealed mean of the probability distribution is computed to produce the final colorization. The intuition behind the colorization task is that understanding the coloring scheme of the objects in the input images will result in learning rich representations about them.

Split-brain auto-encoder

Split-brain auto-encoder is another pretext task proposed by *Zhang, Isola & Efros (2017)* and extended their earlier work on image colorization. The main idea behind the split-brain auto-encoder is to obtain useful representations by learning to generate a portion of the data from the remaining data. By translating this idea to the image data in Lab* color space, the gray-scale channel L can be generated from the color channels a and b and vice versa. This process is accomplished through modifying the traditional auto-encoder architecture by adding two splits to the network as shown in Fig. 10—where each disjoint split learns the underlying representations from the input data as described previously. Eventually,
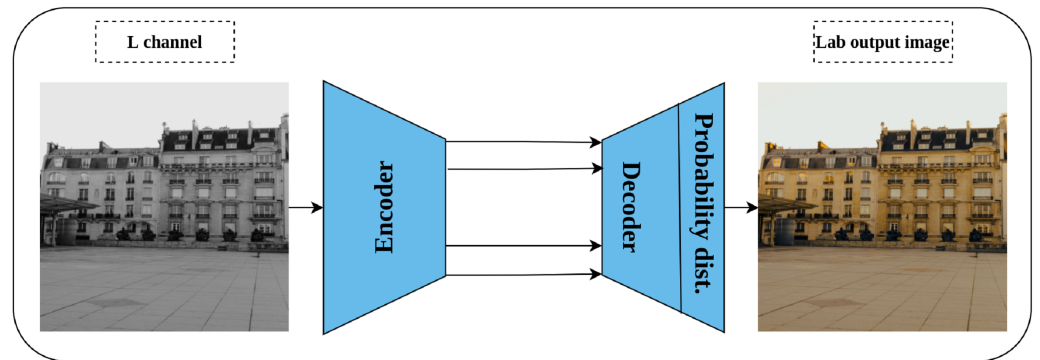
Figure 9 **Illustration of image colorization pretext task.** An encoder–decoder model is trained to predict the colored image from a gray scale image. The input is the L channel in Lab color space, while the channels a and b are used as supervisory signals. The last block indicates the color probability distribution for each pixel in the output image. Image credit: Celine.
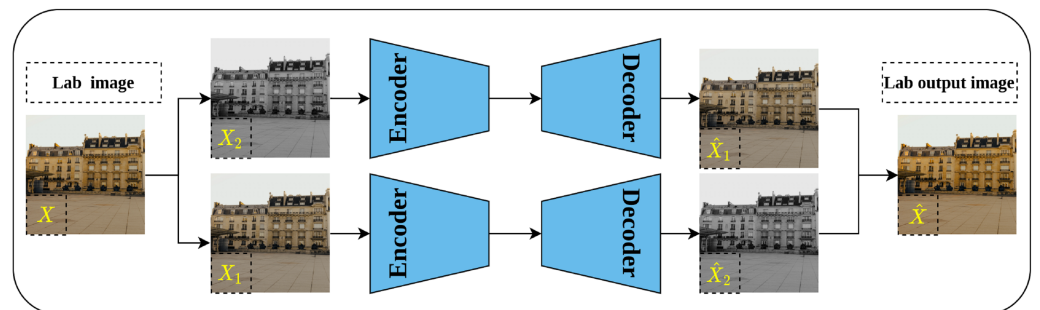
Figure 10 **Illustration of split-brain auto-encoder pretext task.** The input image X is separated by channels as color channels X_1 and gray-scale channel X_2. Two disjoint networks F_1 and F_2 are trained to predict the missing components in their inputs. F_1 predicts the gray-scale channel $\hat{X_2}$ from the color channels X_1, while F_2 predicts the color channels $\hat{X_1}$ from the gray-scale channels X_2. The outputs of both networks are grouped to produce the recolored image $\hat{X}$. Image credit: Celine.

the output of both splits is aggregated throughout concatenation to produce the final output of the network. The authors stated that learning from both gray-scale and color channels simultaneously rather than single-channel as in colorization problems would enable better representations learning. This is because the split-brain architecture is able to learn color-related information which is not the case in the colorization task which learns features only from gray-scale input.

Deep Convolutional GAN

Generative adversarial networks (GAN) are a class of deep learning generative models that use random noisy input to generate new data which mimics the real training data. Typically, GAN architecture consists of two networks, namely, the generator network and the discriminator network. The role of the generator is to convert the random noisy input into an imitation of the real data while the role of the discriminator is to distinguish
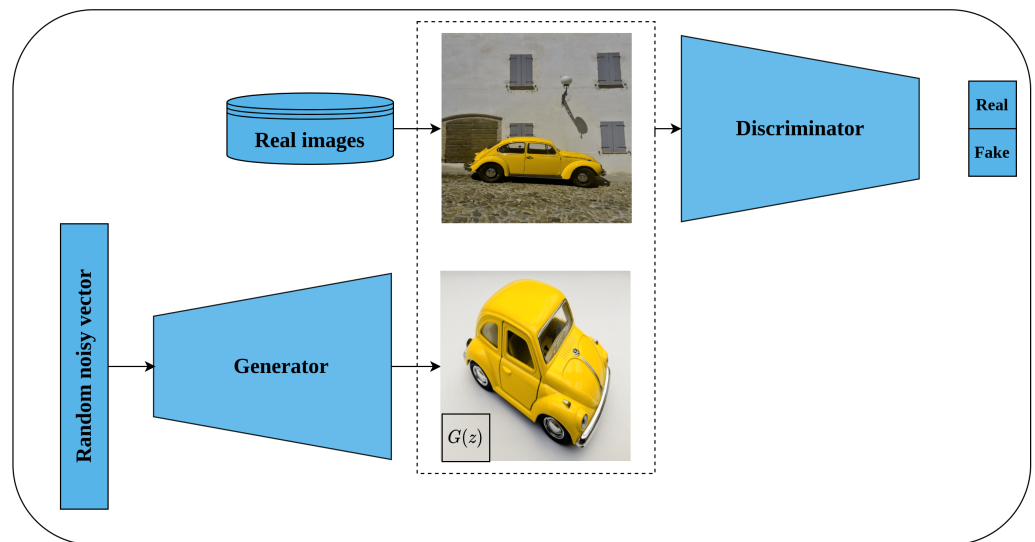
Figure 11 **Illustration of deep convolutional GAN architecture.** (left): A generator network tries to generate fake images using random noisy vector. (right): A discriminator network takes the generated images from generator network as well as real images from the same distribution and classifies them as real or fake until being not able to discriminate both sources. Image credit; upper: Beatrice Gemmi, lower: Mike B.

whether the generator output is real or fake. Both networks are trained in a competing way until the discriminator is not able to differentiate between real and fake images (*Goodfellow et al., 2014*).

Deep convolutional GAN, or DCGAN for short, is an extension of GAN proposed by *Radford, Metz & Chintala (2016)* as an unsupervised representations learning architecture for image data. DCGAN is considered the first successful attempt to scale GAN with convolutional neural networks as opposed to the earlier work of *Goodfellow et al. (2014)* which is based on multi-layer perceptron architecture. Further, the authors provided architectural guidelines for designing a stable DCGAN, such as replacing the pooling layer with a stridden convolutional layer for discriminator, and fractionally strided convolution for generator. Also, employing batch normalization (*Ioffe & Szegedy, 2015*) in the generator and discriminator networks, removing fully connected layers, using ReLU activation (*Nair & Hinton, 2010*) for all generator layers except the output layer which is Thanh activation. LeakyReLU activation (*Maas et al., 2013*) was recommended for all layers in the discriminator network. Figure 11 depicts the generator network architecture as designed by the authors. The authors evaluated the quality of the learned features by DCGAN discriminator by performing an image classification task which showed superior performance in comparison to other unsupervised methods and opened the door for exploiting GAN-based models in pretext tasks.

Bi-directional GAN

Bi-directional GAN (BiGAN) is another generative unsupervised learning architecture proposed by *Donahue, Krähenbühl & Darrell (2016)* that extended the earlier work of

PeerJ Computer Science
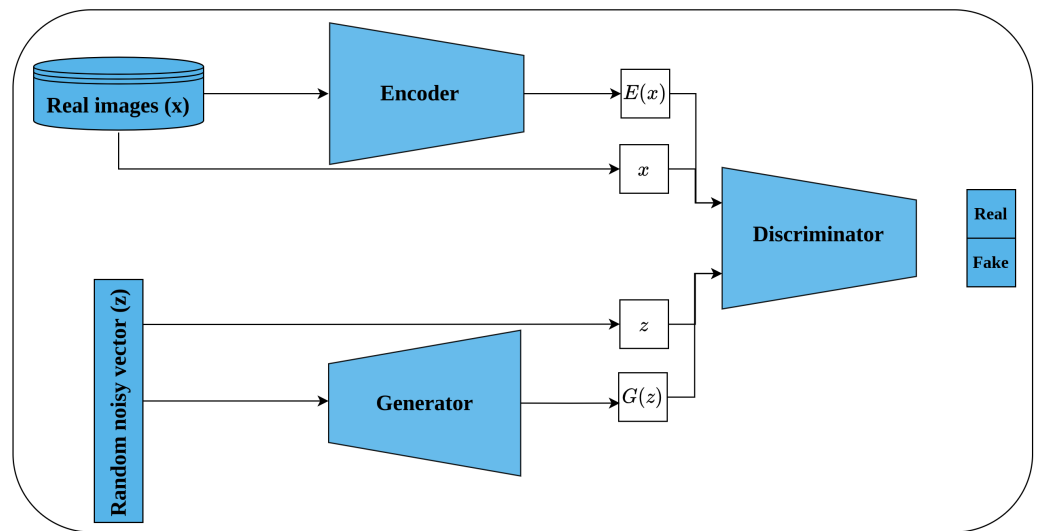
Figure 12 Illustration of self-supervised features learning using bi-directional GAN. (lower left): A generator network that generates a fake image $G(z)$ from random latent space z. (upper left): An encoder network that maps real image x into a latent space $E(x)$. (right): The discriminator network takes as input a tuple of latent space and an image; and classify them as real or fake.

Radford, Metz & Chintala (2016). BiGAN introduces an encoder E which maps an image x back to latent space $E(x)$ (called inverse mapping). The generator decodes random latent space z to produce a fake image $G(z)$. Consequently, the discriminator D takes, as an input, a tuple of latent space and an image which may be either $(G(z), z)$ or $(x, E(x))$ as shown in Fig. 12. The role of the discriminator is to discriminate whether its input tuple is real or fake. The intuition behind incorporating the latent space along with the input image is to serve as free labels generated from the data without supervision in a similar way to learning representations by full supervision. The authors stated that both E and G are completely separated modules that do not communicate with each other during the training. Hence, both modules should learn to invert each other to be able to beat the discriminator. When training is complete, the learned representations, by the encoder, can be transferred to the downstream tasks.

Contrastive self-supervised learning

Contrastive self-supervised learning is a recent representation learning approach that aims at developing robust representations from the input data by learning to differentiate between the similar (*positive*) pairs and the dissimilar (*negative*) pairs or by maximizing the agreement between a pair of positive views depending on the design specifications of the contrastive learning architecture. Positive examples are generated by applying a set of random augmentations to an input image which results in two transformed views of the same image, while negative examples are any other images different from the transformed views. The positive examples are assumed to be slightly different but preserve the global features of the input image which makes the similarity between them higher. Lastly, a contrastive model is trained to maximize the similarity between the positive pairs and

Shurrab et al. (2022), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.1045   **16/51**
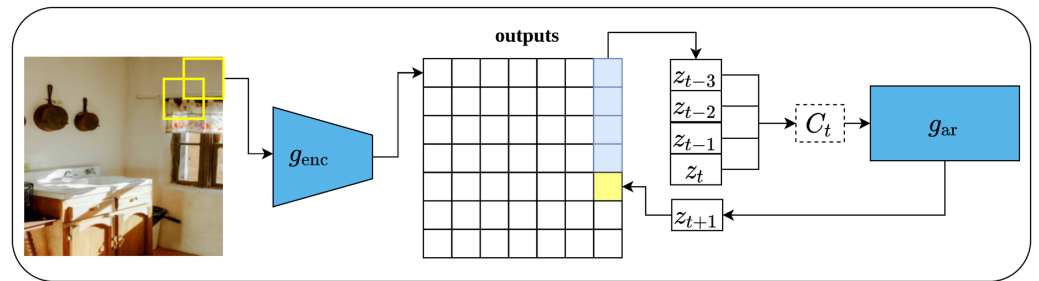
Figure 13 Illustration of features learning using contrastive predicting coding applied on image data.
(left): The input image is rearranged into a grid of overlapping patches of size 7×7. Each crop is then encoded *via* a convolutional network g_{enc} (right): An auto-regressive model is used to make the predictions in top-to-bottom fashion. Image credit: Ali Alcántara.

minimize it with the negative pairs in case of using them. The next sections illustrate five contrastive learning approaches.

Contrastive predictive coding

Contrastive predictive coding (CPC), is a contrastive unsupervised representations learning proposed by *Van den Oord, Li & Vinyals (2018)* that can fit not only image data but also text, and audio. The main intuition behind CPC is to develop a compact representation that maximizes the mutual information between the context C and the target X, rather than predicting the target X directly from C as it is the case with generative models. Such approach enables learning rich representations as it ignores low-level information about the objects in the input data. The architecture of the CPC network consists of three components, namely, an encoder network which is responsible for converting the input into a compact latent variable Z_t; an auto-regressive network which is responsible for producing the context C_t out of the encoded latent variables and generating future predictions; and the contrastive loss function, which is called InfoNCE which is formulated based on the noise-contrastive estimation loss function (NCE) (*Gutmann & Hyvärinen, 2010*).

To apply CPC on visual data (images), an input image of size 256×256 pixels is cropped into patches of size 64x64 pixels with an overlap of 32 pixels with respect to the height and width. This results in a grid of patches of size 7×7. Consequently, each patch is encoded *via* ResNet-101-v2 (*He et al., 2016b*) encoder into a vector Z_t of size 1,024 while the whole image forms $7 \times 7 \times 1,024$ tensor as shown in Fig. 13. Following that, a PixelCNN architecture (*Oord et al., 2016*) is employed as an auto-regressor that generates a context vector $C_t \forall Z_{\leq t}$ which generate future predictions Z_{t+k} in a top-down fashion and in a way that maximizes the mutual information between the context and predictions. Lastly, the InfoNCE loss function contrasts between the predicted patch and all other negative patches; these may come from other locations in the input image or other images in the same mini-batch. CPC V2 *Henaff (2020)* is the second version of CPC which comes with several enhancements over the original CPC.

Momentum contrast

Momentum contrast (MoCo) is another self-supervised contrastive learning approach proposed by *He et al. (2020)*. MoCo framework is inspired by *dynamic dictionary-lookup* and *queues* ideas. The main intuition behind MoCo is to perform a lookup operation using query image encoding in a dictionary that contains keys represented as images' encodings. Learning robust representations is enabled by learning to maximize the similarity between the encoding of the query image and the encoding of its matching key; and to minimize the similarity between the encoding of the query image and non-matching keys. Technically, MoCo architecture consists of two networks, namely, query-encoder and momentum-encoder as shown in Fig. 14. The query-encoder role is to generate a features vector q from the query image x^{query}. The momentum-encoder acts as a dictionary of data samples (whole images or patches x_i^{key}) generated from encodings k_i of features' vectors. Moco maintains a dynamic dictionary which should be of large size and consistent. The dictionary is designed as a queue of feature vectors' encodings k_i, where the present mini-batch enters the queue while the outdated mini-batches leave the queue in a First-In-First-Out fashion. Moreover, the dictionary size is not restricted to the mini-batch size but can be larger. On the other side, as the keys of the dictionary are derived from a group of previous mini-batches, they need to be updated regularly to maintain the consistency property. A momentum update of keys based on values of parameters of the query-encoder is proposed by the authors - where only the query-encoder parameters are updated by back-propagation, while the momentum-encoder is updated consequently using moving average; this allows it to be updated slowly and in a smoother fashion than the query-encoder. MoCo network is trained by minimizing the InfoNCE contrastive loss function (*Van den Oord, Li & Vinyals, 2018*).

Simple framework for contrastive learning of visual representations

Another contrastive learning approach is the simple framework for contrastive learning of visual representations, or SimCLR for short, which was proposed by *Chen et al. (2020)*. As its name implies, SimCLR depends mainly on two simple ideas including heavy data augmentation techniques that result in correlated views for the same input, in addition to a large batch size that includes a large set of negative examples. Furthermore, SimCLR omits the need for additional functionalities as seen in CPC (*Van den Oord, Li & Vinyals, 2018*) and MoCo (*He et al., 2020*). To elaborate more on the SimCLR approach, a set of random transformations τ including cropping and resizing, flipping, rotation, color distortion, and Gaussian blur are applied to the input image x which results in a pair of positively correlated views of the same image $(\tilde{x}_i, \tilde{x}_j)$ as shown in Fig. 15. Consequently, both views are passed into a pair of convolutional encoders $f(.)$, ResNet50 (*He et al., 2016a*), to learn representations of both views, which are denoted as (h_i, h_j), respectively. Following that, the generated representations are passed to a pair of projection heads $g(.)$ which consist of two dense layers with ReLU activation (*Nair & Hinton, 2010*) for the first layer; and linear activation for the second layer. This results in a pair of feature vectors (z_i, z_j). Lastly, the InfoNCE contrastive loss (*Van den Oord, Li & Vinyals, 2018*; *He et al., 2020*) named as normalized temperature-scaled cross-entropy loss (*NT-Xent*) by the authors is employed to
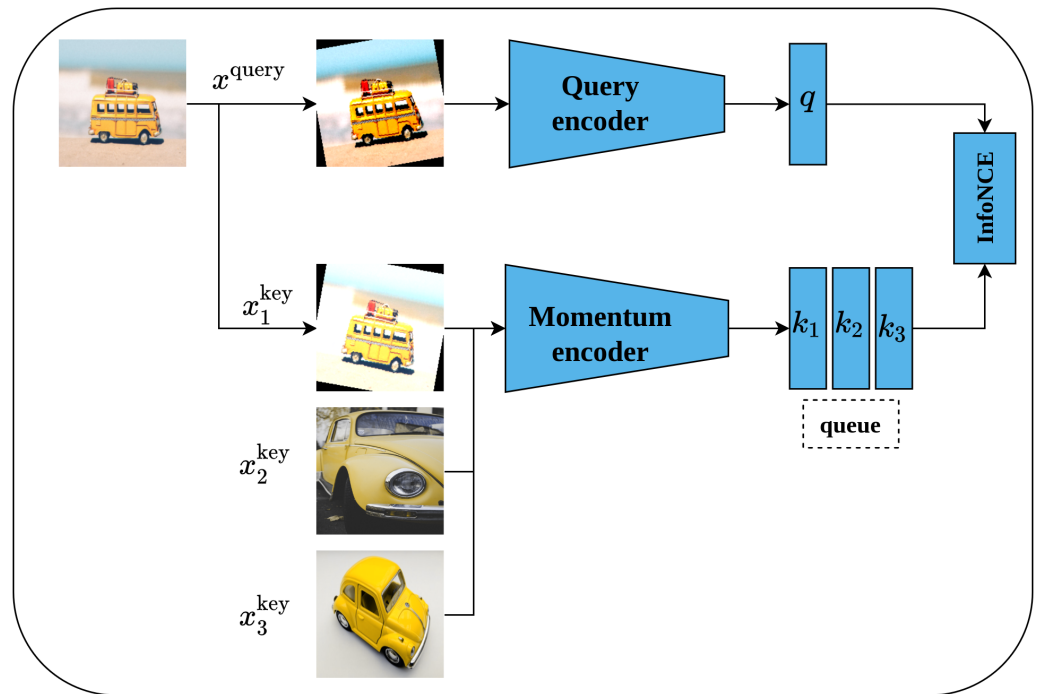
Figure 14 **Illustration of momentum contrast framework.** Image credit; upper: Nubia Navarro, middle: Lilartsy, lower: Mike B.

optimize the whole architecture based on the generated embedding (z_i, z_j) by maximizing the agreement between the positive pairs of augmented images, while minimizing it for other images in the same batch (negative samples). when training is complete, the dense layers are discarded while the convolutional encoders are kept to be utilized in downstream tasks.

Bootstrap your own latent

Bootstrap Your Own Latent (BYOL) is an implicit contrastive learning approach proposed by *Grill et al. (2020)* that omits the need for negative samples during the training. More clearly, BYOL architecture consists of two networks as shown in Fig. 16. The first network is a trainable network, called Online Network, denoted by θ which consists of a representation head f_θ, a projection head g_θ and a prediction head q_θ. The second network is a non-trainable and randomly initialized network, called Target Network, denoted by ξ and has the same architecture as the Online Network except for the prediction head. Target Network acts as a slow-moving average of the Online Network and is updated based on the gradients update in the Online Network *via* the moving average. To train BYOL architecture, two augmented views (x_1, x_2) are generated from the input image x by applying two different augmentation operations $(t, \acute{t})$. Consequently, both augmented views pass to the two networks for encoding (y_θ, y_ξ) and representations generation (z_θ, z_ξ), while z_θ passes through the prediction head to produce the prediction w_θ for the subsequent computation. Following that, both w_θ and z_ξ are normalized *via* L2 norm and accordingly fed into
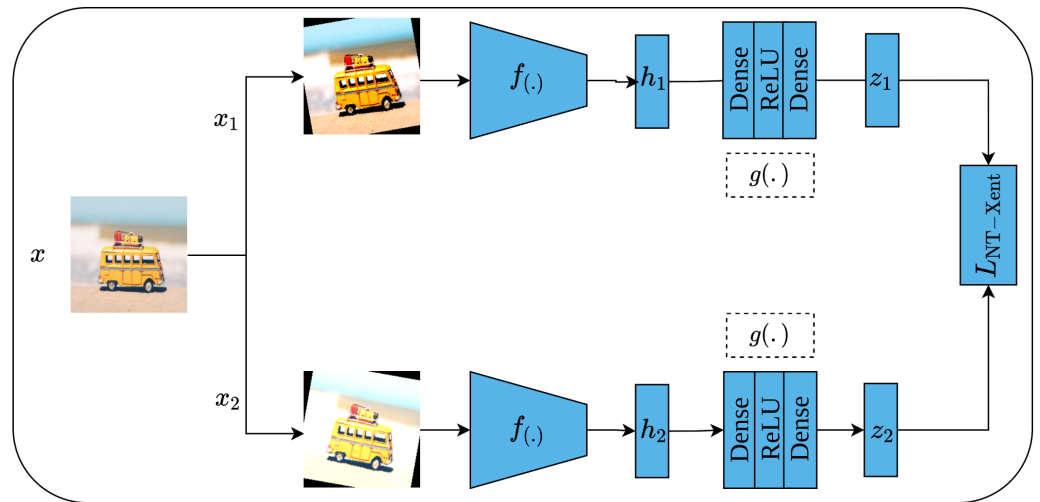
Figure 15 **Self-supervised features learning by SimCLR.** Image credit: Nubia Navarro.

mean squared error (MSE) loss function for optimization rather than contrastive loss. It is worth noting that the gradients flow back only over the Online Network and stopped for Target Network as indicated in Fig. 16 by the term stop-grad which is updated with the momentum equation as a function of the Online Networks parameters θ. Since the target network acts as the moving average of the online network, the online representations should be predicted of the target representations and vice versa. BYOL can learn semantic features by minimizing similarity metrics between the output of each network. Hence, both networks learn interactively from each other from the same image while omitting the need for negative samples.

Swapping assignments between multiple views

While the previous contrastive methods are instance-discrimination-based methods, Swapping Assignments between multiple views (SwAV) is a cluster-discrimination-based method proposed by *Caron et al. (2020)*. Two major elements form the core of the SwAV method including a multi-crop augmentation strategy and the online clustering assignment. The multi-crop strategy aims at generating multiple views of the same image without increasing the memory and computational requirements during the training. This is achieved through generating two global views with standard resolution crops, *e.g.*, 224 × 224, and *V* local views with smaller resolution crops, *e.g.*, 96 × 96. This way, a multi-crop strategy enables producing multiple views rather than just pairs without affecting the computational and memory cost. Besides, each generated view undergoes additional random transformations such as those implemented in SimCLR (*Chen et al., 2020*). On the other side, unlike offline clustering assignment methods which require a complete pass over the dataset to compute the clusters' assignment which becomes computationally intensive in the case of large datasets; Online clustering allows computing clusters' assignment by mapping the encoded views to a prototype vector *C* on the current
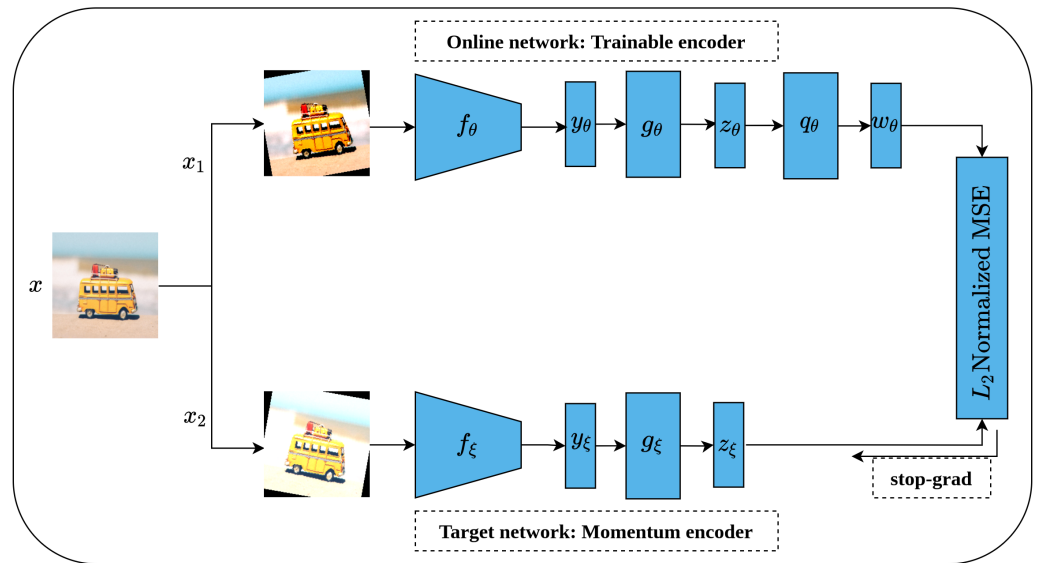
Figure 16 Illustration of BYOL architecture. Image credit: Nubia Navarro.

batch by treating it as an optimal transportation problem (Sinkhorn–Knopp (*Cuturi, 2013*)).

Figure 17 depicts the complete SwAV architecture. Given an input image x, multiple views of the same image are generated by applying a set of random transformation T according to a multi-crop augmentation strategy resulting in x_{nt} augmented views. For simplicity, we will consider one global view x_1 and one local view x_2. Consequently, the generated views are passed into convolutional encoders f_θ, ResNet50 (*He et al., 2016a*) in SwAV case, followed by two dense layers with ReLU activation (*Nair & Hinton, 2010*) to generate feature vectors (Z_1, Z_2). In fact, the initial steps in SwAV do not differ significantly from those of SimCLR (*Chen et al., 2020*) except in the augmentation strategy. Following that, the feature vectors (Z_1, Z_2) are passed through a dense layer with linear activation called prototype layer C which is responsible for mapping the feature vectors into K learnable prototypes (clusters) grouped in a matrix such that $C = [c_1, c_2, ..., c_K]$. It is worth noting that K value is not inferred but user-defined, while the C values represent the weights matrix of the prototype layer. To compute clusters assignments online, only the features of the current batch are used where Sinkhorn–Knopp algorithm is employed to generate the cluster assignments (codes) (Q_1, Q_2) that represent the mapping of feature vectors into clusters in a way that maximizes the similarity between them. Further, Sinkhorn–Knopp enforces the equipartition constraint which prevents assigning all features into a single cluster. Eventually, a swapped prediction problem is performed upon codes generation. Intuitively, given two different views of the same image, they should maintain similar information. Therefore, it is possible to predict the codes of one view from the features vector of the other. This is achieved by minimizing the cross-entropy loss between the code of one view and the softmax of the similarity of the features vector to all clusters. This way, SwAV takes the advantage of contrasting clusters of data with similar features rather
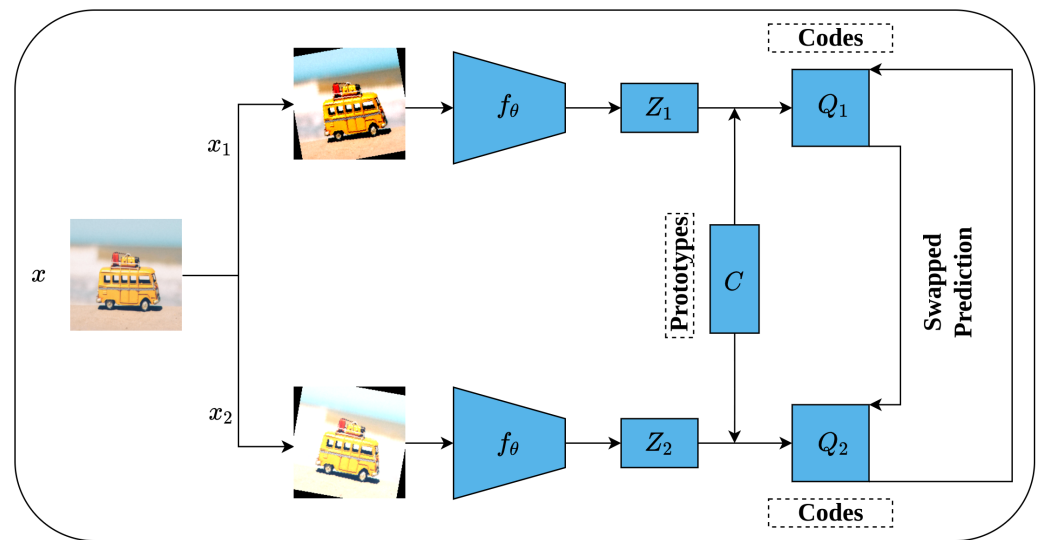
Figure 17 Illustration of SwAV framework. Image credit: Nubia Navarro.

than performing pair-wise comparison over the whole training sets as seen in the previous methods.

To sum up, we opted to provide a high-level overview of each of the previously discussed methods as this article is intended for self-supervised applications in medical imaging which renders it prone to nonspecialist readers from the medical field. One more point to mention is that despite the fact that these methods are developed on natural images, they can be transferred to the medical imaging field as we will see in the next section. Such property encouraged us to briefly discuss them before proceeding with the application of self-supervised learning in medical imaging. Table 1 summarizes the discussed pretext tasks according to their categories, while Table A1 provides an access to the code repository of these works which is provided in Appendix A.

Resources in self-supervised learning

We provided a curated list of pretext tasks that acted as milestones in the history of self-supervised learning in the computer vision field, however, the efforts in this research area are not limited to those methods. As a result, we developed a list of self-supervised learning resources that includes review articles, surveys, and papers as shown in Table 2 for those readers who need to enhance their understanding of the field. For in-depth reviews about self-supervised learning, we highly recommend the readers to refer to one of the following articles: *Jing & Tian (2020)* provided an extensive review of self-supervised learning methods for visual features learning from image and video data, and *Ohri & Kumar (2021)* provided a comprehensive review and performance comparison for a large list of the most recent self-supervised learning approaches developed for image data. Further, *Schmarje et al. (2021)* reviewed various deep learning methods for image classification with fewer labels where self-supervised learning is one of their work dimensions. For Contrastive learning, both *Le-Khac, Healy & Smeaton (2020)* and *Jaiswal et al. (2021)*

Table 1 Summary of self-supervised learning pretext tasks.

No.	Authors	Category	Method
1	*Dosovitskiy et al. (2015)*	Predictive	Exemplar CNN
2	*Doersch, Gupta & Efros (2015)*	Predictive	Relative position prediction
3	*Noroozi & Favaro (2016)*	Predictive	Jigsaw puzzle
4	*Komodakis & Gidaris (2018)*	Predictive	Rotation prediction
5	*Vincent et al. (2008)*	Generative	Denoising auto-encoder
6	*Pathak et al. (2016)*	Generative	Image inpainting
7	*Zhang, Isola & Efros (2016)*	Generative	Image colorization
8	*Zhang, Isola & Efros (2017)*	Generative	Split-brain auto-encoder
9	*Radford, Metz & Chintala (2016)*	Generative	Deep Convolutional GAN
10	*Donahue, Krähenbühl & Darrell (2016)*	Generative	Bi-directional GAN
11	*van den Oord, Li & Vinyals (2018)*	Contrastive	CPC
12	*He et al. (2020)*	Contrastive	MoCo
13	*Chen et al. (2020)*	Contrastive	SimCLR
14	*Grill et al. (2020)*	Contrastive	BYOL
15	*Caron et al. (2020)*	Contrastive	SwAV

Table 2 A summary of self-supervised learning resources.

No.	Authors	Type	Title	Venue
1	*Jing & Tian (2020)*	Survey	Self-supervised visual feature learning with deep neural networks: A survey	IEEE Transactions on Pattern Analysis and Machine Intelligence
2	*Ohri & Kumar (2021)*	Review	Review on self-supervised image recognition using deep neural networks	Knowledge-Based Systems
3	*Schmarje et al. (2021)*	Survey	A survey on semi-, self- and unsupervised learning in image classification	IEEE Acsess
4	*Le-Khac, Healy & Smeaton (2020)*	Review	Contrastive representation learning: A framework and review	IEEE Access
5	*Liu et al. (2021)*	Review	Self-supervised learning: Generative or contrastive	IEEE Transactions on Knowledge and Data Engineering
6	*Jaiswal et al. (2021)*	Survey	Survey on contrastive self-supervised learning	Technologies
7	Jason Ren	Papers list	Awesome self-supervised learning	Github
8	Ashish Jaiswal	Papers list	Awesome contrastive learning	Github

provided a comprehensive survey on contrastive self-supervised methods for different research areas such as computer vision and natural language processing. *Liu et al. (2021)* summarized a set of generative and contrastive self-supervised learning approaches from computer vision, natural language processing, and graph learning. To access these lists of papers, readers may visit the following two repositories: Awesome-self-supervised-learning (https://github.com/jason718/awesome-self-supervised-learning) which covers a curated list of research articles for self-supervised learning from different research areas. In addition, Awesome-contrastive-learning (https://github.com/asheeshcric/awesome-contrastive-self-

supervised-learning) is a curated list of papers that is mainly dedicated to contrastive learning methods.

SELF-SUPERVISED METHODS IN MEDICAL IMAGING

Primarily, there are two paths to follow when employing self-supervised learning in medical images analysis (*Chen et al., 2021a*). The first path is to directly adopt one of the pre-designed pretext tasks from the computer vision field as given in 'Self-Supervised Learning Approaches' or alternatively develop modified versions of these tasks and employ them in the medical context. On the other side, the second path exploits knowledge from the medical domain and computer vision to design novel pretext tasks that target medical images. We prefer to categorize self-supervised learning methods, which are used in the medical field, in the same way we categorized self-supervised learning methods in the computer vision field. Further, after exploring self-supervised learning literature in medical imaging, we discovered that some researchers tend to utilize multiple methods separately or collectively in a multi-tasking fashion. So, we added an additional category called multiple-tasks/multi-tasking to fit such works.

Predictive methods in medical imaging

Inspired by relative position prediction (*Doersch, Gupta & Efros, 2015*) task, *Zhang, Wang & Zheng (2017)* introduced slices ordering pretext task. Knowing that 3D medical images such as CT and MRI scans can be represented as a group of successive 2D slices, such property can be used as an auxiliary supervision signal to learn a good representation. As a result, the authors treated the slice ordering task as a binary classification problem by developing a Siamese convolutional architecture called Paired-CNN that receives two successive slices and predicts their spatial order as below or above. The authors tested their proposed task on fine-grained body part recognition (regression) as a downstream task.

Spitzer et al. (2018) proposed to predict the geodesic distance between two patches located on the brain surface to learn a rich representation of the human brain. They trained a Siamese architecture with two identical branches and weights sharing to accomplish this task. The defined distance between two patches is the Euclidean distance while the ground truth distance is computed manually from the input data. Besides the distance prediction, the authors included the 3D location coordinates prediction of the input patches to the same task which improved the accuracy and convergence of the predicted distances. Lastly, their approach was evaluated on Cytoarchitectonic segmentation as a downstream task.

Bai et al. (2019) proposed anatomical position prediction pretext task from cardiac MRI scans for segmentation purposes. As the cardiac MRI scans provide several cardiac views from different orientations, *e.g.*, short-axis, 2CH long-axis, and 4CH long-axis. Different cardiac anatomical regions, *e.g.*, left and right atrium and ventricle can be expressed using these views. Such properties motivated the authors to define a set of anatomical positions with respect to a certain view as bounding boxes and forced the network to predict these anatomical positions. For the downstream task, a private dataset of 200 annotated cardiac MRI scans was used for evaluation purposes.

Li et al. (2020b) employed self-supervised learning to improve the pseudo-labeling uncertainty estimation in semi-supervised medical images' segmentation by proposing a novel methodology called self-loop uncertainty. They adopted the Jigsaw puzzle pretext task (*Noroozi & Favaro, 2016*) in their approach and introduced random patches' rotation with angles of [0°, 90°, 180°, 270°] to secure learning translation and rotation invariant features. Further, they omitted the need for Siamese architecture as compared to the original Jigsaw puzzle by combining the input patches into a single image for subsequent permutation classification. Besides the labeled data, they leveraged unlabeled data for uncertainty estimation in semi-supervised settings. Two different segmentation tasks were considered for validating the methodology including nuclei segmentation and skin lesion segmentation as down-stream tasks.

Taleb et al. (2021) presented another work that is inspired by Jigsaw puzzle-solving (*Noroozi & Favaro, 2016*) that exploits multiple imaging modalities, e.g.: T1 and T2 scans simultaneously called multi-modal Jigsaw puzzle. A significant improvement has been brought to the original Jigsaw puzzle, besides the multi-modal settings, represented by the employment of Sinkhorn network (*Mena et al., 2018*) for Jigsaw puzzle solving. Sinkhorn network utilizes the Sinkhorn operator as an alternative to the Softmax function which in turn enables solving the Jigsaw puzzle by learning a permutation task rather than a classification task. They also introduced cross-modal synthesis data generation using CycleGAN architecture (*Zhu et al., 2017*) to increase the amounts of data available for self-supervision. On the downstream side, four tasks were utilized for method validation including brain tumor segmentation, prostate segmentation, liver segmentation, and survival days prediction (regression).

Zhuang et al. (2019) proposed a novel pretext task that is inspired by the early work of *Noroozi & Favaro (2016)* on Jigsaw puzzle solving for 3D medical data called Rubik cube recovery. Two operations constitute the Rubik cube recovery pretext task including cube rearrangement and cube rotation. The same logic of the original Jigsaw puzzle task is adopted in the Rubik cube recovery task with 3D input as a Rubik cube partitioned into a 3D grid of $2 \times 2 \times 2$ sub-cubes rather than 2D input with respect to the rearrangement task. To introduce additional complexity, the authors introduced the cube rotation process and limited it to only 180° vertically and horizontally. This way, the authors secured learning translation and rotation invariant features as opposed to the original Jigsaw puzzle task which secures learning only translation-invariant features. Two downstream tasks were used for evaluation purposes including brain hemorrhage classification and brain tumor segmentation which showed competitive performance.

As an extension of the previous work, *Zhu et al. (2020a)* introduced the Rubik cube+ pretext task which adds an additional level of complexity to the Rubik cube recovery problem represented as cube masking identification on the top of both cube rearrangement and cube rotation. Masking identification operation can be viewed as randomly blocking part of the information in a certain cube by masking. The intuition behind masking identification is that robust features learning can be achieved by solving harder tasks. Rubik cube+ was evaluated on the same downstream tasks from the previous work which showed slight improvement.

Table 3 Summary of predictive self-supervised learning methods in medical imaging.

No.	Authors	Pretext task	Down-stream task
1	*Zhang, Wang & Zheng (2017)*	Slices ordering	Body parts recognition
2	*Spitzer et al. (2018)*	Geodesic distance prediction	Brain area segmentation
3	*Bai et al. (2019)*	Anatomical position prediction	Short-axis cardiac MRI segmentation long-axis cardiac MRI segmentation
4	*Li et al. (2020b)*	Jigsaw puzzle	Nuclei Segmentation Skin lesions segmentation
5	*Taleb et al. (2021)*	Jigsaw puzzle	Brain tumor segmentation Liver segmentation Prostate segmentation
6	*Zhuang et al. (2019)*	Rubik cube	Brain tumor segmentation Brain hemorrhage classification
7	*Zhu et al. (2020a)*	Rubik cube+	Brain tumor segmentation Brain hemorrhage classification
8	*Nguyen et al. (2020)*	Spatial awareness	Organ at risk segmentation Intracranial Hemorrhage detection

Nguyen et al. (2020) proposed spatial awareness pretext task that is able to learn semantic and spatial representations from volumetric medical images. Spatial awareness is inspired by the context restoration framework (*Chen et al., 2019*) but was treated as a classification problem. For a certain 3D image, a single slice is selected in addition to a neighboring slice in the range $[-2, 2]$ where this range represents the spatial index. Following that, two patches of predefined dimensions are selected randomly and swapped between the two slices T times. Lastly, a classification network is trained to predict if the input slice is corrupted or not to learn semantic representations. Further, the network is trained to learn the spatial index which enables learning spatial features.

Table 3 summarizes the predictive self-supervised learning methods in medical imaging.

Generative methods in medical imaging

Ross et al. (2018) adopted image colorization pretext task (*Zhang, Isola & Efros, 2016*) for solving endoscopic medical instruments segmentation task from endoscopic video data. The authors did not utilize the original architecture as in the colorization task, but instead, a conditional GAN architecture was employed to encourage generating more realistic colored images, while six datasets from medical and natural domains were used in the evaluation of downstream tasks.

Chen et al. (2019) proposed a novel generative pretext task, called context restoration, that is inspired by the early works of relative position prediction (*Doersch, Gupta & Efros, 2015*) and context encoder (*Pathak et al., 2016*). The authors described the context restoration task as a simple and straightforward method in which two isolated patches are selected randomly and their positions are swapped. The swapping process repeats itself iteratively to produce a corrupted version of the input image but preserves the input image's overall distribution. Following that, a generative model is employed to restore the corrupted image to its original version. Three downstream tasks were used to test the context restoration feasibility including fetal standard scan plane classification, abdominal multi-organ localization, and brain tumor segmentation.

Another work that is built on the same idea of context restoration is called Models Genesis and is performed by *Zhou et al. (2019)* for 3D medical images. As opposed to the context restoration pretext task, models genesis introduced four distortion operations, namely, non-linear transformation using the Bézier transformation function, local pixel shuffling which is similar to the swapping operation in context restoration but in 3D settings, in-painting which is similar to context encoder method and out-painting which is the inverse operation of in-painting. It is worth noting that each input volume undergoes the first two operations and only one of the remaining operations. Consequently, a generative model is built to restore the distorted image to its original context. Six downstream tasks were used to evaluate their method in terms of segmentation and classification tasks.

Matzkin et al. (2020) designed a self-supervised approach for bone flab reconstruction that results from decompressing craniectomy (DC) operations using normal CT scans rather than DC post-operative annotated CT scans. DC is the surgical procedure of removing part of the skull due to different causes such as stroke and traumatic brain injury. The authors designed a virtual craniectomy approach to simulate the DC from normal CT scans that generate DC post-operative CT scans with bone flaps removed from different parts of the upper head which in turn serve as input for the reconstruction model. Consequently, two strategies were proposed to reconstruct the bone flab including direct estimation as well as reconstruction and subtraction. Further, two architectures were employed including U-Net (*Ronneberger, Fischer & Brox, 2015*) and denoising auto-encoder (*Vincent et al., 2008*).

Hervella et al. (2020b) proposed a multi-modal reconstruction task as a self-supervised approach for retinal anatomy learning. The main assumption is that different modalities for the same organ can provide complementary information which enables learning useful representations for the subsequent tasks. The authors proposed to reconstruct fundus fluorescein angiography photos from color fundus photos using aligned pairs from both modalities for the same patient's eye. Further, U-net architecture (*Ronneberger, Fischer & Brox, 2015*) is employed for the sake of completing the reconstruction task along with structural similarity index map(SSIM) (*Wang et al., 2004*) as a loss function. Subsequent research by the same authors experimented with their approach on different ophthalmic oriented down-stream tasks such as retinal vascular segmentation (*Morano et al., 2020*), joint optic disc and cup segmentation (*Hervella et al., 2020a*) and diagnosis of retinal diseases (*Hervella et al., 2021*).

Holmberg et al. (2020) suggested that designing an effective pretext task for medical domains must accurately extract disease-related features which are typically present in a small part of the medical image. Hence, such condition makes traditional pretext tasks that are dominated by the presence of larger objects in natural images ineffective for the medical context. As a result, they developed a novel pretext task for ophthalmic diseases diagnosis called cross-modal self-supervised retinal thickness prediction that employs two different modalities including optical coherence tomography scans (OCT) and infrared fundus images. Initially, retinal thickness maps are extracted from OCT scans by developing a segmentation model using a small annotated dataset which then serves as ground-truth labels for the actual pretext task. Following that, a model is trained to predict the thickness

maps using unlabeled infrared fundus images and the predicted thickness map from the previous step as labels. Learning disease-related features using the proposed approach has been validated by three experienced ophthalmologists. Further, the quality of their task was assessed on diabetic retinopathy grading using color fundus as a downstream task.

Prakash et al. (2020) adopted image denoising approach as a pretext task for nuclei images' segmentation. A special denoising architecture called Noise2Void (*Krull, Buchholz & Jug, 2019*) was employed as a self-supervised pretraining method. Further, four scenarios are evaluated for segmenting nuclei images including random initialization with noisy images, random initialization with denoised images, fine-tuning with noisy images, and fine-tuning with denoised images. The results showed the superiority of self-supervised denoising as opposed to the random initialization.

Hu et al. (2020) adopted context encoder framework (*Pathak et al., 2016*) as a pretext task along with DICOM meta-data as a weak supervision method to learn robust representations from ultrasound imaging. On the top of the context encoder, the authors introduced additional projection discriminator (*Miyato & Koyama, 2018*; *Lučić et al., 2019*) network that produces a feature vector of the inpainted image which to be fed into the classification head and projection head. The classification head classifies the context encoder output as real or fake; while the projection head acts as a conditional classifier that incorporates the DICOM meta-data as weak labels. For DICOM meta-data, two tags were employed including the prop type and the study description as they directly relate to the ultrasound semantic context.

Another extension to Rubik cube pretext tasks is performed by *Tao et al. (2020)* as Rubik cube++ which introduced two substantial changes to the original Rubik cube problem. On the first hand, they introduced the concept of volume-wise transformation which bounds the sub-cubes rotation operation to the neighboring sub-cubes as in playing a real Rubik cube game and as opposed to *Zhuang et al. (2019)* where the sub-cubes are rotated individually. On the second hand, rather than treating the Rubik cube as a classification problem, it has been treated as a generative problem using GAN-based architecture where the generator's role is to restore the original order of the Rubik cube before applying the transformation, while the discriminator role is to discriminate between the correct and wrong arrangement of the generated cubes. As a downstream task, Rubik cube++ has been tested on two segmentation tasks including pancreas segmentation and brain tissues segmentation.

Table 4 summarizes the generative self-supervised learning methods in medical imaging.

Contrastive learning in medical imaging

Jamaludin, Kadir & Zisserman (2017) exploited longitudinal spinal MRI scans as a self-supervised contrastive learning task. This is supported by the fact that time-separated scans from the same patient will share similar representations. As a result, they trained a Siamese network that contrasts two vertebral bodies (VB) MRI scans separated by a period of time. regardless of whether the two images belong to the same patient or not by employing a contrastive loss function *Chopra, Hadsell & LeCun (2005)*. Along with the

Table 4 Summary of generative self-supervised learning methods in medical imaging.

No.	Authors	Pretext task	Down-stream task
1	*Ross et al. (2018)*	Image Colorization	Surgical instruments segmentation
2	*Chen et al. (2019)*	Context restoration	Fetal image classification Abdominal multi-organ localization Brain tumour segmentation
3	*Zhou et al. (2019)*	Models Genesis	Lung nodule segmentation FPR for nodule detection FPR for pulmonary embolism Liver segmentation pulmonary diseases classification RoI, bulb, and background classification Brain tumor segmentation
4	*Matzkin et al. (2020)*	Skull reconstruction	Bone flap volume estimation
5	*Hervella et al. (2020b)*	Multi-modal reconstruction	Fovea localization Optic disc localization Vasculature segmentation Optic disc segmentation
6	*Holmberg et al. (2020)*	Cross modal retinal thickness prediction	Diabetic retinopathy grading
7	*Prakash et al. (2020)*	Image denoising	Nuclei images segmentation
8	*Hu et al. (2020)*	Context encoder	Quality score classification Thyroid nodule segmentation Liver and kidney segmentation
9	*Tao et al. (2020)*	Rubik cube++	Pancreas segmentation Brain tissue segmentation

contrastive loss, they employed a categorical cross-entropy loss to classify the VB scans into seven classes T1-S1. For the downstream task, they tested the pre-trained model on the disc degeneration grading task which showed superior performance in comparison to the random initialization.

Lu, Chen & Mahmood (2020) adopted contrastive predictive coding (*van den Oord, Li & Vinyals, 2018*) along with multiple instance learning (MIL) (*Ilse, Tomczak & Welling, 2018*) for the classification of breast cancer histology images. As a first stage, CPC is employed to learn rich representations from breast cancer histopathological images rather than learning features from scratch using the MIL network. The results showed superior performance as compared to training from scratch and the ImageNet pre-trained model.

Contrastive predictive coding (*Van den Oord, Li & Vinyals, 2018*) is originally designed for 2D data, *Zhu et al. (2020b)* extended the early work on contrastive predictive coding in a way that enables handling 3D data by developing a new method called Task-related CPC. Initially, supervoxels are generated using a simple linear iterative clustering method (SLIC) (*Achanta et al., 2012*) from the input volume to detect the potential lesion areas. Consequently, the sub-volumes that surround the generated supervoxels are cropped to act as the input of the TCPC encoder network. A U-shape path around the center of the generated supervoxel is employed as compared to the original CPC which employs a straight path to achieve better characterization of the lesion. Further, the recurrent neural network acts as the auto-regressor which generates the future predictions, while the whole architecture is optimized using the InfoNCE loss (*van den Oord, Li & Vinyals, 2018*). Brain hemorrhage classification and lung nodule classification tasks were utilized as downstream tasks.

Xie et al. (2020) stated that self-supervised approaches, in general, and contrastive approaches, in specific, are known to consider the global consistency of the input data while ignoring the local consistency. The authors introduced the Prior-Guided Local

(PGL) algorithm for 3D medical images' segmentation which extended the early work on the BYOL method (*Grill et al., 2020*) to consider the local consistency between the different views of the same region. To achieve this, an additional block, called a prior-guided aligner, is added on top of the projection head for both online and target networks used in the original BYOL architecture. The role of the prior-guided aligner is to exploit the augmentation information applied to the input image prior to guide-aligning the features extracted from different views of the same region. Lastly, a local consistency loss function is employed to minimize the difference between the aligned local features. Four downstream segmentation tasks were employed for evaluation purposes, including liver tumors, kidney tumors, spleen, and abdominal organs.

Li et al. (2020a) proposed patient's feature-based Softmax embedding loss function to learn modality and transformation invariant features as well as patients' similarity features using ophthalmic data in contrastive settings. Modality invariant features are learned by combining color fundus photos with a synthesized fundus fluorescein angiography photo of the former photo, while transformations invariant is represented by the ordinary augmentation techniques of the color fundus photo. Such triplet of photos is assumed to share similar features for the same patient. Consequently, to learn patients' similarity features, the triplet of each patient image is considered as a contrasting basis where the features of the same patients are pulled together, while features from other patients are pushed apart using the proposed loss function.

Sowrirajan et al. (2021) adopted the MoCo (*He et al., 2020*) approach to build self-supervised pre-trained models for chest X-ray classification problem. They used pre-trained models on ImageNet (*Deng et al., 2009*) in a supervised fashion as initialization weights for the self-supervised training to speed up the convergence. Further, they suggested that not all augmentation strategies implemented in the original MoCo paper can fit into gray-scale images. Instead, they settle only to use random partial rotation and horizontal flipping. In addition, they tested their work on an external chest X-ray dataset to examine the generalizability of their work on tasks from the same domain, which showed the possibility of transferring the self-supervised learned knowledge to other related tasks.

Vu et al. (2021) proposed the MedAug approach as an augmentation strategy that benefits from the patient meta-data when training MoCo framework (*He et al., 2020*) as an extension of the early work performed by *Sowrirajan et al. (2021)*. More clearly, MedAug requires that the different views must come from the same patient, as such images are expected to be rich in pathological features. In addition, MedAug considers studying number and laterality as two additional conditions derived from the patient meta-data. For the same patient, the study number represents images taken in different sessions, while laterality represents the orientation as frontal or lateral. This way, MedAug leveraged medical knowledge to the learning algorithm rather than depending merely on the transformations obtained by ordinary augmentation techniques to generate positive views. MedAug was tested on pleural effusion classification from chest X-ray as a downstream task.

Sriram et al. (2021) purely adopted MoCo (*He et al., 2020*) as an approach for COVID patients deterioration prediction tasks. They used non-COVID chest X-ray images from

different public datasets to train MoCo for the subsequent tasks. On the other side, the authors defined three prediction tasks that indicate COVID patient deterioration including single image prediction, oxygen requirements prediction, and multiple image prediction as downstream tasks. The first two tasks are ordinary classification problems from a single image; while the third one requires multiple time-indexed radiographs. A continuous positional embedding module was employed to obtain representations from a set of time-indexed radiographs. Another similar work performed by *Chen et al. (2021b)*, which adopted MoCo as a pretraining method, uses chest CT scans for COVID diagnosis *via* a few-shot learning prototypical network (*Snell, Swersky & Zemel, 2017*) as a down-stream task. Similarly, public non-COVID chest CT was utilized for MoCo training; and two public COVID datasets were utilized for evaluation.

Chaitanya et al. (2020) provided two significant improvements to the SimCLR (*Chen et al., 2020*) contrastive learning approach for 3D images segmentation by developing domain-specific and problem-specific knowledge simultaneously. To elaborate more on the domain-specific knowledge, the original contrastive loss (NT-Xent) maximizes the similarity between a pair of transformed versions of the input image by augmentation alone to obtain a global representation. Also, 3D medical images consist of a set of sequential images that depict similar anatomical regions. Hence, such sequences can be exploited as a positive pair to learn a global representation. On the other side for problem-specific knowledge, a segmentation task that is considered a pixel-wise prediction problem requires local representation. As a result, the authors introduced a local contrastive loss function that helps learn a local representation based on the similarity between the local regions within the input volume. It is worth noting that the proposed approach employs encoder–decoder architecture, where the encoder is optimized with global loss while the decoder is optimized with the local loss. Further, cardiac segmentation and prostate segmentation were employed as downstream tasks.

Azizi et al. (2021) adopted a self-supervised contrastive learning approach in a medical context in a way that combines learning features from both unlabelled natural images and unlabelled medical images in a sequential fashion. To elaborate more, they adopted SimCLR (*Chen et al., 2020*) and introduced novel contrastive learning called multi-instance contrastive learning (MICLe) which is built on the same logic of SimCLR with minor modifications. The main idea behind MICLe is to leverage the availability of multiple views of a certain pathology from the same patient as the foundation for contrastive learning. Such correlated views of the same patient are considered as positive pairs rather than generating multiple views from the same image as in SimCLR. In their experiments, the authors tested SimCLR on chest X-ray images dataset with fourteen classes, while MICLe was tested on a Dermatology dataset with twenty-seven classes as a downstream task.

Table 5 summarizes the contrastive self-supervised learning methods in medical imaging.

Multiple-tasks/Multi-tasking in medical imaging

Tajbakhsh et al. (2019) experimented with three pretext tasks, namely, rotation prediction (*Komodakis & Gidaris, 2018*), image colorization (*Larsson, Maire & Shakhnarovich, 2017*)

Table 5 Summary of contrastive self-supervised learning methods in medical imaging.

No.	Authors	Pretext task	Down-stream task
1	*Jamaludin, Kadir & Zisserman (2017)*	Longitudinal spinal MRI	Disc degeneration grading
2	*Lu, Chen & Mahmood (2020)*	CPC	Breast cancer classification
3	*Zhu et al. (2020b)*	TCPC	Brain hemorrhage classification Lung Nodule classification
4	*Xie et al. (2020)*	BYOL	Liver segmentation Spleen segmentation Kidney tumour seg. Abdominal organs seg.
5	*Li et al. (2020a)*	Feature-based softmax embedding	PM classification AMD classification Diabetic retinopathy detection
6	*Sowrirajan et al. (2021)*	MoCo	Tuberculosis detection Pleural effusion classification
7	*Vu et al. (2021)*	MoCo	Pleural effusion classification
8	*Sriram et al. (2021)*	MoCo	COVID patient prognosis
9	*Chen et al. (2021b)*	MoCo	COVID few-shot classification
10	*Chaitanya et al. (2020)*	SimCLR	Cardiac segmentation Prostate segmentation
11	*Azizi et al. (2021)*	SimCLR	Chest X-ray classification Skin lesions classification

and 3D Patch reconstruction (*Arjovsky, Chintala & Bottou, 2017*) on three different medical image analysis tasks including false positive reduction for nodule detection in chest CT scans, diabetic retinopathy severity classification, lung lobe segmentation and skin segmentation. Due to the substantial differences among the utilized imaging modalities, each of them was assigned a specific pretext task. More clearly, image rotation was employed for both lung lobe segmentation and diabetic retinopathy classification, while colorization was employed for the skin segmentation task and finally 3D patch reconstruction was employed for nodule detection. *Jiao et al. (2020)* proposed temporal order correction and spatio-temporal transformation prediction pretext tasks to learn good representations from fetal ultrasound videos. For the first task, the order of the ultrasound video frames is shuffled and the role of the task is to predict the correct order of the shuffled frames. For the second task, certain affine transformations are applied to the input video and the role of the task is to predict the applied transformations. To train both tasks jointly, the authors proposed two strategies including a Siamese network with partial weights sharing that learns two tasks simultaneously with one branch for each task. The second strategy is called objective disentanglement which enables incorporating the proposed task into the same input video and training the network to recognize both of them.

Li, Chen & Zheng (2020) combined two colorization-based pretext tasks into a single multi-tasking framework called ColorMe to learn useful representations from scopy images. In a similar way to the original colorization task (*Zhang, Isola & Efros, 2016*), the authors proposed to predict red and blue channels from the green channel in an RGB scopy images to obtain local features. On the other side, the authors proposed color distribution estimation of the red and blue channels to force learning of global features. Then, both tasks are trained jointly and evaluated on two downstream tasks, namely, cervix type classification and skin lesion segmentation.

Taleb et al. (2020) suggested that rich representations can be learned from medical images with 3D nature instead of 2D images. For this reason, they applied five pre-designed

pretext tasks, namely, CPC (*Van den Oord, Li & Vinyals, 2018*), exemplar CNN (*Dosovitskiy et al., 2015*), rotation prediction (*Komodakis & Gidaris, 2018*), relative position prediction (*Doersch, Gupta & Efros, 2015*) and Jigsaw puzzle (*Noroozi & Favaro, 2016*) to be adaptive with medical images of 3D nature. Their methods were tested on two 3D downstream tasks which are brain tumor segmentation and pancreas tumor segmentation. *Luo et al. (2020)* proposed a self-supervised fuzzy clustering network as a pretext task for color fundus photo classification. The proposed approach consists of auto-encoder architecture which is responsible for initial features learning from the input data as a first stage. In addition, a clustering module that guides the self-supervision process is employed as a second stage. After gaining the initial representations, the Fuzzy C-means algorithm is utilized (*Bezdek, Ehrlich & Full, 1984*) on top of the encoder network to cluster similar inputs into predefined clusters and update the encoder weights accordingly. The learned weights, after the clustering phase is complete, are transferred to the downstream task.

Haghighi et al. (2020) introduced Semantic Genesis as an extension to the previous work on Models Genesis framework (*Zhou et al., 2019*). Besides features learning by restoration, the authors introduced two additional functionalities called self-discovery and self-classification. Self-discovery is the first stage of the Semantic Genesis framework, where an auto-encoder is trained to reconstruct the input images. Such steps help in discovering a set of semantically similar patients who share similar anatomical patterns by comparing their encoding vectors. Consequently, a random number of crops with fixed coordinates are derived from those patients and assigned a numerical label that denotes their positions. For the self-classification stage, a classification head, on the top of the framework encoder, is employed to classify the extracted batches according to their assigned labels. In addition, the same intuition of Models genesis is adopted in the self-restoration phase but applied to the extracted patches rather than the whole image. This way, Semantic Genesis enables learning semantically rich representations from similar anatomical patterns. Seven downstream tasks were utilized for evaluation as classification and segmentation tasks.

Zhang et al. (2020) introduced scale-aware restoration pretext task for 3D medical images segmentation as an extension of Models Genesis framework (*Zhou et al., 2019*). In addition to the transformation restoration as in Models Genesis, the authors introduced scale discrimination property to the original model depending on the fact that desired objects, *e.g.*, tumors, appear in different sizes across different patients. And hence, cubes of predefined sizes as small, medium, and large are generated and resized into a unified size and then labeled according to their original cropping size. Consequently, the classification head is included on the top of the encoder to accomplish the scale classification task; while the whole architecture is responsible for the transformation restoration task. Brain tumor segmentation and pancreas organs and tumors segmentation were used as downstream tasks.

Dong, Kampffmeyer & Voiculescu (2021) developed a multi-task self-supervised learning approach that combines both generative modeling and instance discrimination using sequential medical data. Given a sequence of medical images for the same patient *e.g* CT, an auto-encoder architecture, with a single encoder and two decoders, is responsible for

learning representations by predicting the T steps precedent and successor slices of the input slice which in turn enables learning the anatomical structural similarity between different slices. In addition, an instance discrimination task is included to avoid learning trivial features by generative modeling. To achieve this, an additional encoder is introduced to the whole architecture that takes another input slice from the same patient and tries to contrast it with the generative model input by minimizing the negative cosine similarity between both inputs. It is worth noting that the second encoder shares the same weights with the generative model whereas it does not go through the back-propagation process.

Koohbanani et al. (2021) proposed a self-path framework for histopathology images which comprises three pathology-specific tasks, namely, magnification prediction, magnification Jigsaw puzzle (JigMag), and Hematoxylin channel prediction in multi-tasking settings. For the first one, patches with different predefined levels of magnification are extracted whereas the task role is to predict the right magnification of the input image. For JigMag, the generated puzzles for training include patches with different magnifications levels for the same image, while the task role is to predict the right order of the puzzle. For the latter task, out of a histopathological image stained with Hematoxylin and Eosin, the role of the task is to predict the first channel from the stained image. Lastly, all proposed tasks along with the downstream tasks are trained jointly in a multi-tasking fashion.

Zhang et al. (2021) developed a semi-supervised multi-tasking approach that combines rotation prediction (*Komodakis & Gidaris, 2018*), Jigsaw puzzle (*Noroozi & Favaro, 2016*) and SimCLR (*Chen et al., 2020*) in a unified framework called twin self-supervision based semi-supervised learning (TS-SSL) for spectral-domain optical coherence tomography (SD-OCT) classification. For the Jigsaw puzzle, the authors introduced patch rotation as given in *Li et al. (2020b)*, while for SimCLR the authors introduced supervised category-wise contrastive loss, which considers all samples for a certain label as positive examples. Consequently, the proposed approach is trained in an end-to-end fashion and semi-supervised multitasking setting to learn representations by performing rotation prediction, Jigsaw puzzle-solving, contrastive and supervised contrastive learning. The methods were evaluated on multi-class and binary OCT classification tasks.

Li et al. (2021) suggested that rotation-oriented collaborative features learning would provide a potent representation for fundus disorders. They simultaneously combined rotation prediction (*Komodakis & Gidaris, 2018*) with multi-view instance discrimination (*Wu et al., 2018*) to learn rotation-related and rotation-invariant features using fundus color photography in an end-to-end fashion. Their approach was tested on two ophthalmic diseases, namely, pathological myopia (PM) and age-related macular degeneration (AMD) as a binary classification downstream task. Further, their approach showed that the collaborative approach provided better results than using a single pretext task at a time.

Lu, Li & Ye (2021) designed two domain-specific pretext tasks for white matter tract segmentation from diffusion MRI scans. The first task is concerned with predicting the fiber streamlines density map of the white matter in the human brain which represents the number of streamlines that pass through a voxel. On the other side, the second task is concerned with imitating registration-based white matter tract segmentation by registering the input data to a predefined white matter tract registration atlas. Further, both tasks

Table 6 Summary of multiple-tasks/multi-tasking self-supervised learning methods in medical imaging.

No.	Authors	Pretext task	Down-stream task
1	*Tajbakhsh et al. (2019)*	Colorization rotation prediction 3D patch reconstruction	Lung lobe segmentation FPR for nodule detection skin lesions segmentation diabetic retinopathy grading
2	*Jiao et al. (2020)*	Temporal order correction transformation prediction	Standard plane detection saliency prediction
3	*Li, Chen & Zheng (2020)*	ColorMe	Cervix type classification skin lesion segmentation
4	*Taleb et al. (2020)*	CPC Jigsaw puzzle Exemplar CNN Rotation Prediction Relative position prediction	Brain tumors segmentation pancreas tumor segmentation
5	*Luo et al. (2020)*	Self-supervised fuzzy clustering	Color fundus classification diabetic retinopathy classification
6	*Haghighi et al. (2020)*	Semantic Genesis	Lung nodule segmentation FPR for nodule detection liver segmentation chest diseases classification brain tumor segmentation pneumothorax segmentation
7	*Zhang et al. (2020)*	Scale-aware restoration	Brain tumor segmentation pancreas segmentation
8	*Dong, Kampffmeyer & Voiculescu (2021)*	Multi-task self-supervised learning	Whole heart segmentation
9	*Koohbanani et al. (2021)*	Self-path	histopathology image classification
10	*Zhang et al. (2021)*	SimCLR Jigsaw puzzle	Binary OCT classification multi-class OCT classification
11	*Li et al. (2021)*	Rotation prediction multi-view instance discriminate	PM classification AMD classification
12	*Lu, Li & Ye (2021)*	Fiber streamlines density map prediction Registration imitation	White matter tract segmentation

are employed sequentially rather than independently as each of the proposed methods focuses on part of the white matter properties, and hence, integrating them may provide complementary information.

Table 6 summarizes the multiple-tasks/multi-tasking self-supervised learning methods in medical imaging.

PERFORMANCE COMPARISON

This section compares the performance of the proposed self-supervised learning approaches that have been discussed in the previous section. Mainly, the emphasis, in this section, is on two tasks which are images classification and semantic segmentation as these two tasks are the most common tasks in the discussed works. Further, this section reports the performance of the proposed self-supervised learning approaches in medical images in comparison to random initialization and transfer learning from ImageNet where applicable. Lastly, this section considers only the benchmarks that have more than two works evaluated on them.

Table 7 Performance comparison on LUNA 2016 dataset. Pretext tasks indicated with an asterisk (*) are reproduced by the same author. Pretext tasks indicated with two asterisks (**) are implemented using different backbones.

No.	Author	Pretext task	Category	Random init.: AUC	SSL: AUC
1	*Zhu et al. (2020b)*	TCPC **	Contrastive	0.982	0.996
2	*Zhu et al. (2020b)*	TCPC **	Contrastive	0.911	0.987
3	*Haghighi et al. (2020)*	Semantic Genesis	Multi-tasking	0.943	0.985
4	*Zhou et al. (2019)*	Models Genesis	Generative	0.942	0.982
5	*Haghighi et al. (2020)*	Rubik Cube*	Predictive	0.943	0.955
6	*Haghighi et al. (2020)*	Context Restoration*	Generative	0.943	0.919
7	*Haghighi et al. (2020)*	Image Inpainting*	Generative	0.943	0.915
8	*Haghighi et al. (2020)*	Auto-encoder*	Generative	0.943	0.884
9	*Tajbakhsh et al. (2019)*	3D patch Reconstruct.	Generative	0.724	0.739

Table 8 Performance comparison on brain hemorrhage classification dataset. Pretext tasks indicated with two asterisks (**) are implemented using different backbones.

No.	Author	Pretext task	Category	Random init.: Acc (%)	SSL: Acc (%)
1	*Zhu et al. (2020b)*	TCPC **	Contrastive	81.08	88.17
2	*Zhu et al. (2020a)*	Rubik Cube+ **	Predictive	79.73	87.84
3	*Zhuang et al. (2019)*	Rubik Cube	Predictive	72.60	83.80
4	*Zhu et al. (2020a)*	Rubik Cube+ **	Predictive	72.30	78.68
5	*Zhu et al. (2020b)*	TCPC **	Contrastive	72.30	78.38

Table 9 Performance comparison on CheXpert dataset.

No.	Author	Pretext task	Category	ImageNet: AUC	SSL: AUC
1	*Sowrirajan et al. (2021)*	MoCo	Contrastive	0.949	0.953
2	*Vu et al. (2021)*	MoCo	Contrastive	0.858	0.906
3	*Azizi et al. (2021)*	SimCLR	Contrastive	0.763	0.767

Classification tasks performance comparison

For the sake of performance comparison of the proposed methods from the previous section on the classification tasks, three datasets are selected in which two of them are public datasets which are Lung Nodule Analysis LUNA (*Setio et al., 2016*), and CheXpert dataset (*Irvin et al., 2019*), while the third one is a private dataset which is Brain Hemorrhage classification dataset. For consistent comparison, the most common performance measure among all proposed approaches is reported. The area under curve(AUC) for receiver operating characteristic curve (ROC) is reported for both LUNA and CheXpert datasets, while overall accuracy is reported for the brain hemorrhage classification dataset. Further, all reported results are in terms of fine-tuning the whole model on the downstream task rather than the results of the fine-tuning the classification layer. Table 7 reports the results on the LUNA dataset while Table 8 reports the results on the brain hemorrhage classification dataset, and Table 9 reports the results on the CheXpert dataset.

Table 10 Performance comparison based on BraTS dataset.

No.	Author	Pretext task	Category	Random init.: DSC (%)	SSL: DSC (%)
1	*Zhou et al. (2019)*	Models Genesis	Generative	90.68	92.58
2	*Taleb et al. (2021)*	Jigsaw Puzzle	Predictive	80.54	89.74
3	*Zhu et al. (2020a)*	Rubik Cube+	Predictive	85.47	89.6
4	*Chen et al. (2019)*	Context Restoration	Generative	84.41	85.57
5	*Zhang et al. (2020)*	Scale Aware Rest.	Multi-Tasking	74.35	84.92
6	*Chen et al. (2019)*	Image Inpainting*	Generative	84.41	84.54
7	*Taleb et al. (2020)*	3D Relative Pos. Pred.	Predictive	76.38	81.28
8	*Taleb et al. (2020)*	3D CPC	Contrastive	76.38	80.83
9	*Taleb et al. (2020)*	3D Rotation	Predictive	76.38	80.21
10	*Taleb et al. (2020)*	3D Jigsaw	Predictive	76.38	79.66
11	*Taleb et al. (2020)*	3D Exemplar	Predictive	76.38	79.46

It can be clearly observed from Tables 7, 8 and 9 that self-supervised learning approaches provide better performance when compared to either training from scratch or using ImageNet pre-trained models. In addition, designing new approaches or modifying existing methods in a way that considers the properties of medical images provides better performance than directly adopting pretext tasks from the computer vision field. Contrastive learning based algorithms tend to show better results when compared to other categories as shown in Table 7. However, this is not always the case where the TCPC approach showed the best and the worst results on the brain hemorrhage classification dataset when using different backbones as shown in Table 8. In Table 9, it can be clearly seen that directly adopting pretext task from computer vision to the medical images analysis field provides marginal improvements on the down-stream task as compared to the ImageNet pre-trained models as given by both *Sowrirajan et al. (2021)* and *Azizi et al. (2021)*. On the other side, modifying computer vision tasks by incorporating medical knowledge, as given by *Vu et al. (2021)*, provided significant improvements on the performance when compared to ImageNet pre-trained models.

Segmentation tasks performance comparison

Performance comparison on semantic segmentation tasks is reported on brain tumor segmentation dataset (BraTS 2018) (*Bakas et al., 2018*). All listed works are compared with respect to the Dice Score (DSC) as it is the most common performance among all works on the same dataset. Lastly, all reported results are in terms of fine-tuning the whole model on the downstream task.

Table 10 reports the results of the BraTS dataset. It can be clearly seen that the performance of self-supervised learning based approaches significantly outperforms training from scratch. In addition, it can be observed that the adopted methods from the computer vision field do not provide better performance as compared to those methods designed especially to suit the nature of medical images.

DISCUSSION AND FUTURE RESEARCH DIRECTIONS

A variety of methods that employed self-supervised learning in medical imaging analysis have been discussed in the previous section with respect to each category. Some of these works adopted methods from the computer vision field. Other researchers proposed novel approaches by incorporating medical knowledge into the design of the pretext task or exploiting the unique properties of the medical images. This section discusses the salient insights that can be derived from the previously discussed works.

Computer vision task in medical imaging

Despite the fact that computer vision and medical images analysis fields deal with image data, there are fundamental differences in the characteristics of natural images and medical images in terms of the number of channels, intensity, location, scale, and orientation. For the number of channels, natural images are mainly 2D RGB images, while medical images may be 2D gray-scale, 3D volumes, or 4D as volume over time dimension. For intensity, the same object will nearly possess the same features under different intensity levels, *e.g.*, the human face is the same under different intensities. On the other side, intensity holds meaningful information in medical images, *e.g.*, different tissues have different values according to the Hounsfield scale in CT scans. For the location, objects in natural images are not affected by changing locations, *e.g.*, the human face holds the same features for the same person in different locations. In medical images, object location has significant indication with respect to certain pathology, *e.g.*, Diabetic Macular Edema severity is diagnosed by examining the Oedema presence with respect to the Fovea in OCT scans. For scale, an object's features in natural images are not significantly affected by the scale, *e.g.*, the human face will not change significantly by changing magnification levels. In contrast, scale is an important factor in some medical imaging modalities, *e.g.*, in histopathological images, different information can be obtained at different magnification levels. Eventually, orientation in natural images is a significant factor for some applications, *e.g.*, texts and numbers orientation in optical character recognition applications. In medical images, orientation may not be a decisive factor, *e.g.*, tumors may have different non-predefined shapes which in turn make them agnostic for orientations. In summary, medical images have unique properties that distinguish them from the natural images that need to be considered (*Zhou et al., 2021*).

The direct adoption of pretext tasks from the computer vision field, which have achieved state-of-the-art results on natural images, may not necessarily give the same performance in the medical images analysis field. Hence, knowing that medical images have unique properties as compared to natural images, these properties must be taken into consideration when adopting pretext tasks from computer vision.

A variety of the discussed works considered the unique properties of medical images when adopting pretext tasks from the computer vision field and modified these methods accordingly. For instance, several works modified the existing methods to be able to deal with the volumetric nature of the medical images rather than 2D images such as in *Taleb et al. (2020)*, *Zhuang et al. (2019)*, *Zhu et al. (2020a)* and *Zhu et al. (2020b)*. Other researchers provided modifications to the existing methods to suit the nature of medical images in

terms of loss functions such as in *Chaitanya et al. (2020)*, *Li et al. (2020a)*, *Xie et al. (2020)*, and positive pairs selection for contrastive learning algorithms such as in *Jamaludin, Kadir & Zisserman (2017)*, *Azizi et al. (2021)*, *Chaitanya et al. (2020)*, *Vu et al. (2021)* and *Li et al. (2020a)*. Lastly, some researchers combined more than one computer vision task together to enable robust representations learning such as in *Zhang et al. (2021)* and *Li et al. (2021)*. To sum up, pretext tasks adopted from computer vision need to be modified when adopted in medical imaging analysis in a way that fits the unique characteristics of medical images. Further, the previously mentioned differences between natural images and medical images can be considered as design considerations for the research in the field.

Pretext tasks based on medical knowledge

Most of the presented works that proposed novel pretext tasks tend to be based on the manipulation of the input image as well as the property of the images. Fewer works tend to incorporate medical knowledge into their approaches such as in *Hu et al. (2020)*, *Lu, Li & Ye (2021)*, *Hervella et al. (2020a)*, *Holmberg et al. (2020)* and *Vu et al. (2021)*. This may be attributed to the fact that incorporating medical knowledge such as patient metadata, cross-modal images, and disease-specific knowledge may limit the applicability of the proposed self-supervised learning approach to a certain imaging modality and specific disease and may limit its transferability to other tasks without the need to modify the core of the proposed approach. On the other hand, exploiting medical images' properties as well as images manipulation as the bases for the design of pretext tasks provides a wider range of applications for different imaging modalities that possess common attributes. Medical knowledge incorporation with the design of pretext tasks is another research direction that needs to be further explored to benefit from such available knowledge in designing self-supervised learning approaches that are able to provide robust representations empowered by the medical knowledge.

Pretext tasks design with multiple imaging modalities

Diagnosis of a certain disease or capturing a certain organ in the clinical practice may be performed by using more than one imaging modality as they provide complementary information. As an example, OCT scans and fundus color photos are used to diagnose retina diseases. Several works that have been discussed in this survey considered such property in their designs of pretext tasks such as in *Holmberg et al. (2020)*, *Hervella et al. (2020b)*, *Li et al. (2020a)* and *Taleb et al. (2021)*. While learning from a single imaging modality can produce good representations, incorporating multiple imaging modalities, in the design of pretext tasks, can offer learning rich representations. Hence, additional research efforts need to be performed in this direction.

Data availability

Most of the presented works utilize either public or private datasets for the training phase of the pretext tasks. Public datasets are known to be of small size except for some modalities such as X-ray (*Irvin et al., 2019*; *Wang et al., 2017*), fundus (https://www.kaggle.com/c/diabetic-retinopathy-detection) color photo, and optical coherence tomography (*Kermany, Zhang & Goldbaum, 2018*) which are available with a considerable

number of images. On the other side, private datasets are not available to the research community and are not easy to reach. Hence, there is a need for building large unlabeled data pools that cover a wide range of imaging modalities to be available for the research community to accelerate the application of self-supervised learning in the field. An important point to consider, when developing an unlabeled medical images dataset, is the data bias. More clearly, medical images datasets in general and medical images datasets in specific tend to be biased toward the healthy cases, while fewer images represent the abnormalities. Data bias must be avoided when developing self-supervised learning methods to guarantee learning representations that are rich in pathological features.

CONCLUSION

Machine learning applications in medical imaging analysis require large amounts of high-quality annotated data to develop robust models in a supervised fashion, which may not be always available at our disposal. Annotated medical images are scarce and this acts as a major problem that researchers, in the field of machine learning, encounter. Self-supervised learning methods can significantly alleviate the problem of scarcity in annotated data, in the field of medical images analysis, as it enables learning robust representations from unlabeled data.

This is the first survey, to the best of our knowledge, that covers recent self-supervised learning methods and their applications in the field of medical imaging analysis and cast them into four categories, namely, predictive, generative, contrastive, and multi-tasking. This survey extensively reviews 15 state-of-the-art self-supervised learning methods from the computer vision field that have been extensively employed in the context of medical imaging analysis. In addition, the survey covers the 40 most prominent self-supervised learning applications in the field of medical imaging analysis for different imaging modalities and medical conditions. Further, a comparative analysis is conducted to highlight the best performers among the reviewed self-supervised learning approaches in the medical images field when compared on a unified benchmark. Finally, this survey summarizes the major patterns that can be observed from the discussed self-supervised learning applications in medical imaging. Moreover, this survey emphasizes some of the open issues in the field that requires attention from the research community.

APPENDIX A

Table A1 lists the implementation of the previously listed works from both computer vision and medical image analysis who render their code publicly available. Further, starred implementations represents the authors' official code.

Table A1 Implementation codes list.

No.	Authors	Pretext task	Implementation
1	*Dosovitskiy et al. (2015)*	Exemplar CNN	caffe*
2	*Doersch, Gupta & Efros (2015)*	Relative position prediction	caffe*—pytorch
3	*Noroozi & Favaro (2016)*	Jigsaw puzzle	pytorch
4	*Komodakis & Gidaris (2018)*	Rotation prediction	pytorch*
5	*Vincent et al. (2008)*	Denoising auto-encoder	theano—pytorch
6	*Pathak et al. (2016)*	Image inpainting	caffe*—tensorflow
7	*Zhang, Isola & Efros (2016)*	Image colorization	pytorch*
8	*Zhang, Isola & Efros (2017)*	Split-brain auto-encoder	caffe*—tensorflow
9	*Radford, Metz & Chintala (2016)*	Deep Convolutional GAN	pytorch—tensorflow
10	*Donahue, Krähenbühl & Darrell (2016)*	Bi-directional GAN	theano*—tensorflow
11	*van den Oord, Li & Vinyals (2018)*	CPC	pytorch—tensorflow
12	*He et al. (2020)*	MoCo	pytorch*—tensorflow*
13	*Chen et al. (2020)*	SimCLR	tensorflow*—pytorch
14	*Grill et al. (2020)*	BYOL	tensorflow*—pytorch
15	*Caron et al. (2020)*	SwAV	pytorch*—tensorflow
16	*Taleb et al. (2020)*	CPC Jigsaw puzzle Exemplar CNN Rotation Prediction Relative position prediction	tensorflow*
17	*Li et al. (2021)*	Rotation prediction multi-view instance discriminate	tensorflow*
18	*Sriram et al. (2021)*	MoCo	pytorch*
19	*Sowrirajan et al. (2021)*	MoCo	pytorch*
20	*Xie et al. (2020)*	BYOL	pytorch*
21	*Chaitanya et al. (2020)*	SimCLR	tensorflow*
22	*Zhou et al. (2019)*	Models Genesis	tensorflow*
23	*Haghighi et al. (2020)*	Semantic Genesis	tensorflow*
24	*Li et al. (2020a)*	Feature-based softmax embedding	pytorch*
25	*Holmberg et al. (2020)*	Cross modal retinal thickness prediction	tensorflow*
26	*Matzkin et al. (2020)*	skull reconstruction	pytorch*
27	*Zhang et al. (2021)*	TS-SSL	tensorflow*
28	*Prakash et al. (2020)*	Images denoising	tensorflow*

Funding

This research was supported by Jordan University of Science and Technology, Grant no. 20210418. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Jordan University of Science and Technology: 20210418.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Saeed Shurrab conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Rehab Duwairi conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

   This is a literature review article and doesn't have data.

REFERENCES

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34(11)**:2274–2282 DOI 10.1109/TPAMI.2012.120.

Altaf F, Islam SM, Akhtar N, Janjua NK. 2019. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access* **7**:99540–99572 DOI 10.1109/ACCESS.2019.2929365.

Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. 2018. Medical image analysis using convolutional neural networks: a review. *Journal of Medical Systems* **42(11)**:1–13 DOI 10.1007/s10916-017-0844-y.

Arjovsky M, Chintala S, Bottou L. 2017. Wasserstein generative adversarial networks. In: *International conference on machine learning*. PMLR, 214–223.

Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T. 2021. Big self-supervised models advance medical image classification. ArXiv preprint. arXiv:2101.05224.

Bai W, Chen C, Tarroni G, Duan J, Guitton F, Petersen SE, Guo Y, Matthews PM, Rueckert D. 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In: *International conference on medical image computing and computer-assisted intervention*. Cham: Springer, 541–549.

Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. ArXiv preprint. arXiv:1811.02629.

Bengio Y, Lamblin P, Popovici D, Larochelle H. 2007. Greedy layer-wise training of deep networks. In: *Advances in neural information processing systems*. 153–160.

Bezdek JC, Ehrlich R, Full W. 1984. FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences* **10(2–3)**:191–203 DOI 10.1016/0098-3004(84)90020-7.

Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in neural information processing systems*. Vol. 33. Red Hook, NY: Curran Associates, Inc, 9912–9924.

Chaitanya K, Erdil E, Karani N, Konukoglu E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: *Advances in neural information processing systems*. Vol. 33. Red Hook, NY: Curran Associates, Inc, 12546–12558.

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

Chen T, Kornblith S, Norouzi M, Hinton G. 2020. A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. Proceedings of Machine Learning Research, 1597–1607.

Chen X, Wang X, Zhang K, Zhang R, Fung K-M, Thai TC, Moore K, Mannel RS, Liu H, Zheng B. 2021a. Recent advances and clinical applications of deep learning in medical image analysis. ArXiv preprint. arXiv:2105.13381.

Chen X, Yao L, Zhou T, Dong J, Zhang Y. 2021b. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. *Pattern Recognition* **113**:107826 DOI 10.1016/j.patcog.2021.107826.

Chopra S, Hadsell R, LeCun Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1*. Piscataway: IEEE, 539–546.

Cuturi M. 2013. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems* **26**:2292–2300.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 248–255.

Doersch C, Gupta A, Efros AA. 2015. Unsupervised visual representation learning by context prediction. In: *Proceedings of the IEEE international conference on computer vision*. Piscataway: IEEE, 1422–1430.

Donahue J, Krähenbühl P, Darrell T. 2016. Adversarial feature learning. ArXiv preprint. arXiv:1605.09782.

Dong N, Kampffmeyer M, Voiculescu I. 2021. Self-supervised multi-task representation learning for sequential medical images. In: *European conference on machine learning and principles and practice of knowledge discovery in databases, lectures notes in computer Science, Volume 12977*. 779–794.

Dosovitskiy A, Fischer P, Springenberg JT, Riedmiller M, Brox T. 2015. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38(9)**:1734–1747.

Goodfellow I, Bengio Y, Courville A, Bengio Y. 2016. *Deep learning*. vol. 1. Cambridge, USA: MIT press.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial nets. *Advances in neural information processing systems* **27**.

Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, Piot B, kavukcuoglu K, Munos R, Valko M. 2020. Bootstrap your own latent—a new approach to self-supervised learning. In: *Advances in neural information processing systems*. vol. 33. Red Hook, NY: Curran Associates, Inc, 21271–21284.

Gutmann M, Hyvärinen A. 2010. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

Haghighi F, Taher MRH, Zhou Z, Gotway MB, Liang J. 2020. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: *International conference on medical image computing and computer-assisted intervention*. Cham: Springer, 137–147.

He K, Fan H, Wu Y, Xie S, Girshick R. 2020. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Piscataway: IEEE, 9729–9738.

He K, Zhang X, Ren S, Sun J. 2016a. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 770–778.

He K, Zhang X, Ren S, Sun J. 2016b. Identity mappings in deep residual networks. In: *European conference on computer vision*. Cham: Springer, 630–645.

Henaff O. 2020. Data-efficient image recognition with contrastive predictive coding. In: *International conference on machine learning*. Proceedings of Machine Learning Research, 4182–4192.

Hervella ÁS, Ramos L, Rouco J, Novo J, Ortega M. 2020a. Multi-modal self-supervised pre-training for joint optic disc and cup segmentation in eye fundus images. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Piscataway: IEEE, 961–965.

Hervella ÁS, Rouco J, Novo J, Ortega M. 2020b. Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. *Applied Soft Computing* **91**:106210 DOI 10.1016/j.asoc.2020.106210.

Hervella ÁS, Rouco J, Novo J, Ortega M. 2021. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications* **185**:115598.

Holmberg OG, Köhler ND, Martins T, Siedlecki J, Herold T, Keidel L, Asani B, Schiefelbein J, Priglinger S, Kortuem KU. 2020. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence* **2**(11):719–726 DOI 10.1038/s42256-020-00247-1.

Hu S-Y, Wang S, Weng W-H, Wang J, Wang X, Ozturk A, Li Q, Kumar V, Samir AE. 2020. Self-supervised pretraining with DICOM metadata in ultrasound imaging. In: *Machine learning for healthcare conference*. Proceedings of Machine Learning Research, 732–749.

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 4700–4708.

Ilse M, Tomczak J, Welling M. 2018. Attention-based deep multiple instance learning. In: *International conference on machine learning*. Proceedings of Machine Learning Research, 2127–2136.

Ioffe S, Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*. Proceedings of Machine Learning Research, 448–456.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K. 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence, volume 33*. 590–597.

Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. 2021. A survey on contrastive self-supervised learning. *Technologies* **9**(1):2.

Jamaludin A, Kadir T, Zisserman A. 2017. Self-supervised learning for spinal MRIs. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Cham: Springer, 294–302.

Jiao J, Droste R, Drukker L, Papageorghiou AT, Noble JA. 2020. Self-supervised representation learning for ultrasound video. In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. Piscataway: IEEE, 1847–1850.

Jing L, Tian Y. 2020. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(11):4037–4058.

Karpathy A. 2016. Convolutional neural networks for visual recognition. Online course. Stanford University.

Ker J, Wang L, Rao J, Lim T. 2017. Deep learning applications in medical image analysis. *IEEE Access* **6**:9375–9389.

Kermany D, Zhang K, Goldbaum M. 2018. Large dataset of labeled optical coherence tomography (oct) and chest x-ray images. Mendeley Data, version 3. DOI 10.17632/rscbjbr9sj.3.

Komodakis N, Gidaris S. 2018. Unsupervised representation learning by predicting image rotations. In: *International conference on learning representations (ICLR)*.

Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. 2021.
Self-Path: self-supervision for classification of pathology images with limited
annotations. *IEEE Transactions on Medical Imaging* **40(10)**:2845–2856.

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep con-
volutional neural networks. *Advances in Neural Information Processing Systems*
25:1097–1105.

Krull A, Buchholz T-O, Jug F. 2019. Noise2void-learning denoising from single noisy
images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern
recognition*. Piscataway: IEEE, 2129–2137.

Larsson G, Maire M, Shakhnarovich G. 2017. Colorization as a proxy task for visual
understanding. In: *Proceedings of the IEEE conference on computer vision and pattern
recognition*. Piscataway: IEEE, 6874–6883.

Le-Khac PH, Healy G, Smeaton AF. 2020. Contrastive representation learning: a
framework and review. *IEEE Access* **8**:193907–193934.

Li X, Hu X, Qi X, Yu L, Zhao W, Heng P-A, Xing L. 2021. Rotation-oriented collabo-
rative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on
Medical Imaging* **40(9)**:2284–2294.

Li X, Jia M, Islam MT, Yu L, Xing L. 2020a. Self-supervised feature learning via exploit-
ing multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical
Imaging* **39(12)**:4023–4033 DOI 10.1109/TMI.2020.3008871.

Li Y, Chen J, Xie X, Ma K, Zheng Y. 2020b. Self-loop uncertainty: a novel pseudo-
label for semi-supervised medical image segmentation. In: *International conference
on medical image computing and computer-assisted intervention*. Cham: Springer,
614–623.

Li Y, Chen J, Zheng Y. 2020. A multi-task self-supervised learning framework for scopy
images. In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*.
Piscataway: IEEE, 2005–2009.

Liu X, Zhang F, Hou Z, Mian L., Wang Z, Zhang J, Tang J. 2021. Self-supervised
learning: generative or contrastive. *IEEE Transactions on Knowledge and Data
Engineering*.

Lu MY, Chen RJ, Mahmood F. 2020. Semi-supervised breast cancer histology clas-
sification using deep multiple instance learning and contrast predictive coding
(Conference Presentation). In: *Medical imaging 2020: digital pathology, volume 11320*.
International Society for Optics and Photonics, SPIE.

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

Lučić M, Tschannen M, Ritter M, Zhai X, Bachem O, Gelly S. 2019. High-fidelity
image generation with fewer labels. In: *International conference on machine learning*.
Proceedings of Machine Learning Research, 4183–4192.

Luo Y, Pan J, Fan S, Du Z, Zhang G. 2020. Retinal image classification by self-supervised
fuzzy clustering network. *IEEE Access* **8**:92352–92362.

Maas AL, Hannun AY, Ng AY. 2013. Rectifier nonlinearities improve neural network acoustic models. In: *ICML workshop on deep learning for audio, speech and language processing, volume 30.* Citeseer, 3.

Mao HH. 2020. A survey on self-supervised pre-training for sequential transfer learning in neural networks. ArXiv preprint. arXiv:2007.00800.

Matzkin F, Newcombe V, Stevenson S, Khetani A, Newman T, Digby R, Stevens A, Glocker B, Ferrante E. 2020. Self-supervised skull reconstruction in brain CT images with decompressive craniectomy. In: *International conference on medical image computing and computer-assisted intervention.* Cham: Springer, 390–399.

Mena G, Belanger D, Linderman S, Snoek J. 2018. Learning latent permutations with Gumbel-Sinkhorn networks. In: *International conference on learning representations.*

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems.* 3111–3119.

Mitchell BR. 2021. Chapter 3 - overview of advanced neural network architectures. In: Cohen S, ed. *Artificial intelligence and deep learning in pathology.* Elsevier, 41–56 DOI 10.1016/C2018-0-02465-2.

Miyato T, Koyama M. 2018. cGANs with projection discriminator. In: *International conference on learning representations.*

Morano J, Hervella ÁS, Barreira N, Novo J, Rouco J. 2020. Multimodal transfer learning-based approaches for retinal vascular segmentation. In: *ECAI 2020 - 24th european conference on artificial intelligence, Santiago de Compostela, Spain.* IOS Press, 1866–1873 DOI 10.3233/FAIA200303.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted boltzmann machines. In: *International conference on machine learning.*

Nguyen X-B, Lee GS, Kim SH, Yang HJ. 2020. Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access* **8**:162973–162981 DOI 10.1109/ACCESS.2020.3021469.

Noroozi M, Favaro P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: *European conference on computer vision.* Cham: Springer, 69–84.

Ohri K, Kumar M. 2021. Review on self-supervised image recognition using deep neural networks. *Knowledge-Based Systems* **224**:107090 DOI 10.1016/j.knosys.2021.107090.

Oord AVD, Kalchbrenner N, Vinyals O, Espeholt L, Graves A, Kavukcuoglu K. 2016. Conditional image generation with PixelCNN decoders. In: *Proceedings of the 30th international conference on neural information processing systems, NIPS'16.* Red Hook, NY, USA: Curran Associates Inc, 47974805.

Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. 2016. Context encoders: feature learning by inpainting. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* Piscataway: IEEE, 2536–2544.

Prakash M, Buchholz T-O, Lalit M, Tomancak P, Jug F, Krull A. 2020. Leveraging self-supervised denoising for image segmentation. In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI).* Piscataway: IEEE, 428–432.

Radford A, Metz L, Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In: *4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, conference track proceedings.*

Raghu M, Zhang C, Kleinberg J, Bengio S. 2019. Transfusion: understanding transfer learning for medical imaging. In: *Advances in Neural Information Processing Systems, volume 32.* Red Hook, NY: Curran Associates, Inc.

Ronneberger O, Fischer P, Brox T. 2015. U-net: convolutional networks for biomedical image segmentation. In: *International conference on medical image computing and computer-assisted intervention.* Cham: Springer, 234–241.

Ross T, Zimmerer D, Vemuri A, Isensee F, Wiesenfarth M, Bodenstedt S, Both F, Kessler P, Wagner M, Müller B. 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery* **13(6)**:925–933 DOI 10.1007/s11548-018-1772-0.

Sarhan MH, Nasseri MA, Zapp D, Maier M, Lohmann CP, Navab N, Eslami A. 2020. Machine learning techniques for ophthalmic data processing: a review. *IEEE Journal of Biomedical and Health Informatics* **24(12)**:3338–3350 DOI 10.1109/JBHI.2020.3012134.

Schmarje L, Santarossa M, Schröder S-M, Koch R. 2021. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access.*

Setio AAA, Traverso A, de Bel T, Berens MSN, Van den Bogaard C, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B, van der Gugten R, Heng P, Jansen B, de Kaste MMJ, Kotov V, Lin JY, Manders JTMC, Sónora-Mengana A, García-Naranjo JC, Prokop M, Saletta M, Schaefer-Prokop C, Scholten ET, Scholten L, Snoeren MM, Torres EL, Vandemeulebroucke J, Walasek N, Zuidhof GCA, van Ginneken B, Jacobs C. 2016. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. CoRR. ArXiv preprint. arXiv:1612.08012.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. In: *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings.*

Snell J, Swersky K, Zemel R. 2017. Prototypical networks for few-shot learning. In: *Advances in neural information processing systems, volume 30.* Red Hook, NY: Curran Associates, Inc.

Sowrirajan H, Yang J, Ng AY, Rajpurkar P. 2021. MoCo pretraining improves representation and transferability of chest x-ray models. In: Heinrich M, Dou Q, de Bruijne M, Lellmann J, Schlfer A, Ernst F, eds. *Proceedings of the fourth conference on medical imaging with deep learning, volume 143 of proceedings of machine learning research.* Proceedings of Machine Learning Research, 728–744.

Spitzer H, Kiwitz K, Amunts K, Harmeling S, Dickscheid T. 2018. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: *International conference on medical image computing and computer-assisted intervention.* Cham: Springer, 663–671.

Sriram A, Muckley M, Sinha K, Shamout F, Pineau J, Geras KJ, Azour L, Aphinyanaphongs Y, Yakubova N, Moore W. 2021. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. ArXiv preprint. arXiv:2101.04909.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 1–9.

Tajbakhsh N, Hu Y, Cao J, Yan X, Xiao Y, Lu Y, Liang J, Terzopoulos D, Ding X. 2019. Surrogate supervision for medical image analysis: effective deep learning from limited quantities of labeled data. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. Piscataway: IEEE, 1251–1255.

Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. 2020. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Medical Image Analysis* **63**:101693 DOI 10.1016/j.media.2020.101693.

Taleb A, Lippert C, Klein T, Nabi M. 2021. Multimodal self-supervised learning for medical image analysis. In: *International conference on information processing in medical imaging*. Cham: Springer, 661–673.

Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, Lippert C. 2020. 3D self-supervised methods for medical imaging. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in neural information processing systems*. Vol. 33. Red Hook, NY: Curran Associates, Inc, 18158–18172.

Tao X, Li Y, Zhou W, Ma K, Zheng Y. 2020. Revisiting Rubiks cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Cham: Springer, 238–248.

Torrey L, Shavlik J. 2010. Transfer learning. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 242–264.

Tschannen M, Bachem O, Lucic M. 2018. Recent advances in autoencoder-based representation learning. ArXiv preprint. arXiv:1812.05069.

Van den Oord A, Li Y, Vinyals O. 2018. Representation learning with contrastive predictive coding. CoRR. ArXiv preprint. arXiv:1807.03748.

Vincent P, Larochelle H, Bengio Y, Manzagol P-A. 2008. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

Vu YNT, Wang R, Balachandar N, Liu C, Ng AY, Rajpurkar P. 2021. MedAug: contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. ArXiv preprint. arXiv:2102.10663.

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4):600–612 DOI 10.1109/TIP.2003.819861.

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and

localization of common thorax diseases. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 2097–2106.

Wu Z, Xiong Y, Yu SX, Lin D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 3733–3742.

Xie Y, Zhang J, Liao Z, Xia Y, Shen C. 2020. Pgl: prior-guided local self-supervised learning for 3d medical image segmentation. ArXiv preprint. arXiv:2011.12640.

Yamashita R, Nishio M, Do RKG, Togashi K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging* **9(4)**:611–629 DOI 10.1007/s13244-018-0639-9.

Zeiler MD, Krishnan D, Taylor GW, Fergus R. 2010. Deconvolutional networks. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. Piscataway: IEEE, 2528–2535 DOI 10.1109/CVPR.2010.5539957.

Zhang R, Isola P, Efros AA. 2016. Colorful image colorization. In: *European conference on computer vision*. Cham: Springer, 649–666.

Zhang R, Isola P, Efros AA. 2017. Split-brain autoencoders: unsupervised learning by cross-channel prediction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 1058–1067.

Zhang Y, Li M, Ji Z, Fan W, Yuan S, Liu Q, Chen Q. 2021. Twin Self-supervision Based Semi-supervised Learning (TS-SSL): retinal anomaly classification in SD-OCT images. *Neurocomputing* **462(C)**:491–505.

Zhang P, Wang F, Zheng Y. 2017. Self supervised deep representation learning for fine-grained body part recognition. In: *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. Piscataway: IEEE, 578–582.

Zhang X, Zhang Y, Zhang X, Wang Y. 2020. Universal model for 3D medical image analysis. ArXiv preprint. arXiv:2010.06107.

Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A., Prince JL, Rueckert D, Summers RM. 2021. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*.

Zhou Z, Sodha V, Siddiquee MMR, Feng R, Tajbakhsh N, Gotway MB, Liang J. 2019. Models genesis: generic autodidactic models for 3d medical image analysis. In: *International conference on medical image computing and computer-assisted intervention*. Cham: Springer, 384–393.

Zhu J, Li Y, Hu Y, Ma K, Zhou SK, Zheng Y. 2020a. Rubiks cube+: a self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis* **64**:101746 DOI 10.1016/j.media.2020.101746.

Zhu J, Li Y, Hu Y, Zhou SK. 2020b. Embedding task knowledge into 3D neural networks via self-supervised learning. ArXiv preprint. arXiv:2006.05798.

Zhu J-Y, Park T, Isola P, Efros AA. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. Piscataway: IEEE, 2223–2232.

51/51

Zhuang X, Li Y, Hu Y, Ma K, Yang Y, Zheng Y. 2019. Self-supervised feature learning for 3d medical images by playing a rubiks cube. In: *International conference on medical image computing and computer-assisted intervention.* Cham: Springer, 420–428.

51/51