Archives of Computational Methods in Engineering (2023) 30:2761–2775
https://doi.org/10.1007/s11831-023-09884-2

REVIEW ARTICLE

Self-supervised Learning: A Succinct Review

Veenu Rani[1] · Syed Tufael Nabi[1] · Munish Kumar[1] · Ajay Mittal[2] · Krishan Kumar[2]

Received: 26 December 2022 / Accepted: 5 January 2023 / Published online: 20 January 2023

Abstract
Machine learning has made significant advances in the field of image processing. The foundation of this success is supervised learning, which necessitates annotated labels generated by humans and hence learns from labelled data, whereas unsupervised learning learns from unlabeled data. Self-supervised learning (SSL) is a type of un-supervised learning that helps in the performance of downstream computer vision tasks such as object detection, image comprehension, image segmentation, and so on. It can develop generic artificial intelligence systems at a low cost using unstructured and unlabeled data. The authors of this review article have presented detailed literature on self-supervised learning as well as its applications in different domains. The primary goal of this review article is to demonstrate how images learn from their visual features using self-supervised approaches. The authors have also discussed various terms used in self-supervised learning as well as different types of learning, such as contrastive learning, transfer learning, and so on. This review article describes in detail the pipeline of self-supervised learning, including its two main phases: pretext and downstream tasks. The authors have shed light on various challenges encountered while working on self-supervised learning at the end of the article.

Keywords Contrastive learning · Machine learning · Self-supervised · Supervised learning · Un-supervised learning

1 Introduction

Machine learning (ML) and deep learning (DL) algorithms have revealed incredible performance in computer vision applications such as image recognition, object detection, image segmentation, and so on. These models are trained using either labeled data or un-labeled data. Supervised learning works on labeled data and unsupervised learning works on un-labeled data. Manually labeling data is a time-consuming and labor-intensive process [1]. Self-Supervised Learning (SSL) is a solution to the aforementioned issues that has emerged as one of the most promising techniques that does not necessitate any expensive manual annotations. The term "self-supervised learning" was first used in robotics, where labels were automatically assigned to training data to exploit the relationships between input signals and sensors. The basic idea behind SSL's operation is that while providing input, some parts are hidden, and the visible parts are used to predict the hidden parts. SSL differs from unsupervised learning in that it requires labels but does not require human labeling [2]. The fundamental process of SSL is depicted in Fig. 1, with a large unlabeled corpus of images as input. The ConvNet model is trained to predict the image's hidden portion based on the visible portion. The concept of self-supervised learning is inspired by the way infants teach [3]. Infants always learn through observation, common sense, their surroundings, and little interaction. All these factors contribute to his or her ability to self-learn. An infant's surroundings become a source of supervision for him, which helps in his understanding of how things work without constant supervision. The same idea is replicated in the machine via SSL, where the data supervises itself for training the model rather than having annotated labels, instructing the network on what is right or wrong. SSL employs two main concepts: auxiliary pretext tasks and contrastive learning [4]. In pretext tasks,

✉ Munish Kumar
munishcse@gmail.com

Veenu Rani
veenugoyal7@gmail.com

Syed Tufael Nabi
ertufail32@gmail.com

Ajay Mittal
ajaymittal@pu.ac.in

Krishan Kumar
k.saluja@pu.ac.in

[1] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

[2] University Institute of Engineering and Technology, Panjab University, Chandigarh, India

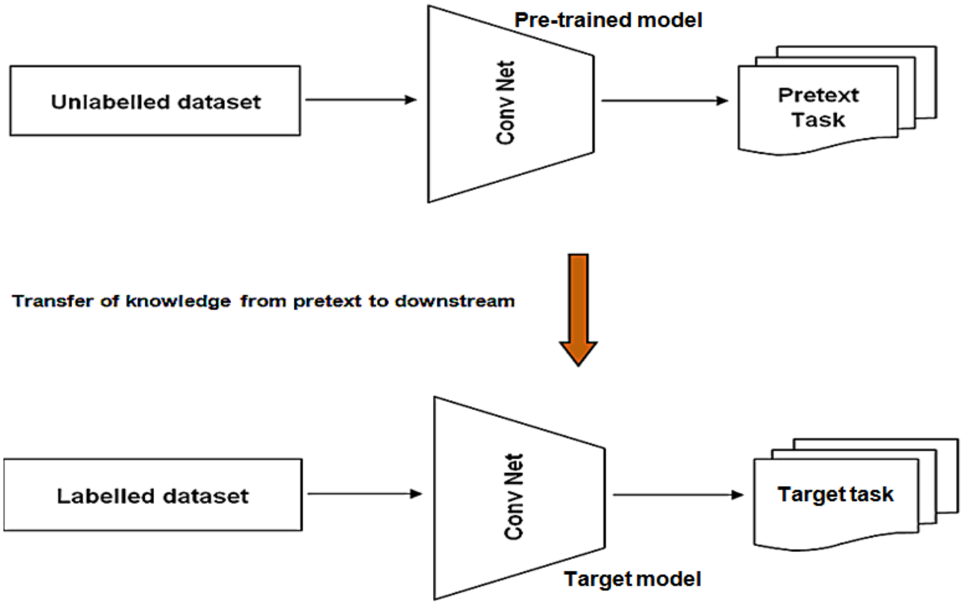


Fig. 1 Self-supervised learning process

pseudo labels are used for representations that were generated automatically by taking into account the attributes of the dataset. These pseudo labels are then used for classification, detection, and segmentation [5]. The auxiliary pretext is primarily used to fill in missing parts of an image [6], convert it to gray scale [7], predict hidden parts, and many other tasks. Contrastive learning, on the other hand, distinguishes between augmented image features. For example, in one image, a close-up view is captured, while in another, a distant view is captured. Considering the differences in perspectives between the two helps the model in learning. Before being integrated into computer vision, SSL had its origins in the NLP (natural language processing domain. Bidirectional Encoder Representations from Transformers (BERT is the most widely used SSL method in NLP. Neighbour sentence prediction, Neighbor word prediction, auto-regressive language modeling, and other NLP methods are used for SSL. As we all know, shuffling a single word in NLP can change the semantics of a sentence,the method described above can help with this.

2 Motivation Behind the Work

One important distinction between humans and machines is that humans can learn faster than machines by observing and sensing their surroundings. Machines, on the other hand, can take hours or even days to simulate. A large dataset is required to train machines to predict. Most artificial intelligence techniques require labelled datasets to make predictions. Labeling each data point manually or with data labelling software is a costly and time-consuming task. Unlabeled data, on the other hand, is available in plenty and easily. Thus, the motivation behind SSL is to learn useful depictions of the data from unlabeled data by making use of the self-supervision concept and then fine-tune these depictions with some labels for the supervised downstream task. These downstream tasks may range from simple to complex, like from image classification to semantic segmentation and object detection, etc. Therefore, there was a need for an hour to thoroughly discuss the SSL in detail. This article aims to provide a detailed illustration of SSL with applications in various areas, along with the literature review. The latter section also discusses the benefits and drawbacks of SSL. Also, the critical analysis along with future directions has been provided to make researchers familiar with the future research scope of SSL.

The article is organized as follows: Sect. 1 gives a brief introduction to SSL; Sect. 2 discusses the motivation for this work; and Sect. 3 discusses various learning algorithms. Section 4 discusses in detail the SSL algorithm, it's types, its learning tasks, the SSL applications and related work and the various terms used in the SSL paradigm. Section 5 discusses about various SSL datasets used in the medical domain and computer vision. Section 6 summarizes the critical analysis of the literature review, and at the end, Sect. 7 concludes the article by discussing the domain's future prospects.

3 Learning Algorithms

3.1 Supervised Learning

Supervised learning necessitates manual labelling of data, which slows the system's performance [8]. Supervised learning attempts to map input variables to output variables. Supervised learning is similar to a classroom setting in which a teacher is present to guide the students [2]. These models are created for a specific task and contain a large amount of manually annotated data. This data is divided into three categories at random: training data, testing data, and validation data. The success of computer vision programs is dependent on this annotated data, which is a time-consuming and expensive process to acquire. Figure 2 depicts the operation of supervised learning.

3.2 Semi-supervised Learning

Semi-supervised learning uses both labelled and unlabeled data to perform specific tasks [9]. First, some of the systems are trained using manually labelled data, and then the system is used to predict the remaining portion using unlabeled data, as shown in Fig. 3. Finally, a full dataset containing both labelled and pseudo-labeled datasets is used to train the network.

The goal of semi-supervised learning is to combine the benefits of both supervised and unsupervised learning techniques. Semi-supervised learning algorithms' main goal is to use unlabeled data to build a reliable model. These algorithms do not guarantee that including only unlabeled data will improve prediction performance because unlabeled data is only useful when it contains information useful for label prediction.

3.3 Weakly-Supervised Learning

Weakly-supervised learning is the process of learning from noisy or poorly labelled data. Because of the high cost of manually labelling data, it is difficult to obtain strong supervision information. In this scenario, it is necessary to build machines that can operate with minimal supervision. Weak supervision is classified into three types: incomplete supervision, inexact supervision, and inaccurate supervision [10]. Incomplete supervision is one of the, in which only a portion of the training data is labeled, and the rest is left unlabeled. The second type of supervision is inexact supervision, which contains only coarse-grained labels, and the last type is inaccurate supervision, in which the given labels do not depict the truth. Instagram is an example of weakly-supervised learning, in which users' hash tags are used to label data [2].

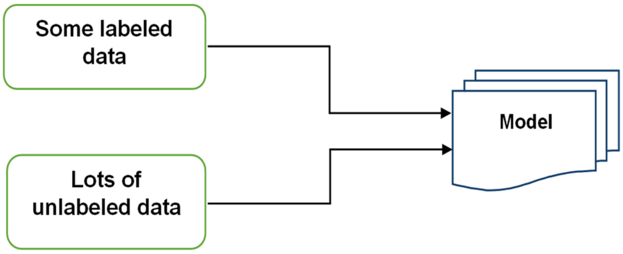


Fig. 3 Semi-supervised learning

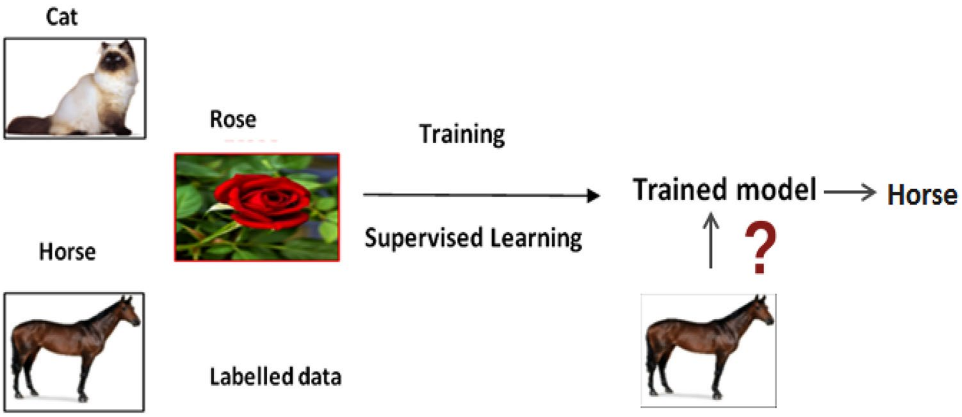
3.4 Semi-weak Supervised Learning

The combination of semi-supervised and weakly-supervised learning techniques is known as semi-weak supervised learning [11]. It adheres to the "student–teacher" framework, in which a weakly supervised dataset is first trained with noisy hash tags, referred to as the teacher model. This teacher model is further refined using an ImageNet labelled dataset, and the refined labels are used to train the target student model.

3.5 Unsupervised Learning

Unsupervised learning, on the other hand, seeks to discover implicit patterns in data that has not been labeled. Manual annotations are not required for unsupervised learning. Unsupervised learning is done through clustering. Unsupervised learning is the process of teaching a computer to operate on unlabeled data and allowing the algorithm to act on it without supervision. The machine's task is to sort unsorted data into groups based on similarities, patterns, and differences, as illustrated in Fig. 4.

Fig. 2 Supervised learning



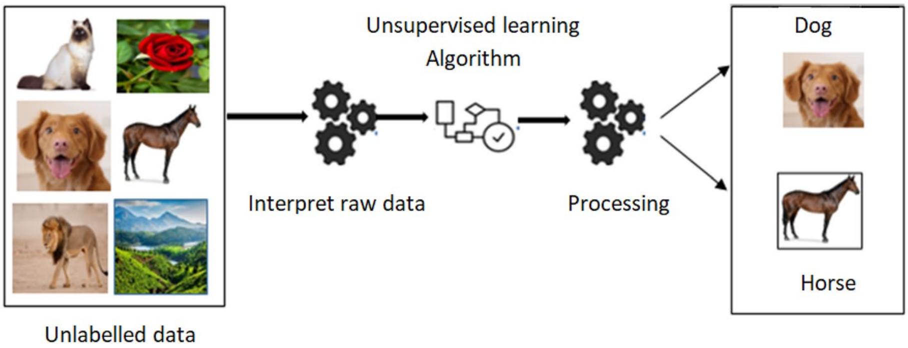


Fig. 4 Unsupervised learning

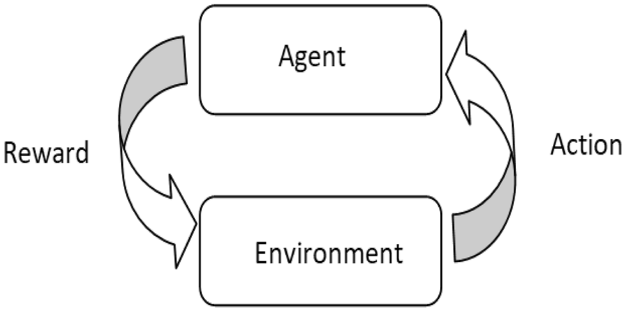


Fig. 5 Reinforcement learning

3.6 Reinforcement Learning (RL)

Reinforcement learning is a subset of machine learning in which an agent learns from trial-and-error feedback [12]. Feedback can be either a punishment or a reward. A variety of software and computers use it to determine the best possible solution in a given situation. The agent decides what to do with the given task in reinforcement learning. It is based on previous experience. Xin et al. [13] proposed self-supervised and reinforcement learning for sequential recommendation tasks. Their proposed model includes two output layers: one for self-supervised learning and one for reinforcement learning. The RL component acts as a regularize, directing the supervised layer's attention to specific rewards. Figure 5 depicts the operation of RL.

3.7 Increment Learning (IL)

As illustrated in Fig. 6, the goal of incremental learning is to learn new knowledge from new samples [14] and solve new tasks on a continuous basis without forgetting previous tasks by using new data. IL is a subset of machine learning technology that can handle more consistent applications with human behavior and thought. When learning with new knowledge, a back propagation method adjusts parameter weights based on losses on available sequential data. The model's performance on previously learned knowledge will suffer as a result. This is referred to as catastrophic forgetting (CF), and it is the primary issue with incremental learning.

3.8 Transfer Learning

Transfer learning is used to help learners improve their knowledge. Transfer learning refers to the transfer of knowledge from one domain to another [15]. For example, two people want to learn how to play the guitar. One has no musical knowledge, while the other has extensive musical knowledge. A person with a musical background

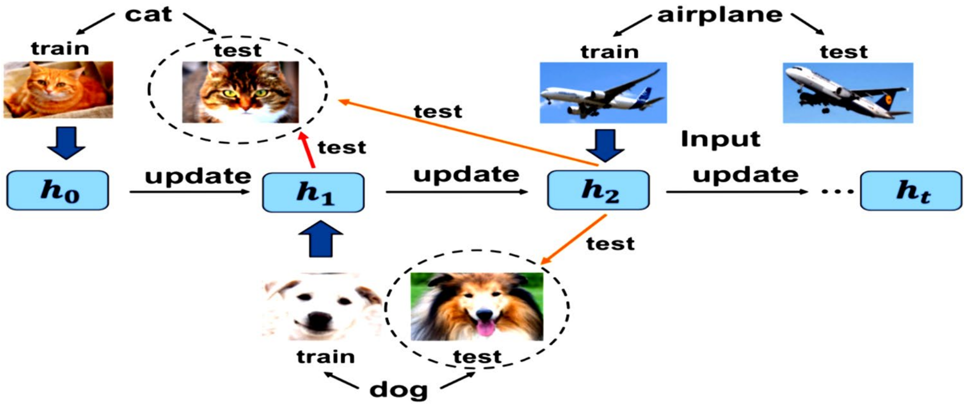


Fig. 6 Incremental learning

will be able to learn guitar more quickly by applying prior knowledge to a new domain [16]. The basic idea behind using transfer learning is to transfer the information contained in a trained model on a task with a large amount of data to an objective task with less data.

One of the most important learning algorithms is self-supervised learning (SSL) which has been discussed in the upcoming section.

4 Self-Supervised Learning (SSL)

SSL is considered to be the bridge between supervised learning and unsupervised learning. The SSL model trains itself using one part of the input data to learn the other part of the input data. This is also called predictive or pretext learning. The SSL algorithm has the ability to auto-generate the labels for un-labelled data, which converts the un-supervised model to a supervised model. Figure 7 shows the block diagram of the SSL algorithm.

4.1 Self-Supervised Learning (SSL) Tasks

SSL is a new approach that differs from other techniques. The main distinction between them is that SSL does not require manual labeling. SSL tasks are divided into two categories: pretext and downstream tasks. The former employs supervised learning to learn representations, with labels generated from the data itself. When this learning is complete, the model applies the previously learned representations to the subsequent tasks. Figure 8a, b depict various tasks performed by pretext and downstream tasks, respectively.

4.1.1 Pretext Task Learning Framework

In pretext tasks, the hidden portion of data is predicted using the visible portion. The pretext task can be applied to any type of data, such as images, audio, video, and so on [5]. Figure 9a–d show examples of pretext tasks such as colorizing an image [17], predicting a missing patch [6], estimating the rotation angel [2], jigsaw puzzle [18], and so on. This task allows machines to learn automatically by obtaining supervision directly from the data, without the use of annotations. Designing an appropriate pretext task necessitates domain knowledge.

4.1.1.1 Image Colorization Image colorization, as shown in Fig. 9a refers to the process of converting a colour image to a black-and-white image. Each pixel's full-color information is stored by the trained model. It is a pretext for learning vis-

Fig. 7 Block diagram of SSL

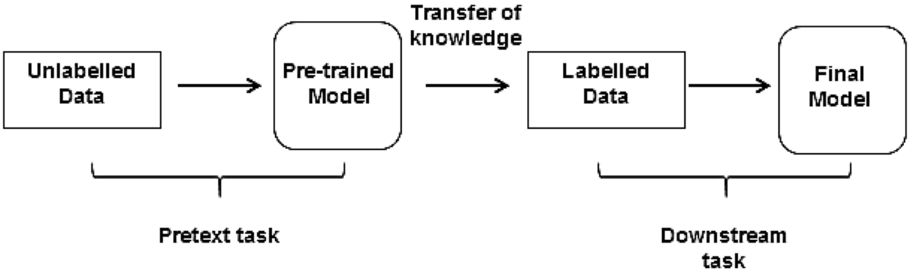
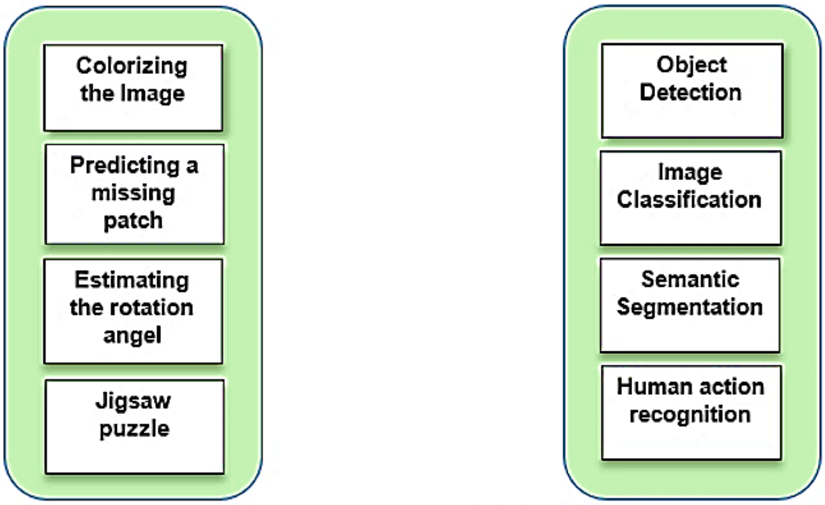


Fig. 8 Various tasks performed by **a** pretext and **b** downstream



(a) Various tasks performed by pretext

(b) Various tasks performed by downstream ual features. Treneska et al. [19] used an image colorization model based on a Generative Adversarial Network (GAN). The GAN model can produce the most realistic results. The extracted knowledge was transferred to two downstream tasks after applying the GAN model: multi-label image classification and semantic segmentation. In multi-label classification, a single image is assigned to multiple classes. The authors used 11,987 training images and 1713 testing images in their study. Semantic labels are assigned to each pixel of an image in semantic segmentation. They conducted experiments using the PascalVOC 2012 dataset.

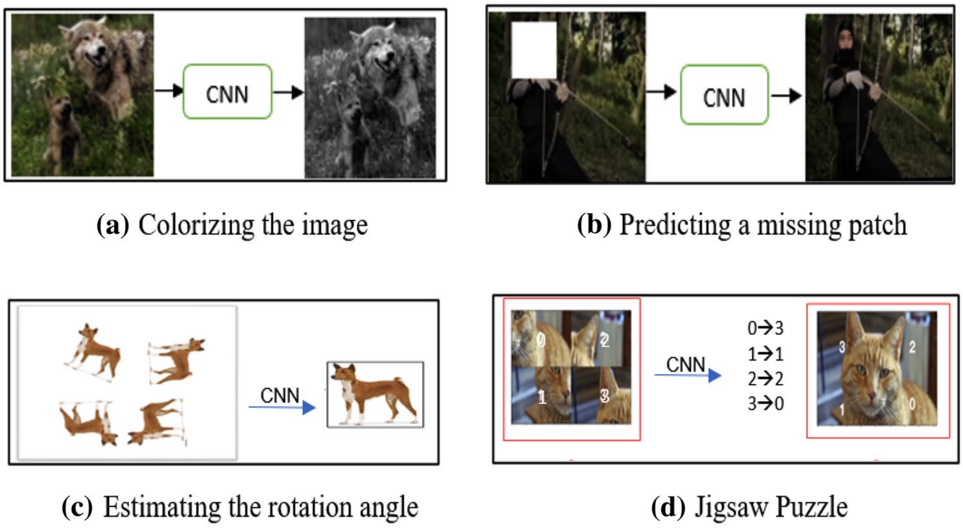
4.1.1.2 Predicting a Missing Patch A pretext task that can predict the position of an image patch is predicting a missing patch from an image. To predict the relative position of a patch within an image, models should be trained. Doersch et al. [18] proposed using SSL to predict a missing patch. They chose a random patch and predicted the relative position of an image's second and central patches, as shown in Fig. 9b. The patches are numbered 1 through 8. Following the selection of a patch, each patch is fed into a CNN that adheres to the AlexNet architecture. Both architectures are completely interconnected and share weights between layers. A gap was added between patches to prevent over-fitting. Softmax was used as a final layer to predict the relative position of each patch configuration.

4.1.1.3 Estimating the Rotation Angle Estimating the rotation is a task requiring instance discrimination [20]. Predicting picture rotations is a simple but effective method for identifying rotation discriminative features. If an image is rotated at any angle (0, 90, 180, or 270 degrees) and fed into a CNN model, the network model is pre-trained on pairs of rotated images from an unlabeled dataset. The network must understand the location, type, and pose of an object in an image in order to estimate the rotation angle of the original image. In Fig. 9c, an image is rotated three times, and the CNN model recovers the original image by learning the position of the object. Devgon et al. [21] developed a self-supervised learning model to estimate an object's rotation between the desired rotation and the current rotation. They used a trained model to estimate the rotation between two depth images.

4.1.1.4 Jigsaw Puzzle Solving a jigsaw puzzle requires not only knowledge of a single patch, but also knowledge of the relationships between different patches of the same image, as shown in Fig. 9d [22]. Understanding the patches discriminatory features helps in solving the puzzle. There is no shortcut method for predicting the right position of a patch, so multiple permutation functions are generated for each patch to find its right position [23]. Li et al. [24] solved Jigsaw puzzles using the GAN model. They created a multitask pipeline for solving unpaired image jigsaw puzzles. They classified jigsaw permutations into a separate branch and then used a GAN model to recover features from images in the correct order. The classification branch concentrated on the pseudo labels generated by shuffling the image pieces, whereas the GAN model concentrated on the semantic information of the pieces.

4.1.2 Downstream Task

Downstream tasks are primary tasks that define the model's purpose. Pretext tasks, also known as secondary tasks, allow the model to learn useful feature representation information that is used to complete downstream tasks. To ensure quality in downstream tasks, feature representation learned in pretext tasks should be calculated. The primary task downstream is to perform classification or object detection



(**a**) Colorizing the image

(**b**) Predicting a missing patch

(**c**) Estimating the rotation angle

(**d**) Jigsaw Puzzle

Fig. 9 Several exapmples of pretext task (**a**) colorizing the image (**b**) Predicting a missing patch (**c**) Estimating the rotation angle (**d**) Jigsaw Puzzle

with insufficient data labels, semantic segmentation, and action recognition. Down streaming can be accomplished in two ways: fine-tuning or using a linear classifier [25]. To achieve good performance, a small amount of data labeling is required in the downstream task. When the domain gap between self-supervised pre-training and the downstream task is smaller, performance on the downstream task is usually better.

4.1.2.1 Image Classification Image classification is the process of recognizing the category of each object in an image. Many networks are used for image classification, including AlexNet, ConvNet, ResNet, DenseNet, GoogLeNet, VGG, and others [26]. An image may contain multiple objects of different classes, but only one class label is used for each image. Image classification is used as a downstream task to estimate the quality of an image's features. To extract features from each image, a self-supervised learning model is used, which is then used to train a classifier.

Liu Object detection is a downstream task that recognizes the category of an object as well as its relative position in an image. This task is extremely important in computer vision applications such as robotics, autonomous driving, scene text detection, and so on [27]. The two most popular datasets for object detection are MOSOCO and OpenImage. Many ConvNet models, such as CNN, R-CNN [28], Fast-RCNN, Fast YOLO, and others, have been proposed to achieve high performance [29].

4.1.2.2 Semantic Segmentation Semantic segmentation is the process of assigning semantic labels to each pixel in an image. This task is critical in a variety of applications, such as human–machine interaction, robotics, and autonomous driving. Many networks have been used in downstream tasks, and semantic segmentation is no exception. These networks include VGG, ResNet, AlexNet, CNN, and FCN (fully connected network). FCN is a watershed moment in semantic segmentation because it employs a fully convolutional network to solve the problem.

4.1.2.3 Human Action Recognition (HAR) Human action recognition is the task of recognizing what people are doing in videos for a set of pre-defined action classes. This task necessitates the use of both spatial and temporal features. HAR has enabled the field to address one of its most pressing issues: using unlabeled data to build reliable recognition systems with only a few labeled training samples [30]. HAR is frequently used to evaluate the quality of video features learned using self-supervised methods. You and Wang [31] proposed using a view-enhanced jigsaw puzzle (VEJP) to recognize 3D human actions. VEJP captures multi-view data and forces the encoder to obtain view independent high-level features from the human skeleton. VEJP-extracted features are more robust and distinguishable.

4.2 Types of SSL Learning Techniques

The ConvNet model serves as the foundation of SSL, and its architecture influences the quality of visual representations learned through the pretext task. There are numerous pre-trained convolutional neural network models available, including ResNet50, ResNet50 v1, and upgraded versions. When a low-capacity model, such as AlexNet, is used for pre-training, little improvement can be obtained when compared to the Resnet50 family [26]. The major categories of SSL learning are Contrastive learning and non-contrastive learning.

4.2.1 Contrastive Learning

Contrastive learning is used to learn a representation or feature space that attracts and repels representations from similar images. Contrastive learning has a wide range of applications in computer vision and natural language processing. For example, in NLP, changing the position of a single word can change the semantics of a sentence. The goal of contrastive learning is to bring semantically related samples closer together while keeping dissimilar samples separate [32]. In contrastive learning one sample from the training data is used as an anchor, its augmented form is labeled as a positive sample, and the remaining examples in the training batch are labeled as negative samples. Because people can distinguish items without remembering every detail about them, their cognitive learning patterns give rise to the concept of contrastive learning. The contrastive learning framework is depicted in Fig. 10, and it consists of a large set of unlabeled data, accurate data augmentation, an encoder (ResNet-50, ConvNet, CNN, etc.), and a handcrafted pretext task. Data augmentation is the use of stochastic transformations to map an image into different perspectives, such as resizing, random flipping, color distortion, Gaussian blurring, and so on. Data augmentation can alter an image's visual appearance without changing its semantics. An encoder is used to extract features from images. The quality of features increases as the number of layers in a network increases. In contrastive learning, three types of encoders are used: image encoders, momentum encoders, and dictionaries [33]. To predict the unlabeled data, pretext tasks generate pseudo labels from these extracted features.

4.2.2 Non-contrastive Learning

Non-contrastive learning techniques only depends on positive sample pairs which means that the training data contains only related representations, for example the data may

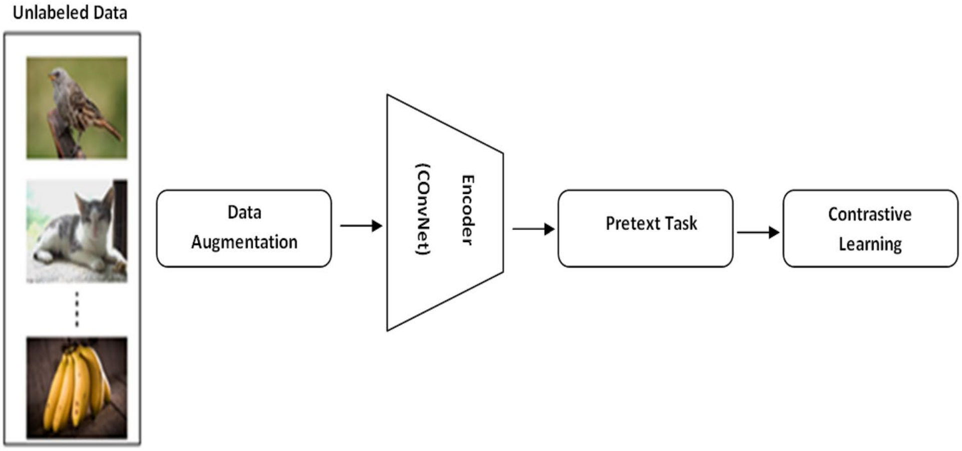


Fig. 10 Contrastive learning in computer vision

contain the two versions of same image of a dog like color and black and white but not the image of some unrelated data or negative samples like a picture of a building. This may be thought that the contrastive learning model being trained on positive samples only may be prone to collapse, but FAIR found that the model has the ability to learn good representations despite learning from only positive samples. The non-contrastive SSL model uses stop-gradient and extra-predictor operations to achieve better learning results. BYOL and SimSiam proved that using these operations, the non-contrastive learning model does not suffer representation collapse. The Table 1 given below mentions some differences between contrastive and non-contrastive learning.

4.3 Applications of SSL and Its State of the Art Work

4.3.1 Self-supervised Learning and NLP

SSL has become a hot research topic for researchers since 2018, when Google introduced the natural language processing (NLP) model. SSL's most fruitful results in the field of NLP are BERT and T5. A substantial amount of research has been conducted in this area. This section goes over a few of them. Zhou et al. [34] used the SSL regularization technique in NLP for text classification. They defined text classification as a key concept in NLP. Authors provided training texts that were encoded using a text encoder as input. In the encoded text, the authors defined two tasks: prediction and self-supervision. Both tasks use the same encoded text. They tested their model on 17 text classification datasets, attempting to minimize classification and regularization losses. Chen et al. [35, 36] used a hybrid SSL approach to regularize the training of a text classification task. Task adaptive pre-training (TAPT) and domain adaptive pre-training (DAPT) have been proposed by Gururangan et al. [37]. TAPT continues to train it on the target task's training dataset, after pre-training RoBERTa on large-scale corpora. DAPT is still pre-training RoBERTa on datasets with minor domain differences from datasets in target tasks. The difference between the SSL-Reg technique and TAPT and DAPT is that SSL-Reg uses a self-supervised task (for example, mask token prediction) to regularize RoBERTa fine-tuning, whereas TAPT and DAPT use separate tasks. TAPT and DAPT, on the other hand, use a self-supervised task for pre-training, with the text classification task coming first, followed by the self-supervised task. SSL-Reg method and TAPT are similar in that they both use texts in target tasks to conduct self-supervised learning. Sun et al. [38] proposed ERNIE 2.0, a framework for learning multiple tasks in incremental increments. They created several tasks and trained the ERNIE 2.0 model to extract lexical, syntactic, and semantic information from the training dataset. They used General Language Understanding Evaluation (GLUE) to assess the model's performance. As the size of the model in natural language processing grows, so

Table 1 Difference between contrastive learning and non-contrastive learning

Contrastive learning	Non-contrastive learning
Complex NN	Simple NN as compared to CL
Prone to dimensional collapse	Lesser dimensional collapse
Takes into consideration both positive and negative data samples	Takes into consideration only positive data samples
Minimises the distance between positive data and maximizes the distance between negative data	Leverages the positive data points from the dataset

does the need for large memory. To address this issue, Lan et al. [39] proposed a two-parameters reduction technique called ALBERT (A lite BERT) that use fewer parameters than traditional BERT. ALBERT has two parameters: factorized embedding parameterization and cross-layer parameter sharing. One large vocabulary matrix is decomposed into two small matrices in the first step. The size of the vocabulary embedding is kept separate from the hidden layers. If the number of parameters is increased, this separation will not increase the size of the memory. Multiple techniques are used to share parameters in cross-layer parameter sharing. When the network's depth increases, cross-layer parameter sharing prevents the parameters from growing.

4.3.2 Self-supervised Learning in Healthcare

Chowdhury et al. [40] discussed SSL implementation in healthcare as well as four other major areas: pixel to scalar pretext task, pixel to pixel pretext task, adversarial learning, and contrastive learning. The findings explain in their article that SSL has the ability to solve the problems caused by supervised learning. Jamaludin et al. [41] used longitudinal MRI scan data to train a Siamese CNN to learn embedding in which pairs of images from the same patient at different times in time are pushed further apart in the latent space and vice versa. These two pretext tasks' loss functions are combined, and a third pretext task is used to forecast vertebral body levels. The authors collected data from 1016 subjects for their experimental work. To address the issue of high traffic at network stations, Isravel et al. [25] incorporating SSL techniques into the software-defined networking domain. By sensing the channel, SSL can estimate the traffic behavior. Chen et al. [42] used SSL for context restoration in medical image analysis. They validated context restoration in three medical imaging problems: classification, localization, and segmentation. They used 2D ultrasound images for classification and CT images of abdominal organs for localization. They considered MRI of brain tumors to segment the imaging problem. The semantic features of images are taken into account by using SSL-based restoration. Kwasigroch et al. [43] proposed a fusion method that combines transfer and self-supervised learning. They used this method in skin treatment, which is the most delicate application in the medical domain. They divided the dermoscopic images into two categories: benign and malignant. The dataset used for this experiment includes 2000 training images, 150 validation images, and 600 testing images. Because this dataset is unbalanced, they encountered various problems while conducting experiments. They demonstrated in their proposed method that SSL can perform better even when only a small number of images are labeled. Ghesu et al. [44] proposed a contrastive learning and online feature clustering method. They collected 100 million 2D and 3D medical images from a variety of modalities, including radiography, computed tomography (CT), magnetic resonance imaging (MR), and ultrasonography. They extracted features from these images and using these extracted features to train the model in both supervised and unsupervised modes. They validate three medical problems: chest radiography abnormality assessment, brain metastasis detection in MR, and brain hemorrhage detection in CT image data. Spathis et al. [45] discussed the role of self-supervised learning in the medical domain. They collected ECG signal data and applied prominent SSL methods to it. They demonstrated that self-supervised learning can perform well with large amounts of data without the need for annotations. Nguyen et al. [46] proposed a method using SSL to solve a variety of medical problems. They have worked with image data containing spatial 3D information, such as CT and MR images. The system has been divided into two stages: classification and segmentation. They used Resnet 34 to predict the value of spatial information. This stage's output was used as the initial parameters for the second segmentation stage. The authors considered two medical domain problems for experiment purposes. The first issue is organ risk, and the second is detecting intracranial hemorrhage. The StructSeg dataset, which contains CT images of 60 lung cancer patients, was used for the Organ at Risk problem, and the RSNA dataset was used for Intracranial Hemorrhage detection. This dataset contains 17,079 images of patients.

4.3.3 Self-supervised Learning in Computer Vision

SSL has been widely used in computer vision applications such as object detection, image classification, graph classification, visual question answering, and so on. For object detection, two approaches are used: a one-stage detector and a two-stage detector. In a two-stage detector, the object proposal is generated first, and then the object is located and classified in the second stage, whereas in a one-stage detector, classification and bounding box location are done in the same stage. Self-supervised learning methods are rapidly gaining popularity. The basic idea behind SSL is to develop a model that can solve problems in the field of computer vision. Gidaris et al. [47] proposed a ConvNets-based method for recognizing 2D rotations in images. According to Huang et al. [48], SSL should be used for few shot object detection (FSOD) and instance segmentation. Object detection necessitates dense labeling, which is a time-consuming process,the FSOD method attempts to recognize previously unseen object classes based on a few labels. They discussed various object detection benchmark datasets and their evaluation matrices also in this work. Self-EMD is a method for learning spatial-visual representation proposed by Liu et al. [49]. They used more than one set of images from various perspectives. The image is then cropped to 224 × 224 and resized using various methods such as color distortion, random Gaussian blur, random horizontal flip, and so on. The COCO image dataset was used for experimentation. Amrani et al. [50] proposed a self-supervised learning method for detecting and retrieving an object from an unlabeled video. By listening to the video, this model captures similar frames with a common theme. For background rejection, contrastive learning was used, and new clusters were formed by a high level of label noise. Table 2 lists the major contributions of researchers in various fields, such as health care, natural language processing, image classification, object detection, and so on.

4.4 Taxonomy of SSL

Self-supervised learning has attracted many researchers for its excellent data efficiency. In this method, fewer labels and smaller samples are used to learn more by incorporating a neural network. In this section, we will go over some fundamental SSL terms.

4.4.1 Pseudo Labels

The data curation process must be automated to reduce manual input, and the number of labels required for good performance must be reduced [57]. Pseudo labels are labels that are assigned automatically by the network based on pretext tasks. For example, an image is provided as input, and the system performs some type of transformation on it, such as rotation, colorization, and so on. The transformed image is fed into ConvNet, which predicts the transformation via a pseudo labeling process. Figure 11 depicts this process.

4.4.2 Linear Classification

A typical assessment protocol involves training a linear classifier on top of (frozen) representations learned through self-supervised methods. It is evaluated for the classification accuracy of the learned classifier model on the ImageNet Val/Test set.

4.4.3 Pre-trained Models

A common assessment protocol involves training a linear classifier on top of (frozen) representations learned through self-supervised methods. The classification accuracy of the learned classifier model is evaluated on the ImageNet Val/Test set.

4.4.4 Fine-tuning the Model

The fairness of self-supervised learning is affected by model fine-tuning [58]. The design of all models, including their parameters, is copied into the target model, except for the output layer when fine-tuning models. These parameters are adjusted based on the target datasets. To achieve the highest level of performance, downstream tasks in SSL use fine-tuned models.

5 Sources of Datasets

This section will go over various datasets used for training and evaluating self-supervised visual features, natural language processing, image classification, the medical domain, and so on. These datasets are collected for self-supervised training without the use of human-annotated labels. Table 3 discusses various datasets used for computer vision, medical imaging, and natural language processing, along with their source links.

6 Critical analysis

When compared to supervised models, SSL has achieved remarkable success in the computer vision, NLP, and healthcare domains. SSL has some issues that researchers encounter, just like every coin has two sides. We have discussed a few of the pros and cons of SSL in this section.

- The vast majority of self-supervised pre-trained models, such as those in the ImageNet dataset, are trained on images with a single dominant object. The scene in applications such as self-driving cars contains several items, making it difficult to distinguish between two similar scenes.
- Finding context in satellite and medical images is extremely difficult due to their lack of structure. As a result, approaches like relative patch prediction and jigsaw puzzles are useless when dealing with such images.
- When dealing with less structured datasets containing medical and satellite images, a different augmentation method is required than when dealing with datasets for natural language processing.
- Creating a useful pretext assignment that allows a network to learn meaningful images/text is the most difficult aspect of self-supervised learning.
- After reviewing a large amount of literature, we discovered that as the size of the dataset grows, so does the system's performance. Larger datasets should thus be used whenever possible.
- Because SSL can process large data sets without relying on labels, an incomplete sentence in the NLP domain can be completed with a few words. Later words can be completed by understanding the semantics of the previous sentence.

Table 2 Major contribution by researchers

Authors name	Journal/conference	Highlights of the article	Dataset/number of subjects
Girshick et al. [28]	IEEE conference	Object detection and semantic segmentation	PASCAL VOC 2012
Jamaludin et al. [41]	Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support	Spinal MRI	1016 subjects
Chen et al. [42]	Medical Image Analysis	Used Context restoration strategy in medical imaging	–
Lee et al. [51]	IEEE explore conference on computer vision and pattern recognition	Object detection via recycling of bounding box annotation	PASCAL VOC and COCO
Feng et al. [20]	IEEE Conference CVF	Estimating the rotation angle in computer vision	PASCAL VOC 2007(20,044 testing images and 19,808 testing images)
Yun et al. [52]	IEEE Access	Proposed a human pose estimation method by creating a path for learning new rotational changes based on a self-supervised method	COCO key-point detection dataset
Amrani et al. [50]	IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops	Object detection and retreival from videos	HOW2 dataset
Liu et al. [49]	Computer vision and Pattern recognition	Self-EMD method for Object detection	COCO
Sun et al.[38]	Computation and language	Developed an ERNIE 2.0 framework for natural language processing	General Language Understanding Evaluation (GLUE)
Chen et al. [35, 36]	Machine learning	Propsed Contrastive learning framework for visual representation	ImageNet
Nguyen et al. [46]	IEEE Access	Used Self-supervised learning for medical image analysis	Two dataset: StructSeg contains images of 60 lung cancer patients, RSNA dataset contains 17,079 images. Furthermore, images were added at the time of processing
Kwasigroch et al. [43]	Electronic	Employed SSL to increase the performnace in skin care domain	2000 training, 150 validation, and 600 testing images
Huang et al.[48]	Computer vision and Pattern recognition	Few-shot object detection	–
Chowdhury et al. [40]	Informatics	Review on application of SSL in Healthcare	–
Zhou et al.[34]	Transactions of the Association for Computational Linguistics	Text classification in NLP	17 datasets
Pototzkyet al. [53]	DAGM conference on pattern recognition	Object detection in Autonomous driving	1,200,000 images,and 20,000 unlabeled images
Jain et al.[54]	Proceedings of the ACM on Interactive	SSL for human action recognition by using inertial sensors placed on human body	–
Breiki et al. [23]	Computer vision and Pattern recognition	Image classification in computer vision	Cassava plant disease dataset. That contain 12,000 unlabeled images and 6000 labeled
Ziegler & Asano [55]	Computer vision and Pattern recognition	Semantic segmentation in object detection	COCO
Treneska et al. [19]	Sensors	Image colorization for visual feature learning	COCO and PascalVOC 2012
Ding et al.[56]	IEEE Transactions on Pattern Analysis and Machine Intelligence	Proposed Deeply Unsupervised Patch Re-ID (DUPR) for unsupervised visual represenation learning	PASCALVOC
Ghesu et al. [44]	Computer vision and Pattern recognition	Solve image assessemnet problems in medical domain	100 million images (radiography, CT, MR imaging, ultrasonography etc.)

Fig. 11 Pseudo labeling process

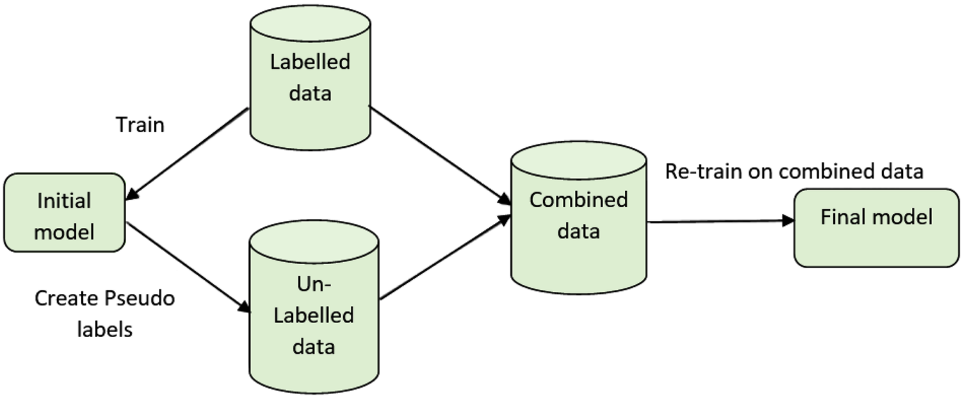


Table 3 Various datasets with their size and source links

Name of the dataset	Size	Classes	Category	Source
CIFAR 10	60,000 images	10	Computer vision	https://www.kaggle.com/datasets/fedesoriano/cifar10/download
CIFAR 100	50,000 images	100	Computer vision	https://www.kaggle.com/datasets/fedesoriano/cifar100/download
PASCAL VOC 2007	9963 images	20	Computer vision	https://www.kaggle.com/datasets/zaraks/pascal-voc-2007/download
PASCAL VOC 2012	11,530	20	Computer vision	https://www.kaggle.com/datasets/huanghanchina/pascal-voc-2012/download
Caltech 101	9146	102	Computer vision	http://www.vision.caltech.edu/Image_Datasets/Caltech101/
Covid-19 images	317		Health domain	https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset/download
MNIST	70,000	10	Handwritten digits	https://data.deepai.org/mnist.zip
Food 101	120,216	251		https://www.kaggle.com/datasets/dansbecker/food-101/download
Stanford Dogs	20,000	120	Images of dogs	https://www.kaggle.com/datasets/jessicali9530/stanford-dogs-dataset/download
Kinetics 400	400video clips for each human action	400	videos	https://academictorrents.com/download/184d11318372f70018cf9a72ef867e2fb9ce1d26.torrent
MS COCO	328,000	80	Computer vision	https://www.kaggle.com/datasets/awsaf49/mscoco-dataset/download
UCF 101	13,320	101	Videos for HAR	https://www.kaggle.com/datasets/pevogam/ucf101/download
HMDB51	6849	51 each category contains 101 clips	Videos for human motions	https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#Downloads
SVHN	73,257	10	Images of house numbers	https://www.kaggle.com/datasets/stanfordu/street-view-house-numbers/download
Flowers 102	2040	102	Images of flowers	https://www.kaggle.com/datasets/nunenuh/pytorch-challange-flower-dataset/download
FGVC aircraft	3334	100	Images of aircraft	https://www.kaggle.com/datasets/seryouxblaster764/fgvc-aircraft/download
Weather image recognition	6862	11	Images of weather	https://www.kaggle.com/datasets/jehanbhathena/weather-dataset/download

7 Conclusion and Future Directions

Self-supervised approaches have dominated supervised learning. They use the vast amounts of unlabeled data that are freely available. Self-supervised learning approaches have been shown to be effective in difficult downstream tasks such as image classification, object detection, image segmentation, and other tasks with little labeled input. The authors of this review article investigated the SSL application area as well as various types of learning. When combined with other learning methods, SSL can achieve greater success. Various pretext tasks generate different supervision signals, which can help the network learn more typical characteristics. In most existing self-supervised visual feature learning algorithms, ConvNet is trained to solve one pretext task. This review article provided a comprehensive overview of various learning schemes, the SSL pipeline, and recent research in this domain. Only a few studies have examined learning multiple pretext tasks for self-supervised feature learning. Self-supervised feature learning on multiple pretext tasks can be investigated further. The majority of self-supervised visual feature learning approaches currently available are focused on learning features for a single modality. If multiple data modalities from other sensors are available, the constraints between them can be used to train networks to learn features. As everyone is busy these days and wants to do most of the work automatically, researchers have a lot of room to explore many new techniques in this domain. SSL provides this level of security without the need for human intervention.

[1] https://research.aimultiple.com/self-supervised-learning/
[2] https://neptune.ai/blog/self-supervised-learning
[3] https://atcold.github.io/pytorch-Deep-Learning/en/week10/10-1/
[4] https://engineering.purdue.edu/ECE/News/2020/incremental-learning-in-online-scenario

Authors Contributions VR and STN: database collection, methodology, original draft. MK and AM: supervision, review and editing. KK: review and final draft.

Funding Nil.

Declarations

Conflict of interest The authors declare that they have no conflict of interest in this work.

References

1. Albelwi S (2022) Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. Entropy 24(4):551. https://doi.org/10.3390/e24040551
2. Ohri K, Kumar M (2021) Review on self-supervised image recognition using deep neural networks. Knowl-Based Syst 224:7090. https://doi.org/10.1016/j.knosys.2021.107090
3. Orhan, AE, Gupta VV, Lake BM (2007) Self-supervised learning through the eues of a child 2020, arXiv e-prints, arXiv-2007
4. Tao L, Wang X, Yamasaki T (2022) An improved inter-intra contrastive learning framework on self-supervised video representation. IEEE Trans Circ Syst Video Technol. https://doi.org/10.1109/tcsvt.2022.3141051
5. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F (2020) A survey on contrastive self-supervised learning. Technologies 9(1):2. https://doi.org/10.3390/technologies9010002
6. Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. IEEE Conf Comput Vis Pattern Recogn 2016:2536–2544. https://doi.org/10.1109/CVPR.2016.278
7. Larsson G, Maire M, Shakhnarovich G (2016) Learning representations for automatic colorization. In: Computer Vision—ECCV 2016, pp 577–593. https://doi.org/10.1007/978-3-319-46493-0_35
8. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ (2019) A systematic review on supervised and unsupervised machine learning algorithms for Data Science. In: Unsupervised and Semi-Supervised Learning, pp 3–21. https://doi.org/10.1007/978-3-030-22475-2_1
9. Engelen JEV, Hoos HH (2019) A survey on semi-supervised learning. Mach Learn 109(2):373–440. https://doi.org/10.1007/s10994-019-05855-6
10. Zhou ZH (2017) A brief introduction to weakly supervised learning. Natl Sci Rev 5(1):44–53. https://doi.org/10.1093/nsr/nwx106
11. Yalniz IZ, Jégou H, Chen K,Paluri M, Mahajan D (2019) Billion-scale semi-supervised learning for image classification. Comput Vis Pattern Recogn. arXivpreprint arXiv:1905.00546
12. Samsuden MA, Diah NM, Rahman NA (2019) A review paper on implementing reinforcement learning technique in optimising games performance. In: 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), pp 258–263. https://doi.org/10.1109/ICSEngT.2019.8906400
13. Xin X, Karatzoglou A, Arapakis I, Jose JM (2020) Self-supervised reinforcement learning for recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 931–940
14. Luo Y, Yin L, Bai W, Mao K (2020) An appraisal of incremental learning methods. Entropy (Basel, Switzerland) 22(11):1190. https://doi.org/10.3390/e22111190
15. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3:9. https://doi.org/10.1186/s40537-016-0043-6
16. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
17. Zhang R, Isola, P, Efros AA (2016) Colorful image colorization. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision–ECCV 2016. ECCV. Lecture Notes in Computer Science, vol 9907. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_40
18. Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: 2015 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2015.167

19. Treneska S, Zdravevski E, Pires IM, Lameski P, Gievska S (2022) Gan-based image colorization for self-supervised visual feature learning. Sensors 22(4):1599. https://doi.org/10.3390/s22041599
20. Feng Z, Xu C, Tao D (2019) Self-supervised representation learning by rotation feature decoupling. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2019.01061
21. Devgon S, Ichnowski J, Balakrishna A, Zhang H, Goldberg K (2020). Orienting novel 3D objects using self-supervised learning of rotation transforms. In: 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), pp 453–1460
22. Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision, pp 69–84. Springer, Cham.
23. Breiki FA, Ridzuan M, Grandhe R (2021) Self-supervised learning for fine-grained image classification. arXiv preprint arXiv:2107.13973
24. Li R, Liu S, Wang G, Liu G, Zeng B (2022) Jigsawgan: auxiliary learning for solving jigsaw puzzles with generative adversarial networks. IEEE Trans Image Process 31:513–524. https://doi.org/10.1109/tip.2021.3120052
25. Isravel DP, Silas S, Rajsingh EB (2021) Self-supervised learning approaches for traffic engineering in software-defined networks. In: Advances in Intelligent Systems and Computing, pp 511–522. https://doi.org/10.1007/978-981-33-6984-9_41
26. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional Neural Networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386
27. Jing L, Tian Y (2021) Self-supervised visual feature learning with deep neural networks: a survey. IEEE Trans Pattern Anal Mach Intell 43(11):4037–4058. https://doi.org/10.1109/tpami.2020.2992393
28. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2014.81
29. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. IEEE Conf Comput Vis Pattern Recogn 2016:779–788. https://doi.org/10.1109/CVPR.2016.91
30. Haresamudram H, Essa I, Plötz T (2022) Assessing the state of self-supervised human activity recognition using wearables. arXiv preprint arXiv:2202.12938
31. You W, Wang X (2022) View enhanced jigsaw puzzle for self-supervised feature learning in 3D human action recognition. IEEE Access 10:36385–36396. https://doi.org/10.1109/access.2022.3165040
32. Bhattacharjee A, Karami M, Liu H (2022) Text transformations in contrastive self-supervised learning: a review. arXiv preprint arXiv:2203.12000
33. He K, Fan H, Wu F, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. IEEE/CVF Conf Comput Vis Pattern Recogn 2020:9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975
34. Zhou M, Li Z, Xie P (2021) Self-supervised regularization for text classification. Trans Assoc Comput Linguist 9:641–656. https://doi.org/10.1162/tacl_a_00389
35. Chen T, Liu S, Chang S, Cheng Y, Amini L, Wang Z (2020a) Adversarial robustness: from self-supervised pre-training to fine-tuning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00078
36. Chen T, Kornblith, Norouzi M, Hinton G (2020b) A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002. 05709
37. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740
38. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H (2020) Ernie 2.0: a continual pre-training framework for language understanding. Proc AAAI Conf Artifi Intell 34(05):8968–8975. https://doi.org/10.1609/aaai.v34i05.6428
39. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
40. Chowdhury A, Rosenthal J, Waring J, Umeton R (2021) Applying self-supervised learning to medicine: review of the State of the art and medical implementations. Informatics 8(3):59. https://doi.org/10.3390/informatics8030059
41. 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, pp 294–302. https://doi.org/10.1007/978-3-319-67558-9_34
42. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D (2019) Self-supervised learning for medical image analysis using image context restoration. Med Image Anal 58:101539. https://doi.org/10.1016/j.media.2019.101539
43. Kwasigroch A, Grochowski M, Mikołajczyk A (2020) Self-supervised learning to increase the performance of skin lesion classification. Electronics 9(11):1930. https://doi.org/10.3390/electronics9111930
44. Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Neumann D, Patel P, Comaniciu D (2022) Self-supervised Learning from 100 Million Medical Images. arXiv preprint arXiv:2201.01283
45. Spathis D, Perez-Pozuelo I, Marques-Fernandez L, Mascolo C (2022) Breaking away from labels: the promise of self-supervised machine learning in intelligent health. Patterns 3(2):1410. https://doi.org/10.1016/j.patter.2021.100410
46. Nguyen XB, Lee GS, Kim SH, Yang HJ (2020) Self-supervised learning based on spatial awareness for medical image analysis. IEEE Access 8:162973–162981. https://doi.org/10.1109/ACCESS.2020.3021469
47. Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728
48. Huang G, Laradji I, Vazquez D, Lacoste-Julien S, Rodriguez P (2021) A survey of self-supervised and few-shot object detection. arXiv preprint arXiv:2110.14711
49. Liu S, Li Z, Sun J (2020) Self-EMD: self-supervised object detection without imagenet. arXiv preprint arXiv:2011.13677
50. Amrani E, Ben-Ari R, Shapira I, Hakim T, Bronstein A (2020) Self-supervised object detection and retrieval using unlabeled videos. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw50498.2020.00485
51. Lee W, Na J, Kim G (2019) Multi-task self-supervised object detection via recycling of bounding box annotations. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2019.00512
52. Yun K, Park J, Cho J (2020) Robust human pose estimation for rotation via self-supervised learning. IEEE Access 8:32502–32517. https://doi.org/10.1109/access.2020.2973390
53. Pototzky D, Sultan A, Kirschner M, Schmidt-Thieme L (2021) Self-supervised learning for object detection in autonomous driving. In: Bauckhage C, Gall J, Schwing A (eds) Pattern Recognition. DAGM GCPR 2021. Lecture Notes in Computer Science, vol 13024. Springer, Cham. https://doi.org/10.1007/978-3-030-92659-5_31

54. Jain Y, Tang CI, Min C, Kawsar F, Mathur A (2022) ColloSSL: collaborative self-supervised learning for human activity recognition. Proc ACM Interact Mob Wearable Ubiquitous Technol 6(1):1–28
55. Ziegler A, Asano YM (2022) Self-supervised learning of object parts for semantic segmentation. arXiv preprint arXiv:2204.13101.
56. Ding J, Xie E, Xu H, Jiang C, Li Z, Luo P, Xia G-S (2022) Deeply unsupervised patch re-identification for pre-training object detectors. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2022.3164911
57. Taherkhani F, Dabouei A, Soleymani S, Dawson J, Nasrabadi NM (2021) Self-supervised Wasserstein pseudo-labeling for semi-supervised image classification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12262–12272. https://doi.org/10.1109/CVPR46437.2021.01209.
58. Ramapuram J, Busbridge D, Webb R (2021) Evaluating the fairness of fine-tuning strategies in self-supervised learning. arXiv preprint arXiv:2110.00538

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.