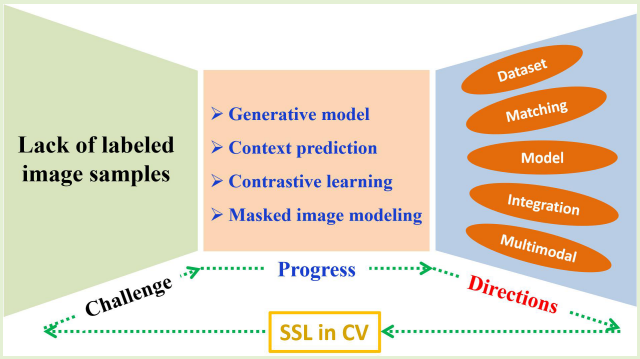
Progress and Thinking on Self-Supervised Learning Methods in Computer Vision: A Review

Zhihua Chen, Bo Hu, Zhongsheng Chen, *Member, IEEE*, and Jiarui Zhang

***Abstract*—Deep learning (DL) methods have been widely studied and applied in the field of computer vision (CV) over the past decades. The biggest disadvantage of classic DL methods is that they strongly rely on a large number of labeled samples. In engineering application, however, it is much expensive and even impossible to generate so many high-quality labeled samples. For this purpose, self-supervised learning (SSL) methods have become a research hot spot in CV in recent years due to their strong ability of learning representation without manually labeled images. So far, SSL has made strides in CV, but it is far from maturity and still faces some underlying challenges. The main purpose of this article is to review the latest development of SSL methods and applications, summarize key technologies and challenges, and discuss the trends. First, the development history of SSL methods in CV is outlined. Then, the existing SSL methods in CV are classified into four main categories and typical applications SSL in CV are summarized. Finally, key technologies of SSL are refined and future trends are discussed. This article can help researchers to quickly understand the current progress of SSL in CV.**



***Index Terms*—Computer vision (CV), deep learning (DL), future trends, key technologies, self-supervised learning (SSL).**

NOMENCLATURE

AE	Autoencoding.
AM	Additive manufacturing.
AR	Autoregressive.
BEIT	Bidirectional encoder representation from image transformers.
BiGAN	Bidirectional generative adversarial networks.
BYOL	Bootstrap your own latent.
CAE	Context autoencoder.
CL	Contrastive learning.
CNN	Convolutional neural network.
CSI	Contrasting shifted instances.
CSSL	Contrastive self-supervised learning.
CT	Computed tomography.
CV	Computer vision.
DIM	Deep infomax.
DL	Deep learning.
GAN	Generative adversarial networks.
GLCNet	Global style and local matching CL network.
GNN	Graph neural networks.
IBOT	Image BERT pretraining with online tokenizer.
LSTM	Long short-term memory.
MAE	Masked autoencoders.
MAE*	Mean absolute error.
MIM	Masked image modeling.
MLP	Multilayer perceptron.
MoCo	Momentum contrast.
MRI	Magnetic resonance imaging.
MSE	Mean squared error.
MSN	Masked Siamese network.
NCE	Noise-contrastive estimation.
NICE	Nonlinear independent components estimation.
NLP	Natural language processing.
PCL	Prototypical contrastive learning.
RSI	Remote sensing image.

Manuscript received 5 August 2024; accepted 11 August 2024. Date of publication 28 August 2024; date of current version 2 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 52377204. The associate editor coordinating the review of this article and approving it for publication was Prof. Changqing Shen. *(Corresponding authors: Zhihua Chen; Zhongsheng Chen.)*

Zhihua Chen, Bo Hu, and Jiarui Zhang are with the National Key Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing 210014, China (e-mail: chenzh@mail.njust.edu.cn; hubo@njust.edu.cn; 12112101576@njust.edu.cn).

Zhongsheng Chen was with the College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China. He is now with the College of Automotive Engineering, Changzhou Institute of Technology, Changzhou 213032, China (e-mail: chenzs@czu.cn).

Digital Object Identifier 10.1109/JSEN.2024.3443885

SiameseIM	Siamese image modeling.
SimCLR	Simple framework for contrastive learning of visual representations.
SimSiam	Simple Siamese.
SLAM	Simultaneous localization and mapping.
SSL	Self-supervised learning.
SwAV	Swapping assignments between views.
3D	Three-dimensional.
UAV	Unmanned aerial vehicles.
VAE	Variational autoencoding.
VICReg	Variance–invariance–covariance regularization.
ViT	Vision transformer.
VQ-VAE	Vector quantized variational autoencoder.

I. Introduction

CV IS the important subcategory of artificial intelligence that focuses on extracting high-level information from digital images or videos using advanced algorithms. By now, the existing CV algorithms can be divided into two main classes: traditional methods and DL-based methods. Traditional CV methods can date back to the past 10–20 years, which are often designed based on direct features of an image, such as edge detection, corner point detection, threshold segmentation, and so on. The biggest disadvantage of traditional CV methods is that it needs more human guidance to define proper features in an image. In consequence, lots of manual attempts and trials must be done in order to determine optimal features corresponding to different classes. In this case, manual feature extraction becomes increasingly difficult and expensive as the number of image classes increases. To overcome it, DL-based CV methods have been developed as a promising alternative, which are based on end-to-end learning without artificially extracting features. By this way, once a labeled-image dataset is given, deep neural networks can be used to automatically find the underlying features of images and then identify specific class of objects. Therefore, DL has been widely studied and applied in the fields of CV, such as image classification [1], [2], object detection [3], [4], and semantic segmentation [5], [6].

According to the development history, DL-based CV methods can be mainly classified into supervised learning, semi-supervised learning, and unsupervised learning methods. Early studies mainly focused on supervised learning algorithms, which strongly depended on labeled image samples. In this case, a large amount of high-quality labeled datasets are always needed to achieve good performance. For example, several famous labeled datasets in different fields have been generated and shared for supervised learning, such as ImageNet [7] (for image classification), VOC [8] (for object detection and segmentation), and COCO [9] (for object detection and segmentation). In engineering applications, however, annotating lots of dataset is a very labor-intensive, time-consuming, and expensive task [10], [11]. In particular, it may be very difficult to collect desirable images in some applications. For instance, some kinds of insect pests in the agricultural field are never seen before [12], and product defects in industry are often rarely and randomly generated. In these cases, it is hard to label images in advance [13]. In order to release the reliance on labeled images, researchers have to develop other new methods, including semi-supervised and unsupervised learning algorithms. Semi-supervised learning is a branch of machine learning that combines a small amount of labeled images and a large amount of unlabeled images to train deep models. In the field of CV, it is relatively easy to obtain lots of unlabeled images, but much difficult or expensive to annotate them. Compared with supervised learning, semi-supervised learning has fewer demands on labeled images. Furthermore, unsupervised learning can learn underlying features from unlabeled data without human supervision. The advantages of unsupervised learning include requiring no manually labeled images, finding unknown patterns, and so on. In a sense, unsupervised learning is much closer to true artificial intelligence, so that it has become an important direction in CV.

SSL is in some sense a subset of unsupervised learning. Up to now, SSL has been studied in CV for ten years. The key idea of SSL methods is to generate supervisory samples, rather than relying on external labeled samples. Once an SSL model learns how to represent images, it can be used for downstream tasks with few labeled images. Different from classical unsupervised learning methods, such as clustering or dimension reduction, SSL methods can learn image representation from supervised signals generated by them. SSL has several outstanding advantages. The first one is that SSL methods do not require manual annotations to train a model, so that a large amount of raw data can be directly utilized to learn useful features. The second one is that SSL models can scale to unprecedented sizes, which can make them have more capacity. The third one is that SSL models can deal with multiple tasks due to the strong generalization ability, which makes it possible to transfer their knowledge to other downstream tasks. Up to now, the superiority of SSL has been widely validated in CV.

Due to the outstanding advantages, SSL has attracted more and more attention in the past and many studies have been done from different perspectives. So far, several review papers have been published, but some of them only focused on SSL algorithms themselves [14], [15], [16], and others paid more attention on the applications of SSL models in specific CV fields [17], [18], [19]. In particular, SSL methods have been developed quickly in the past several years, and new models and applications were increasingly reported. However, SSL is far from maturity and still faces some underlying challenges. In order to quickly grasp the latest development and understand the whole status, it is very significant to give a comprehensive review of SSL methods in CV.

In summary, the main contributions of this work can be summarized as follows.

1) The development history of SSL methods in CV is outlined, and the existing SSL methods in CV are classified into four main categories, which are summarized in detail.
2) Current applications of SSL methods in CV are investigated based on the two aspects of task-related and field-related applications, respectively.

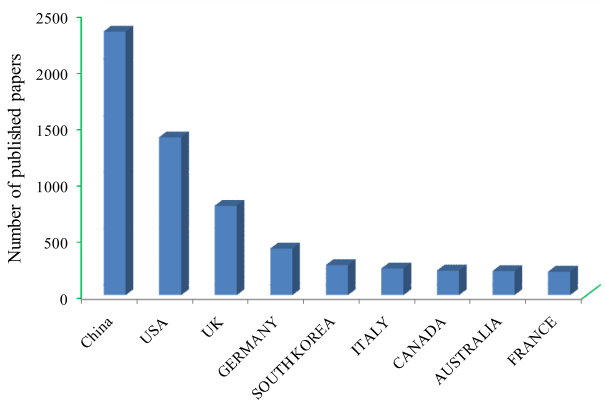


Fig. 1. Published WoS-indexed papers rated by countries.

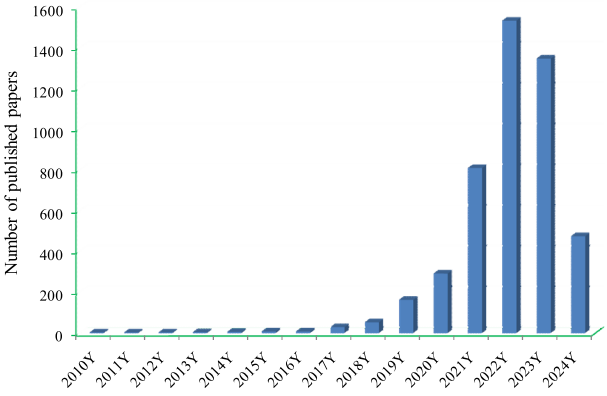


Fig. 2. Published WoS-indexed papers rated by years.

3) Key technologies of SSL in CV are mainly summarized as data augmentation, CV pretext task design, backbone network design, and loss function design. Based on the survey, future trends of SSL in CV are discussed.

The remainder of this article is organized as follows. The definition and development history of SSL is summarized in Section II. In Section III, the SSL methods are divided into different four main categories. Then, different applications of SSL methods in CV are summarized in Section IV, and some key technologies of SSL in CV are addressed in Section V. In Section VI, several potential trends of SSL are discussed. Finally, brief conclusion is presented in Section VII.

II. Progress of SSL Methods in CV

According to the definition in WiKi, SSL refers to a machine learning paradigm, which processes unlabeled data to obtain useful representations contributing to downstream tasks. The greatest feature of SSL methods is that they can first learn useful representations from unlabeled samples using self-supervised algorithms and then fine-tune the representations with few labeled samples for downstream tasks. In the beginning, SSL was widely used in the field of NLP. Later, it was introduced and extended to solve a variety of CV tasks.

The authors have searched published papers from 2010 to 2024 in the Web of Science by using the two subjects of "SSL" and "CV." The detailed results are shown as Figs. 1 and 2. It can be seen that: 1) the leading four countries include China, USA, U.K., and Germany. In particular, rapid progress is being achieved by China in recent years and 2) few papers were published before 2019. However, more and more works have been reported since 2021 and SSL is becoming a research hot spot in the field of CV.

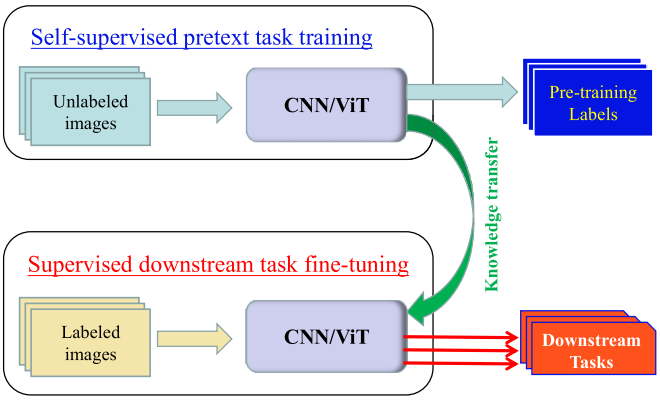


Fig. 3. Classical SSL workflow used for CV.

The development process of SSL methods in CV has gone through several stages. Early works on SSL were investigated as a class of unsupervised learning methods [20], [21], [22], [23], [24], [25], [26], [27], and they were mainly used for image reconstruction. Later, few works pointed out that it was very promising to apply the learned features of these models for new tasks [28], [29]. Then, some researchers began to design specific pretext tasks to train the models, which were used for other downstream tasks [30], [31], [32], [33], [34], [35], [36], [37], [38]. The results showed that the SSL methods performed well and could significantly reduce the gap between unsupervised and supervised learning methods. In 2018, CL began to rise in the field of CV, which focused on extracting meaningful representations by contrasting positive and negative pairs of instances, rather than using labeled samples. Thus, CL is a novel class of SSL methods. Wu et al. [39] proposed the first CL model. Since then, CL has become increasingly popular in the field of SSL. Chen et al. [40] proposed a representative CL method called SimCLR. By using it, the performance of SSL was first proved to be better than that of supervised learning in specific CV tasks. In recent years, another new type of SSL method called MIM was proposed. Typically, MAE was proposed by He et al. [41], which was an excellent representative of MIM methods. In particular, transformers have recently emerged as an alternative to CNN for visual recognition in 2017 [42]. In 2020, ViT is first proposed for CV [43], which outperforms CNN across various domains and settings. The integration of SSL and ViT brings new insights into the field of CV.

Generally speaking, a typical SSL workflow used for CV is shown in Fig. 3, which consists of two main parts: a pretext task and a downstream task. At the pretext-task stage, the model is trained by unlabeled datasets to learn the representation of image features by solving a predefined pretext task. After the pretraining is completed, the model parameters are fine-tuned and applied for downstream tasks, such as classification, detection, and segmentation.

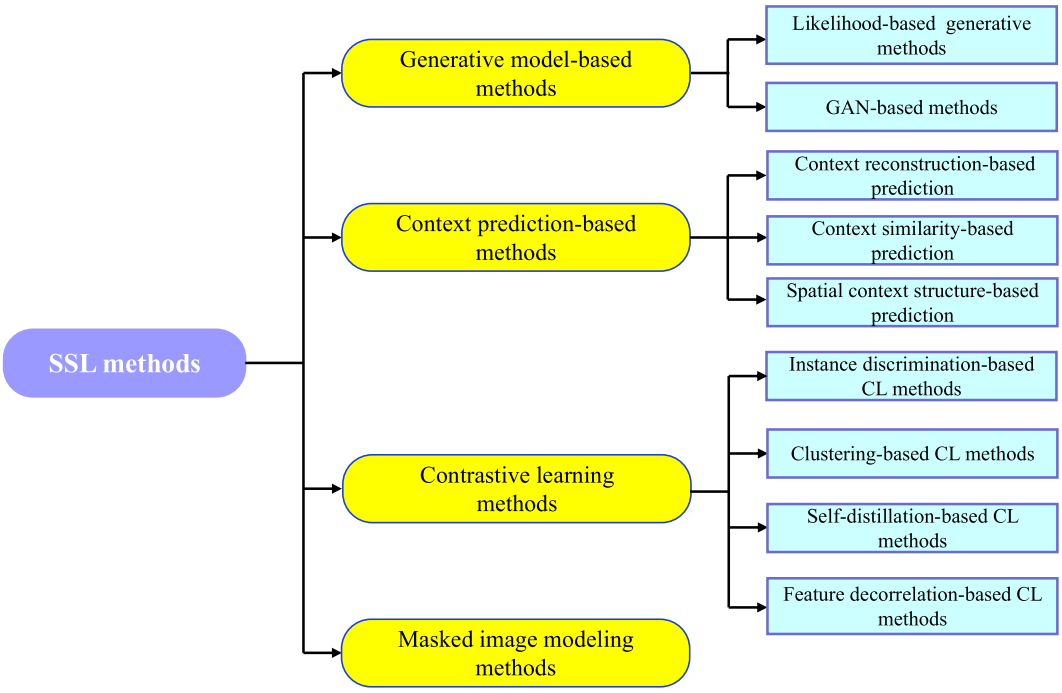


Fig. 4. Classification of SSL methods in CV.

At the same time, SSL has its own disadvantages. The first one is the demand of high computing power. SSL needs significant computing power to scan vast amounts of unlabeled data and then label them. The second one is the low accuracy. SSL generates its own labeled dataset without external knowledge or supervision, so that we cannot tell how accurate these labels are. Then, the accuracy strongly depends on the quality of unlabeled dataset. Thus, SSL may not perform as well as supervised learning on some tasks.

III. Classification of SSL Methods in CV

Due to the outstanding advantages, many SSL methods and their variants have been investigated for different CV tasks. According to the working principle, the existing SSL methods in CV can be classified into four main categories: generative methods, context prediction methods, CL methods, and MIM methods. The detailed classification of SSL methods in CV is shown in Fig. 4. Furthermore, the generative methods mainly include likelihood-based methods and GAN-based methods. The context prediction methods can be realized based on context reconstruction, context spatial structure, or context similarity. The CL methods can be further divided into two subclasses according to whether negative samples are required or not. The next basic principles and features of each SSL method will be outlined.

A. Generative Model-Based SSL

The core of generative models is to generate or reconstruct images from input images. In this way, generative models can learn representations of these images, which then can be used for downstream CV tasks. In nature, this process can also be considered as a kind of pretext tasks. Therefore, generative methods are often considered as one class of SSL methods. By now, two common generative methods are likelihood-based methods and GAN-based methods.

1) Likelihood-Based Generative Methods: Likelihood-based generative methods always depict the distribution of the data directly with a likelihood function, such as AR models [20], [21], flow-based models [22], [23], [44], and VAE models [24], [25], [45].

In *AR-based generative models*, an explicit density function is created to maximize the likelihood of training images. The basic principle is to generate images pixel-by-pixel by using AR connections. The conditional distribution of any pixel is predicted based on its neighborhood pixels. There are two AR-based generative models in CV, namely, PixelRNN [20] and PixelCNN [21]. In the PixelRNN model, the distribution of image pixels is modeled by a 2-D LSTM network. While in the PixelCNN model, the gated CNN is used to model the distribution of image pixels. In both works, the images are scanned both row-by-row and pixel-by-pixel. Generally speaking, PixelRNN models have better performances than PixelCNN models, but the training of PixelCNN models is faster.

Flow-based generative models explicitly represent a probability distribution by leveraging normalizing flow and use the change-of-variable law of probabilities for converting a simple distribution to a complex one. NICE was the first flow-based generative model [22], where multiple additive coupling layers were used to simplify the computation of the Jacobi determinant. By this way, the fitting ability of the model became stronger, so that its probability distribution could be directly fitted. Later, two new models called RealNVP [23] and Glow [44] were developed and refined to synthesize realistic images. RealNVP is a generative model that applies real-valued nonvolume preserving transformations to estimate density [23], which is composed of coupling layers performing

TABLE I
COMPARISON OF SEVERAL LIKELIHOOD-BASED GENERATIVE MODELS IN CV

Methods	Learned density	Parameter number	Train speed	Generated images
AR-based models	Exact density	Least	Highest	Good
Flow-based models	Exact density	Most	Lowest	Excellent
VAE models	Approximated density	Moderate	Moderate	Average

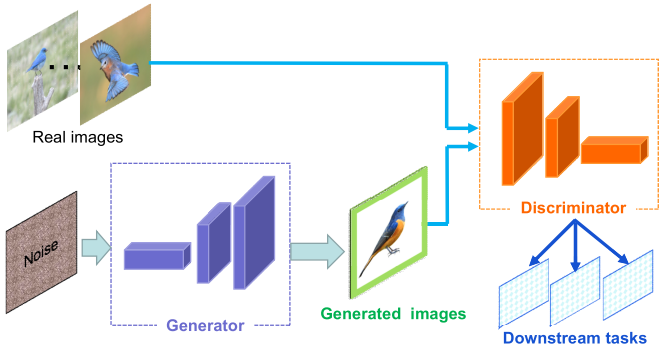


Fig. 5. Schematic of GAN-based SSL.

invertible operations. Glow is the extension of NICE and RealNVP by introducing invertible 1×1 convolutions, so that it can simplify the model structure.

VAE-based generative models are another specific kind of generative models, which use a feed forward to reconstruct the output images from input images. Similar to classical AE models, a VAE model is also composed of an encoder and a decoder [24]. The encoder converts a high-dimensional input image to a low-dimensional feature space (i.e., the latent space). The decoder recovers the encoded information back to the high-dimensional image space. For nonregularized latent space in classical AE models, the VAE model imposes a constraint on the latent distribution by forcing it to be a normal distribution. VQ-VAE model is a powerful variant of VAE model [25], which applies vector quantization to obtain a discrete latent representation. The VQ-VAE model can effectively represent objects spanning many pixels in an image. Based on the VQ-VAE model, the VQ-VAE-2 model was proposed by introducing a self-attention AR model as a prior [45]. In consequence, high-quality images could be synthesized by using hierarchical multiscale latent maps.

Each class of likelihood-based generative methods has its advantages or limitations. The comparison of these methods is shown in Table I. Both AR-based and flow-based generative models can directly learn exact probability distribution, while VAE-based generative models only learn the approximated probability distribution. Among the three classes of models, AR-based generative models have the least number of parameters and training time, while flow-based generative models have the most ones. Generally speaking, flow-based generative models can generate the most realistic images, while the images generated by VAE-based generative models are not real enough.

2) GAN-Based Methods: GANs are a kind of implicit generative models that are based on the game theory [26]. The architecture of GANs-based SSL is shown in Fig. 5, which consists of two neural networks (i.e., the generator and the discriminator). The goal of the generator is to artificially generate fake images and the goal of the discriminator is to identify whether the artificially created images are true or not. In this way, the two neural networks contest with each other according to a zero-sum game, so that the model is trained in an SSL manner. In the field of CV, GANs are generally used to synthesize high-quality images, and the trained discriminator can be used as the pretrained model to perform SSL. For example, Radford et al. [28] proposed deep convolutional GAN, where the discriminator was used as a feature extractor for several classification tasks. The disadvantage of classical GANs is that it is difficult to learn rich features for images with arbitrary distributions because there is no means to map the images back to the latent space. To solve this issue, Donahue et al. [29] proposed BiGAN, where an encoder was added into the standard GAN framework to realize the image mapping. Based on the BiGAN model, the BigBiGAN model was improved to lift the representation learning ability by adding an encoder and modifying the discriminator [46]. The results showed that the BigBiGAN model performed well for image representation learning on the ImageNet dataset.

B. Context Prediction-Based SSL

In the field of CV, context refers to any information related to the objects in an image, which can exist in the image itself or be generated from the image. Context prediction methods focus on how to skillfully design the prediction tasks according to the context. In this way, the model can learn useful and transferable visual features during solving these tasks. At the pretext-task stage, a large number of unlabeled images are processed, so that pseudo-labels are generated according to the designed tasks. Then, these labeled images are used to train the model as supervised samples. Generally speaking, the main aim of these tasks is to predict specific contextual information, such as colorization [31], [32], [35], relative positions [30], and rotation degrees [38]. After pretraining, the model is fine-tuned and applied for other CV tasks. Depending on how contextual information is used, context prediction-based SSL methods can be further divided into three categories: context reconstruction-based SSL, spatial context structure-based SSL, and context similarity-based SSL.

1) Context Reconstruction-Based SSL: The context reconstruction-based prediction methods are mainly utilized in those application scenarios, where some contextual information in an image is usually removed or masked and the missing parts need to be rebuilt. For example, the missing information can be colors of images. In this case, the idea of colorization is to let the model automatically predict original colors of grayscale images, so that more useful features can

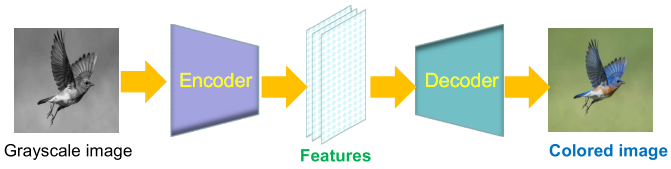


Fig. 6. Schematic of colorization task.

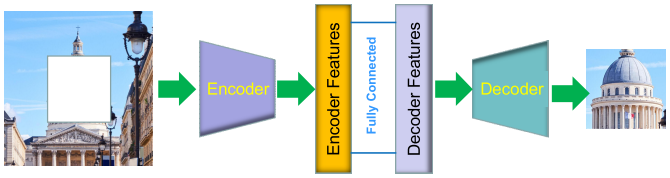


Fig. 7. Architecture of context encoder.

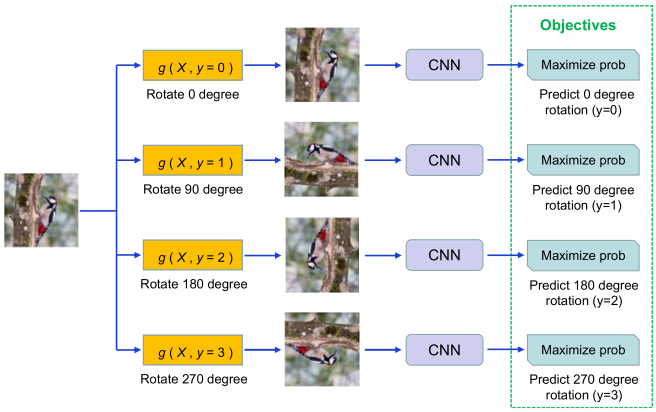


Fig. 8. Schematic of rotation prediction task.

be learned. The principle of the colorization task is shown in Fig. 6. Zhang et al. [31] first proposed to use image colorization as a pretext task, where the colorization task was transformed to a classification task in order to efficiently train the model. Their work showed that colorization was a useful pretext task for SSL. Later, Larsson et al. [32], [35] further improved the colorization models and validated the feasibility of using automatic colorization as a pretext task. Larsson et al. [32] showed that the CNN combined with colorization was utilized to better connect semantic knowledge with color distributions, and then, the color histogram of the image was predicted. Larsson et al. [35] showed that the colorization task in SSL was deeply analyzed by considering the effects of loss, network architecture, and training settings. Another case of missing information is that a port of an image is completely masked or removed. For instance, Pathak et al. [33] proposed a novel pretext task based on image inpainting. A portion of an image was removed, and then, a context encoder was trained to automatically inpaint the missing region based on the remaining portion. The architecture of the context encoder is shown in Fig. 7.

2) Context Similarity-Based SSL: In addition to manually designing tasks related to visual context, it is feasible to directly mine contextual similarity using unsupervised algorithms, such as clustering. Clustering is one classical kind of unsupervised learning methods, which can be used to classify images according to the context similarity without any labels. Caron et al. [47] first introduced the clustering into the field of SSL and proposed the method called DeepCluster. The images were iteratively clustered by using the k-means method, and then, the clustered classes were used as pseudo-labels to train the model. Inspired by this idea, more works were done to combine clustering with SSL. Later, Caron et al. [48] further expanded their work by predicting the rotation degrees. Yan et al. [49] proposed a new framework called ClusterFit, where an additional network trained with pseudo-labeled data was added between the pretraining and downstream tasks. In this case, ClusterFit could generate more general and transferable representations.

3) Spatial Context Structure-Based SSL: Spatial context refers to the spatial relationships among various objects in an image. For spatial context structure-based SSL, the models need to not only recognize spatial relationships but also understand the information contained in the objects. A typical example was to predict relative positions without any extra context [30], where a deep network was trained to predict the relative position between two image patches. Noroozi and Favaro [34] proposed a much complex task called "jigsaw puzzles" to further mine the visual features of an image. They divided a windowed region randomly selected from the image into 3×3 patches and trained a CNN to identify the correct permutations of jigsaw puzzle. Gidaris et al. [38] proposed a simple but effective task, where the image was rotated for four different angles, and then, a CNN was trained to predict the rotated degrees (as shown in Fig. 8). The results showed that the proposed method made a great improvement over previous unsupervised representation learning methods.

C. Contrastive SSL

CSSL is proposed by merging CL into SSL, which builds representations by learning to encode what makes two objects similar or different. In this way, both common and distinguish features between images can be learned. CSSL has been successfully applied for various CV tasks, including image classification, object detection, and image segmentation. Different from context prediction-based methods, CSSL methods focus on contrastive losses among the images in the feature space. According to the literature, CSSL methods in CV can be summarized into four categories, namely, instance discrimination-based CSSL, clustering-based CSSL, self-distillation-based CSSL, feature decorrelation-based CSSL, and ViT-based CSSL.

1) Instance Discrimination-Based CSSL: Instance discrimination is based on the principle that semantic information in different views of the same image is consistent. The goal of instance discrimination-based CSSL methods is to train models to discriminate the views from different images. Wu et al. [39] proposed the first instance discrimination-based CL method, where each image instance was treated as a single class. A CNN was trained to encode each image instance to a feature vector and then distinguish different instances in the feature space. The features of an image instance were saved as positive samples, while the features of all other remaining images were saved as negative samples. To get

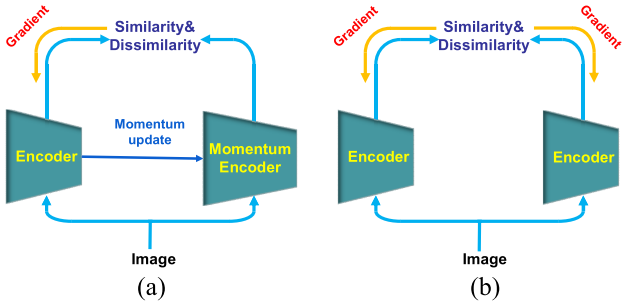


Fig. 9. Contrastive loss mechanisms of (a) MoCo v1 and (b) SimCLR v1.

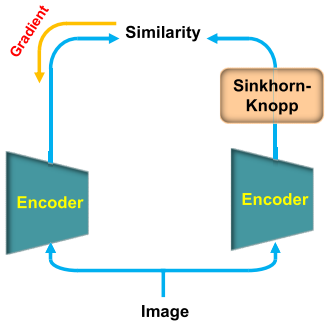


Fig. 10. Contrastive loss mechanism of SwAV.

enough negative samples without additional burden, a memory bank was designed to save the features of each image instance. To reduce the computational complexity, NCE was introduced to compute the similarity between all instances [50]. Hjelm et al. [51] proposed a CL method called DIM, which maximized mutual information to discriminate different instances. First, a CNN was trained to encode an image to $M \times M$ matrix, which then was rearranged to a single feature vector representing the whole image. Finally, good representations are learned by maximizing mutual information between local features and global features of the same image. Based on their work, Bachman et al. [52] extended the local DIM to augmented multiscale DIM. Tian et al. [53] presented a novel CL framework, which was also realized by maximizing mutual information between features of different views of the same image.

The MoCo SSL is one important class of instance discrimination-based CSSL methods [54], [55]. The MoCo v1 model was proposed by He et al. [54], where the instance discrimination was used as the pretext task and a novel MoCo mechanism was designed to replace the memory bank. Generally speaking, the MoCo v1 included two key operations, namely, dictionary and momentum updating. The dictionary maintained a queue of features. During each iteration, the dictionary was updated by adding new encoded features and removing the oldest features. At the same time, the momentum encoder was slowly updated. These two operations enabled to overcome the problem of inconsistency existed in the memory bank.

Another famous CL framework for visual representations is called SimCLR [40], [56]. Chen et al. [40] proposed the SimCLR v1 model, which required no specialized architectures or memory bank. A simple stochastic augmentation module was designed to augment unlabeled image instances. Another important module was an MLP projection head, which could map encoded features to the space of contrastive loss. The contrastive loss mechanisms of MoCo v1 and SimCLR v1 are shown in Fig. 9.

2) Clustering-Based CSSL: Clustering is one class of unsupervised machine learning methods, which can assign groups with similar traits into the same class. Compared with instance discrimination-based CSSL methods, clustering-based CSSL methods can enhance the consistency of clustering different views of the same image. For example, Caron et al. [57] proposed a clustering-based CL method called SwAV, which introduced the multicrop strategy to increase the number of samples without additional burden. The contrastive loss mechanism of SwAV is shown in Fig. 10. Zhuang et al. [58] proposed a local aggregation method, where a CNN was as the embedding function. The CNN is trained to maximize a metric of local aggregation, causing similar image instances to move together in the embedding space, while allowing dissimilar instances to separate. PCL was another novel clustering-based CSSL method [59], which introduced clustering to assign each instance to multiple prototypes with different granularities. The PCL model was trained by minimizing the ProtoNCE loss function.

3) Self-Distillation-Based CSSL: Knowledge distillation is the process of transferring the knowledge from a teacher model to a student model without loss of validity [60]. When both models have the same architecture, this process is called self-distillation. Thus, self-distillation can be considered as knowledge transfer in the same model. Compared with conventional knowledge distillation, self-distillation can achieve higher accuracy, acceleration, and compression. Self-distillation-based CSSL methods have asymmetric network architectures, where different views of the same image are fed to two encoders, and then, an encoder with a predictor is trained to predict the outputs. BYOL was the first self-distillation-based CSSL method, which was consisted of an online network and a target network [61]. Both networks included an encoder and a projector. The online network had an additional predictor, which was trained to predict the feature vectors of the target network. Experiments showed that the BYOL method performed much better than the previous CL methods. Chen and He [62] proposed a simpler self-distillation-based CSSL method called SimSiam, which did not require negative sample pairs, large batches, and momentum encoders. Similarly, SimSiam consisted of two networks, and the model was trained by maximizing the similarity of the output vectors of the two networks. The contrastive loss mechanisms of BYOL and SimSiam are shown in Fig. 11.

4) Feature Decorrelation-Based CSSL: For SSL methods, a common and potential issue is the existence of completely collapsed solutions. Feature decorrelation is an effective operation of avoiding collapses without negative samples. In feature decorrelation-based CSSL methods, special loss functions based on decorrelation mechanism are often applied to handle collapses. For example, The Barlow Twins method used a symmetrical Siamese network to encode different views of images

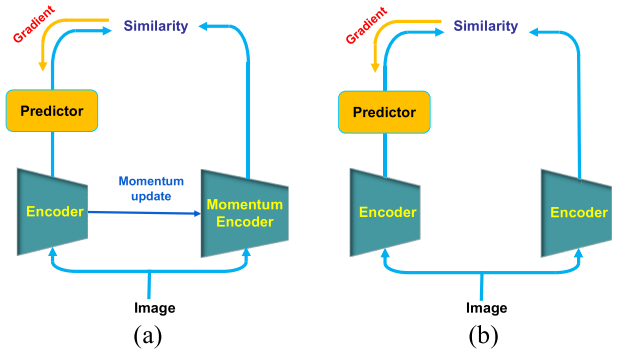


Fig. 11. Contrastive loss mechanisms of (a) BYOL and (b) SimSiam.

into embedding vectors, and the loss function was defined to enforce the cross correlation matrix of the twin embeddings vectors close to the identity matrix [63]. Ermolov et al. [64] defined whitening MSE to constrain the sample representations in a spherical distribution to replace instance contrasting. Similar to the Barlow Twins method, the VICReg method used a Siamese network to obtain embedding vectors [65], which was trained by a joint loss function including three terms: variance, invariance, and covariance regularization.

5) ViT-Based CSSL: Since ViT was proved to greatly improve the performance of encoders [43], many CL methods have utilized the ViT as the backbone network in order to improve their performances in CV tasks. For example, Caron et al. [66] introduced the ViT for the backbone network of a novel self-distillation-based CSSL algorithm called as DINO. Experimental results showed that the DINO model performed better than other CNN-based SSL methods. Based on the DINO model, Li et al. [67] proposed an efficient self-supervised ViT by using a multistage ViT architecture with sparse self-attention mechanism and a region-matching pretraining task.

D. MIM-Based SSL

In recent years, MIM has become popular as a new type of SSL methods in CV. The underlying idea of the MIM method is that a portion of the input image is randomly masked and then reconstructed via the pretext task. Generally speaking, an MIM model consists of a ViT encoder and a ViT decoder, where the encoder is used to encode randomly masked image patches and the decoder is used to reconstruct masked image patches. After pretraining, the encoder will be applied for downstream tasks. The MIM methods have achieved great success in the field of NLP [68], which motivates many researchers to introduce the MIM method into the field of CV. However, it has to face new challenges. He et al. [41] and Xie et al. [69] analyzed and summarized these challenges. First, mask tokens and positional embedding vectors are difficult to be achieved in CNNs, so CNNs are not suitable for MIM tasks. Second, there are significant differences between the visual and language signals. Language signals are generated by humans, which have dense information, but visual signals are sampled by cameras, which have a lot of redundant information. Moreover, images are often locally related. By combining with the ViT model, MIM methods have made a breakthrough in dealing with these challenges [41].

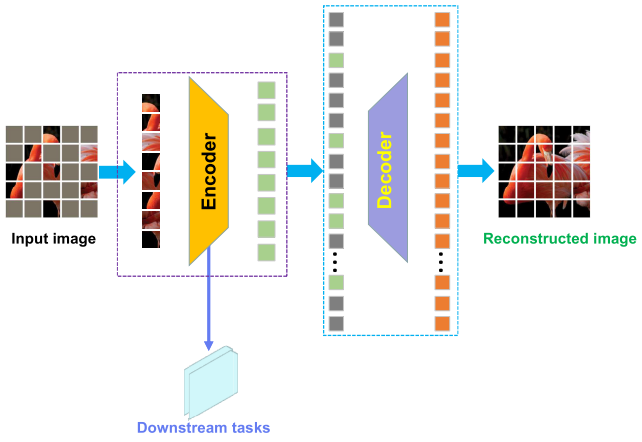


Fig. 12. Architecture of MAE.

In the field of CV, BEIT was the first MIM method [70], where the input images were tokenized to discrete visual tokens by a pretrained discrete VAE, and then, a ViT backbone network was trained to predict the randomly masked visual tokens. Another common MIM method was MAE, which had an asymmetric encoder–decoder architecture [41], which is shown in Fig. 12. The encoder was fed with unmasked patches, while the lightweight decoder was fed with both the encoded patches and the masked tokens to rebuild the original image. Experiments showed that the MAE model performed faster and more accurately in training large models. Chen et al. [71] proposed the CAE, where the functions of the encoder and decoder were further separated. Specifically, a latent contextual regressor with an alignment constraint was designed to make sure that the encoder was only responsible for representation learning and the decoder was only responsible for solving the pretext task.

There are some other methods combining MIM with CL, which also achieve good performances in CV tasks. For example, IBOT used an online tokenizer to perform MIM and then learned to distill knowledge from the tokenizer [72]. MSN was an SSL framework for learning image representations [73], image augmentation was used to generate two views of an image, and MSN randomly masked one view while leaving the other view unchanged. Finally, a ViT was trained to output similar embeddings for the two views. To make up for the shortcomings of instance discrimination and MIM, the SiameseIM method was proposed [74], which was based on a Siamese network with two branches. The online branch encoded the first view and predicted the second view's representation. Then, the target branch produced the object image by encoding the second view.

Despite the outstanding advantages, each class of SSL methods still has its limitations or disadvantages in real-world applications. In the end, the main disadvantages of the above four classes of SSL methods are summarized in Table II.

IV. Applications of SSL Methods in CV

To now, SSL has been widely used in the field of CV. This section will summarize these applications of SSL methods in CV from two different perspectives, namely, task-related and field-related applications.

TABLE II
DISADVANTAGES OF FOUR CLASSES OF SSL METHODS IN CV

Methods	Disadvantages
Generative methods	■ Depend on the quality of the training data ■ Lack of controlling the generated outputs ■ Lack of interpretability ■ Computationally expensive
Context prediction methods	● Lack of general approaches ● Depend on the quality of the training data ● Sensitivity to interrupts
CL methods	■ Limited generalization ■ Sensitivity to hyperparameters ■ Require large amounts of high-quality training data ■ Computational complexity
MIM methods	● Need image preprocessing ● Lack of interpretability ● Difficult to determine the masked ratio ● Rely on the masked regions

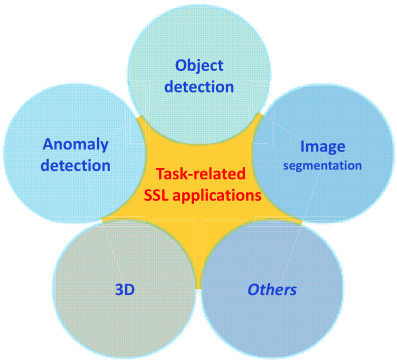


Fig. 13. Task-related classification of SSL applications.

A. Task-Related Classification of SSL Applications

In the field of CV, SSL methods can learn representations by solving different downstream tasks, such as object detection, image segmentation, anomaly detection, and others, which is shown in Fig. 13.

1) SSL Object Detection in CV: Object detection is a basic and important task in the field of CV. Generally speaking, the target detection task consists of classifying and localizing object instances in an image. In order to better apply the SSL method for downstream object detection tasks, researchers successively design new SSL frameworks. Baek et al. [75] proposed a novel pretext task for object detection, where the model was trained to predict the parameters of the geometric transformation. Liu et al. [76] proposed a novel SSL method for object detection, where the earth mover's distance was used to calculate the similarity between two image representations. It was a modified version of BYOL model without the global pooling layers. Dang et al. [77] applied multicrop and box localization pretext tasks for object detection in the SSL framework. Yang et al. [78] proposed a novel object detection-specific pretext task called instance localization. Specifically, two views from different locations and scales of a foreground image instance were copy-pasted onto different background, and then, the model was trained to predict the instance's category based on the composite image and foreground bounding boxes.

2) SSL Image Segmentation in CV: Image segmentation is an important and complex task in the field of CV and has gained wide applications, such as scene understanding, medical image analysis, robotic perception, and autonomous vehicles [6]. The aim of segmentation task is to divide an image into its constituent parts or objects. In traditional way, a large number of labeled datasets are required to train deep segmentation networks. To avoid manually labeling images, researchers investigated specialized SSL algorithms for image segmentation tasks. Larsson et al. [79] proposed the fine-grained segmentation networks, where the segmentation network was trained in a self-supervised manner by using the pixel labels produced by k-means clustering. Hoyer et al. [80] took monocular depth estimation as an SSL pretext task of semantic segmentation, where these estimated depth pseudo-labels were used to generate more training samples. Li et al. [81] proposed a dense cross-image semantic CL framework with semantic decision boundary, which could better learn multigranularity representations. The pretrained model performed excellent in downstream dense prediction tasks, such as semantic image segmentation. Hu et al. [82] proposed region-aware CL for semantic segmentation tasks, which focused on the semantic relationships that existed across the training data. In order to learn representations corresponding to various potential object categories, Ziegler and Asano [83] combined self-supervised ViT with a spatially dense clustering task for unsupervised segmentation. The results testified that the method achieved better performance than ViT in the task of fully unsupervised semantic segmentation.

3) SSL Anomaly Detection in CV: Image anomaly detection aims to detect abnormal samples or events, which do not exist in the available dataset. In nature, SSL methods are much suited for image anomaly detection because they do not need labeled samples. By now, researchers have proposed specific SSL algorithms for image anomaly detection. Li et al. [84] proposed an SSL method to identify and locate anomalies in images, where the superpixel masking and inpainting operations were utilized. Huang et al. [85] proposed an SSL method through random masking and restoring for unsupervised anomaly detection and localization. In this work, a progressive mask refinement approach was proposed to uncover the normal regions and locate the anomalous regions. Li et al. [86] proposed a two-stage SSL framework for building anomaly detectors, where a generative one-class classifier was built on self-supervised learned representations. Long et al. [87] proposed an augmented patch segmentation to train a self-supervised residual de-convolution network. By using this network, the tiny variance between normal and abnormal patterns could be captured at pixel level. In addition, some researchers introduced special abnormal data augmentation strategies into CL methods. For example, Yoa et al. [88] proposed a CL method for anomaly detection with a dynamic local augmentation strategy, which was used to generate negative samples from normal training dataset. Tack et al. [89] presented a CL-based novelty detection method called CSI, which contrasted the image with distributionally shifted augmentations of itself. Cho et al. [90] proposed a task-specific variant of CL for anomaly detection, named masked CL.

4) SSL in 3-D CV: As we all know that SSL methods have achieved great success in 2-D CV tasks in past years. Inspired

by these works, researchers also began to introduce SSL methods for 3-D CV tasks. Different from the representation way characterized by pixels in 2-D images and videos, 3-D image data are usually represented in several different formats, such as depth images, point cloud data, mesh data, volumetric grid data, and so on. In recent years, 3-D sensors have been developed to easily sample 3-D image data, but large amounts of manually labeled 3-D image data are still very scarce. This problem is even more serious than that in the 2-D field. By referring to 2-D SSL algorithms, researchers have built modified pretext tasks for 3-D image data in order to realize 3-D SSL. Zeng et al. [91] gave a comprehensive survey on SSL for point clouds, where the existing SSL methods were classified into four broad categories based on the pretexts' characteristics. For example, Sauder and Sievers [92] proposed an SSL task for DL on raw point cloud data, where the deep network was pretrained to correctly reconstruct point clouds, whose parts were randomly displaced. Poursaeed et al. [93] built 3-D SSL downstream tasks on point clouds with fewer labels, where a point cloud was rotated in many ways to generate label-free dataset for self-supervision. Achituve et al. [94] first studied SSL for domain adaptation on point clouds and a new family of pretext tasks called Deformation reconstruction was proposed. Wang et al. [95] proposed an unsupervised pretraining method called occlusion completion, which could learn representations for point clouds.

In recent years, 3-D CL methods were also adopted to learn representations from point clouds. For example, Sanghi [96] proposed a decoder-free, self-supervised representation learning mechanism by extending the DIM [51] and CL principles on 3-D shapes. Its advantage is that the reconstruction of 3-D shapes was not needed. Janda et al. [97] presented a self-supervised method for extracting visual features from images and then using them as labels to pretrain 3-D models by a contrastive loss. In this way, image and point cloud modalities were combined. Wang et al. [98] proposed a generative variational-CL model, where a variational contrastive module was designed to constrain the feature distribution, rather than feature values corresponding to each image in the latent space. Afham et al. [99] proposed a cross-modal CL approach called CrossPoint to learn transferable 3-D point cloud representations, which could enable a simple 3-D–2-D correspondence of objects in the feature space using CSSL. Sheng et al. [100] combined contrastive predictive coding and reconstruction to build a unified self-supervised framework for dynamic point clouds, where point cloud-based contrastive prediction and reconstruction were designed to collaboratively learn more comprehensive spatiotemporal representations. Li et al. [101] proposed a self-supervised feature learning approach, which directly works on unlabeled point cloud of a complex 3-D scene, where a new CL method called multi-FOV contrasting was applied to predict whether two snapshots are from the same object or not.

B. Field-Related Classification of SSL Applications

DL-based AI has revolutionized the development of many fields of science and engineering. Meanwhile, large amounts of labeled image data are scarce or expensive to obtain in real-world applications. Therefore, SSL is especially valuable in many application scenarios, as shown in Fig. 14, such as industry, agriculture, healthcare, and so on.

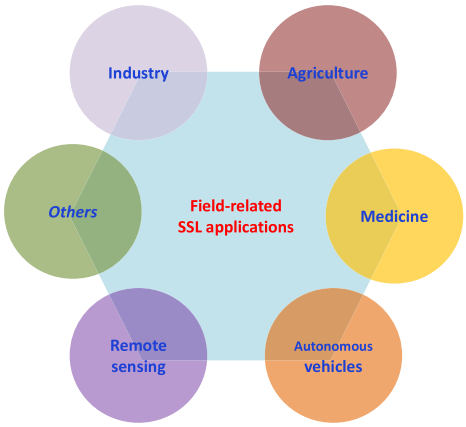


Fig. 14. Field-related classification of SSL applications.

1) SSL in Industrial Applications: With the increasing of automation in manufacturing, DL-based CV technology has brought huge benefits during the industrial production life cycle. In particular, CV-based quality inspection systems can automatically assess product quality by analyzing image data captured by cameras. Compared with traditionally manual inspection methods, it can offer faster speed, higher consistency, and better accuracy with little human intervention. Typical product quality inspections can be classified into preproduction inspection, during-production inspection, and after-production inspection according to the production cycle.

With the rapid development of AM technology, CV-based quality inspection has become one of the most crucial research topics. Yu et al. [102] summarized the application of CV techniques in AM, including material characteristics perception and printing process simulation before printing, real-time monitoring and law exploration during the process, and product evaluation and defect detection after molding. Lui et al. [103] proposed a novel image feature-based SSL model effective quality inspection in AM, where an image fusion approach was used to extract defect-relevant features generating pseudo-labels for SSL. Fernandez-Zelaia et al. [104] proposed an approach for encoding AM layerwise thermal response signatures using self-supervised representation learning.

Defect detection plays a crucial role in the whole production process to ensure product quality. Once undetected bad material or parts gets into the final product, the product quality will decrease sharply. In addition, finished products with defects can potentially lead to irreparable damage to both the economics and the reputation. Defect detection is often hindered by such problems as rare defect data, tiny components, and cluttered background. Xu et al. [105] proposed a high-efficiency defect defector based on SSL strategy with homographic enhancement, so that defective samples with annotations were no longer needed. Zabin et al. [106] proposed a self-supervised representation learning model by using both labeled and unlabeled data. The model was based on a CL framework with an augmentation pipeline and a lightweight

encoder. Min and Li [107] proposed a lightweight two-stage architecture consisted of a railway cropping network and a defect removal VAE. In this way, only normal samples are needed for training to achieve defect detection. Yao et al. [108] presented a defect detection technique for printed circuit boards using the SSL of local image patches.

2) SSL in Agricultural Applications: CV in agriculture is transforming farming into a data-driven, precise, and sustainable industry, but the availability of large annotated datasets is a bottleneck. Nowadays, the potential of SSL methods has being observed in the field of agricultural image analysis. The reason is that it is difficult to generate large amounts of manual labeled images due to unknown objects. SSL methods can effectively deal with this issue, and agriculture-specific SSL algorithms have been widely investigated. Sornapudi and Singh [109] proposed a self-supervised framework for diverse agricultural vision tasks, where the SimCLR was utilized to pretrain the ResNet-50 backbone by using a large of nonannotated real-world agriculture images. The results indicated that the framework was fit for a broad range of downstream agriculture tasks. Güldenring and Nalpantidis [110] validated the potential of self-supervised pretraining in the context of agricultural images, and the CL method was applied to three different agricultural datasets.

Plant diseases or insect pests often bring serious threats to agricultural growth, and it is very necessary to detect and find them as early as possible. Kim et al. [111] proposed a novel CV detection system of plant diseases, where a feature extractor was pretrained on widely used leaf disease datasets in an SSL manner. Yang et al. [112] presented a classification model of tomato diseases called LFC-Net, which composed of a location network, a feedback network, and a classification network. LFC-Net applied the self-supervision mechanism to effectively detect tomato diseases without any manual annotation. Fang et al. [113] proposed a self-supervised clustering framework of plant diseases, where pseudo-labeled training dataset was automatically generated by a clustering algorithm. Monowar et al. [114] proposed a self-supervised clustering system of leaf diseases, which could identify leaf diseases without labeled leaf images. During pretraining, different augmented image pairs were fed to an embedding model to generate embeddings, which were further clustered using the k-means algorithm. The pretrained model was used for leaf disease classification. Kar et al. [12] proposed an SSL framework of pest classification, where raw and segmented insect images were separately fed to the two networks of BYOL [61]. Liu et al. [115] proposed a transformer autoencoder based on latent semantic masking for pest and disease classification.

Apart from identifying pests and diseases, some other SSL applications are also realized in the field of agriculture. For instance, Choi et al. [116] proposed an SSL method for assessing the fruit quality, where randomly permutated RGB channels were used as a new data augmentation strategy. The model learned representations of irregular color patterns by classifying augmented fruit images. Margapuri and Neilsen [117] demonstrated the feasibility of using the SSL framework for seed identification. Three different SSL models were trained and evaluated on five different types of synthetic seed images, including MoCo [54], SimCLR [40], and BYOL [61]. To study plant phenotyping by using CV, Lin et al. [118] proposed a novel self-supervised leaf segmentation framework, where the feature extractor was trained by discriminating whether two local patches from the same image belonged to the same class or not. Based on a novel panoptic pseudo-label generation technique, Siddique et al. [119] proposed an SSL framework for flower image segmentation.

3) SSL in Medical Applications: Medical imaging technologies have become an essential part of modern healthcare, and the workload for radiologists continues to increase. Rapid advancements in DL-based CV provide promising solutions for medical image analysis, including CT, chest and extremity X-rays, MRI, and dermatology images. Different from typical 2-D natural images with three standard color channels, medical images always have different forms, including 2-D grayscale image, 2-D images with four channels, and 3-D volume images. Classical DL methods cannot get rid of labeled images and labeling these medical images is a significant burden for doctors. SSL has achieved remarkable performance in various medical imaging tasks. Huang et al. [120] presented a comprehensive review of SSL strategies for medical image classification. Shurrab and Duwairi [18] reviewed the state-of-the-art researches on SSL applications in the field of medical imaging analysis. According to the literature, the main SSL techniques for medical image analysis can be classified into three categories: generative SSL-based methods, prediction-based SSL methods, and CSSL-based methods.

a) Generative SSL for medicine: Yu and Dai [121] presented a self-supervised multitask learning framework for medical image analysis, which consisted of a discriminative module and a generative module. The generative module learned local fine-grained contextual information by generative adversarial learning. For the interclass imbalance of medical data, Wang et al. [122] proposed an innovative GAN with attention mechanisms to synthesize high-quality fluorescence images and unlabeled fluorescence data were used for downstream classification tasks based on the SSL approach. In order to achieve high-resolution MRI from low-resolution MRI along the longitudinal direction, Xie et al. [123] developed a parallel cycle consistent GAN with a self-supervised strategy. Cheng et al. [124] proposed an SSL cycle-consistent GAN for brain imaging modality transfer, which adopted multibranch inputs for learning the diversity characteristics of multimodal data.

b) Context prediction-based SSL for medicine: Santilli et al. [125] proposed an SSL framework for modeling biochemical signatures of breast cancer from iKnife data. Then, weights were later transferred to the cancer classification task. Spitzer et al. [126] proposed to pretrain a self-supervised Siamese network, which was then applied to predict the 3-D distance between two patches sampled from the same brain. Also, the self-supervised model could implicitly learn to distinguish several cortical brain areas. Bai et al. [127] proposed a novel way to pretrain a cardiac MR image segmentation network, in which features were learned in a self-supervised manner by predicting anatomical positions. Motivated by the inpainting task [35], Chen et al. [128] proposed a novel SSL

strategy for medical images, which used context restoration as a self-supervision task in order to better exploit unlabeled images. The performance of this method was validated on brain MR images, abdominal CT images, and fetal U.S. images.

c) CSSL for medicine: In recent years, CL has been used for self-supervised pretraining CSSL strategy in medical image analysis. The three most used frameworks were SimCLR, MoCo, and BYOL. Chaitanya et al. [129] provided two improved SimCLR method for 3-D medical image segmentation by developing domain-specific and problem-specific knowledge simultaneously. Azizi et al. [130] presented a novel multi-instance CL method for dermatology condition classification, which was built on the same idea of SimCLR with minor modifications. As a variant of the MoCo method, MoCo-CXR was proposed to detect pathologies in chest X-ray images [131]. Vu et al. [132] proposed the MedAug approach as an augmentation strategy when training MoCo framework. MedAug required different views from the same patient. Sriram et al. [133] used MoCo for COVID patients' deterioration prediction tasks, in which non-COVID chest X-ray images from different public datasets were adopted to train MoCo for the subsequent tasks. Feng et al. [134] proposed to integrate SSL into semi-supervised models to enhance medical image recognition, in which the method was pretrained on unlabeled data using the BYOL method.

4) SSL in Autonomous or Unmanned Vehicles: In recent years, autonomous or unmanned vehicles have been well developed, such as self-driving cars and UAV. CV is the core of autonomous vehicle technology, which can enable self-driving vehicles to classify and detect different objects, including pedestrians, vehicles, and road signs. Dong and Cappuccio [135] reviewed publications on CV-based autonomous driving published during the last ten years. However, maintaining a large number of high-quality labeled dataset is especially challenging. Even though the best labeling tools can be utilized, image data can be collected much faster than it can be labeled. Therefore, the applications of SSL in autonomous or unmanned vehicles are very promising, which only apply unlabeled data for representation learning.

a) SSL representation for autonomous driving data: The clustering of autonomous driving scenario data can substantially benefit the autonomous driving. Zhao et al. [136] proposed a comprehensive data clustering framework for a large set of vehicle driving data, where a self-supervised DL approach for spatial and temporal feature extraction to avoid biased data representation. Motion behavior comprehension is very important for autonomous driving. Luo et al. [137] used free supervisory signals from point clouds and paired camera images to estimate motion in a self-supervision manner. Experimental results indicated that the self-supervised approach was superior to the supervised methods. Nunes et al. [138] focused on the problem of representation learning for 3-D point cloud data in the context of autonomous driving and proposed a CL method to learn the structural context of the scene. In order to estimate the traversability of terrain for autonomous driving, in off-road environments, Seo et al. [139] proposed to learn traversability from images without manual labels, so that the model could easily learn traversability in new circumstances.

b) SSL for depth estimation: High-accuracy depth estimation has been an essential task for autonomous driving from the viewpoint of safety, which is the basis of understanding 3-D scene geometry in many tasks. In recent years, self-supervised depth estimation has been studied to achieve great performance improvements. For monocular images or videos, Zhou et al. [140] proposed an SSL framework of estimating monocular depth and camera motion, which could jointly train a single-view depth CNN and a camera pose estimation CNN by using unlabeled video sequences. Fan et al. [141] combined soft attention with hard attention to improve self-supervised monocular depth estimation, in which the hard attention strategy was used to enhance the fusion of multiscale depth predictions. Makarov et al. [142] studied different ways of integrating recurrent blocks and attention mechanisms into a common self-supervised depth estimation pipeline and proposed a modified architecture for monocular depth estimation in a self-supervised manner. In order to increase accuracy and robustness, Yasmina et al. [143] built a self-learned depth prediction network for monocular videos. The model was consisted of a soft attention mechanism and a recurrent block. For stereo images or videos, Tukra and Giannarou [144] proposed contrastive representation learning of left and right views to learn discrete stereo features and then used the trained model to learn disparity by self-supervised training. Guizilini et al. [145] extended monocular self-supervised depth estimation to large-baseline multicamera rigs and built a single network generating dense, consistent, and scale-aware point clouds.

c) SSL for 3-D detection and tracking: For autonomous driving, the main challenge is to understand complex driving environment. To reduce the reliance on labeled data, Shi and Rajkumar [146] investigated SSL for 3-D object detection, in which contrastive and geometric pretext tasks were applied to pretrain the model using unlabeled point cloud data. Kumar et al. [147] proposed a neural network architecture that learns effective object features and their affinities in a self-supervised fashion for multiple-object tracking in 3-D point clouds captured with LiDAR sensors. Xie et al. [148] proposed point cloud MAE, which improved the detection ability of long-distance and occluded objects through SSL.

d) SSL for SLAM: The accurate and robust SLAM systems are crucial for autonomous vehicles to perform missions in unknown environments. Li et al. [149] built a joint feature detection and description network to extract quantized self-supervised local feature for the SLAM to handle environmental interferences in robot localization tasks. Xiu et al. [150] proposed an end-to-end self-supervised monocular visual odometry method based on keypoint heatmap guidance. In this way, the influence of redundant pixels was reduced. Li et al. [151] built a real-time visual SLAM system based on multitask feature extraction CNN and self-supervised feature points. The results showed that the proposed system can achieve high accuracy in a variety of challenging scenes. Xin et al. [152] built an end-to-end network for SLAM preprocessing in an underwater low-light

environment and designed a self-supervised feature point detector and descriptor extraction branch to reduce the reprojection error.

5) SSL in Remote Sensing: Nowadays, CV provides promising analysis tools for RSIs in many real-world tasks, such as forestry monitoring, land surface detecting, disaster prevention, and so on. RSIs are often collected by a satellite, an aircraft, or a drone platform. However, the traditional supervised methods require extensive labeled datasets. It is particularly difficult within the remote sensing domain due to the needs of expert knowledge. Therefore, SSL approaches have been widely applied in the remote sensing domain. Berg et al. [153] reviewed the underlying principles of various self-supervised methods used for scene classification in remote sensing. To now, there are also three mainstream techniques for applying SSL in RSI analysis: generative SSL-based methods, prediction-based SSL methods, and CSSL-based methods.

a) Generative SSL for remote sensing: Xue et al. [154] presented a generative self-supervised feature learning architecture for land cover classification in multimodal RSIs, which could extract high-level feature representations from multiview data without any labeled information. Guo et al. [155] first introduced a gated self-attention module into GANs and proposed a novel scene classification method using self-supervised gated self-attention GANs with the similarity loss. Classical masked image models cannot fully utilize low-level features. To deal with it, Pang et al. [156] proposed a novel masked feature modeling methodology for generative SSL of high-resolution RSIs, which combined both CNN and transformer. The results indicated that the proposed method had better feature extraction capability and higher accuracy on downstream tasks. To solve the data-intensive problem in semantic segmentation, Lu et al. [157] introduced a robust generative SSL model with simple loss for RSIs, which extracted pixelwise representations of RSIs by pretraining on a large number of unlabeled RSIs.

b) Context prediction-based SSL for remote sensing: By using three different pretext tasks of image inpainting, relative position prediction, and instance discrimination, Tao et al. [158] evaluated the feature learning ability of SSL methods for RSI scene classification. Zhao et al. [159] proposed a multitask SSL framework, in which a CNN model was pretrained by predicting the rotation angles of images. Then, the effectiveness of the framework was testified on four commonly used datasets in remote sensing scene classification. Inspired by the colorization task [34], Vincenzi et al. [160] built a novel pretext task for satellite images, where the network was trained to recover the RGB information by using its high-dimensionality spectral bands. Li et al. [161] proposed a novel SSL method for remote sensing semantic segmentation, where three pretext tasks (including image inpainting, transform prediction, and CL) were integrated by a triplet-Siamese network with a multitask loss function.

c) CSSL for remote sensing: Li et al. [162] proposed a GLCNet for semantic segmentation of high-resolution RSIs. In detail, GLCNet consisted of two modules: a global style CL module aiming to learn the image-level representations and a local feature matching CL module aiming to learn the local-region representations. Scheibenreif et al. [163] proposed a novel CSSL method without data augmentation for RSIs, where the model was trained on multimodal data of the same scene obtained with different sensing techniques. This method enabled the model to jointly learn representations from multiple sources, so data fusion could be carried out without supervision. Jain et al. [164] introduced a self-supervised distillation network into the field of remote sensing, called RS-BYOL. In their work, RS-BYOL was trained with multispectral and SAR images, so that the model provided better representation learning.

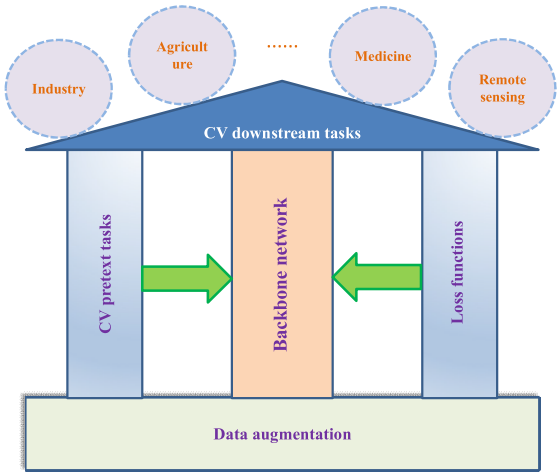


Fig. 15. Key technologies of SSL in CV.

V. Key Technologies of SSL in CV

Technically speaking, SSL involves utilizing pretext tasks to pretrain a model and learn useful feature representations from large amounts of unlabeled image data. Then, the pretrained model is fine-tuned for downstream tasks according to specific requirements, such as image classification, object detection, image segmentation, and so on.

Then, it can be seen that the main components of SSL include lots of unlabeled dataset, pretext tasks, deep networks, and loss functions for pretraining. Therefore, the key technologies of SSL in CV can be mainly summarized as data augmentation, CV pretext task design, backbone network design, and loss function design. The relations among them are shown in Fig. 15. Data augmentation is the basis. Both pretext task design and loss function design are related to backbone network design.

A. Image Data Augmentation

SSL needs a large amount of unlabeled images. However, it is always uneasy and expensive to collect them in real-world applications. The aim of image data augmentation is to artificially create new images from original images by using processing methods, instead of directly collecting image data. Therefore, data augmentation has played a key role in SSL methods. Particularly, during pretraining of CSSL models, different data augmentation methods are selected and applied to generate different views of the same image. By this way, there is consistent semantic meaning within the different

TABLE III
CLASSIFICATION OF COMMON DATA AUGMENTATION METHODS IN CV

Class	Operation	Examples
Geometric transformation	Cropping	MoCo v1 [54], BYOL [61], SimSiam [62], Barlow Twins [63], VICReg [65], Mishra et al. [165], Takahashi et al. [166], etc.
	Flipping	SimCLR v1 [40], BYOL [61], Barlow Twins [63], VICReg [65], Wang et al. [167], etc.
	Rotating	Ishida et al. [168], Chung et al. [169], etc.
	Resizing	SimCLR v1 [40], MoCo v1 [54], BYOL [61], SimSiam [62], Wan et al. [170], Li et al. [171], etc.
Color transformation	Jittering	BYOL [61], SimSiam [62], Puttaruksa and Taeprasartsit [172], He et al. [173], etc.
	Grayscale	Barlow Twins [63], VICREG [65], Wang and Lee [174], etc.
Filtering transformation	Sobel filtering	Khlamov et al. [175], Sandhita and Sonahita [176], etc.
	Gaussian blur	SimSiam [62], Barlow Twins [63], de Vos et al. [177], Sirinukunwattana et al. [178], etc.
DL-based	Feature transfer	Efros and Freeman [179], Gatys et al. [180], etc.
	Adversarial training	Su et al. [181], Goodfellow et al. [182], etc.
	GAN	Karadag and Cicek [183], Motamed et al. [184], etc.

views, so that CSSL can be carried out. To now, common image data augmentation methods can be divided into two categories: traditional and DL-based methods. Furthermore, traditional methods can be divided into three classes. The first class is the geometric transformation of images, such as randomly cropping, flipping, and rotating. The second class is the color transformation of images, such as color jittering (including changing brightness, contrast, and saturation) and converting to grayscale. The third class is the filtering transformation of images, such as Sobel filtering and Gaussian blur. DL-based methods include feature transfer, adversarial training, and GAN. The classification of common data augmentation methods in CV is outlined in Table III, and it can be seen that several methods are always applied in the same work.

B. CV Pretext Task Design

The aim of pretext tasks is to generate labeled data pairs without any human supervision. Thus, pretext tasks play a cornerstone role in the field of SSL. Pretext tasks in CV are often designed according to the downstream tasks. Then, a DNN is trained to solve them, so that it learns visual features from unlabeled data. Finally, the trained DNN can be easily adapted to the downstream tasks. For example, the pretext tasks aim to keep the distance of features in CL methods and the pretext tasks aim to reconstruct the masked images in MIM methods. Nowadays, common pretext tasks in CV mainly include jigsaw puzzle, rotation prediction, and colorization, which are listed in Table IV.

Besides the three tasks, there are some other forms of pretext tasks. For instance, Gidaris et al. [38] proposed to use counting as the pretext task. Lee et al. [188] proposed to sort randomly shuffled frames from a video to learn feature representations.

TABLE IV
COMMON PRETEXT TASKS IN CV

Pretext task	Operation	Examples
Jigsaw puzzle	Reassembling scrambled image patches	Noroozi et al. [34], You and Wang [185], Manna et al. [186], etc.
Rotation prediction	Predicting the degree of image rotation	Gidaris et al. [38], Yamaguchi et al. [187], etc.
Colorization	Predicting missing colors in grayscale images	Zhang et al. [31], Larsson et al. [32], [35], etc.

Jiang et al. [189] used predicting relative depths from a single image as the pretext task. Jenni and Favaro [190] proposed to learn features from spotting artifacts by using a damage and repair strategy. In addition, it is effective to combine several individual pretext tasks for SSL training. For example, Kim et al. [191] proposed a pretext task called completing damaged jigsaw puzzles, which combined colorization, jigsaw puzzles, and missing patch. Chen et al. [192] proposed a pretext task called jigsaw clustering, which combined jigsaw and clustering. Zhu et al. [193] proposed a pretext task called Rubik's cube+, which involved cube ordering, cube rotating, and cube masking. Rubik's cube+ pretrained the 3-D network to learn translation and rotation invariant features from original 3-D medical data.

Designing effective pretext tasks for SSL is an ongoing research direction in CV, which directly affects the performance of solving downstream tasks. In summary, the following main guidelines should be considered when designing pretext tasks in real-world applications: 1) *Task Matching:* any pretext task is always not suitable for all downstream tasks and the pretext task should be designed to well match the target downstream task; 2) *Task Difficulty:* either too easy or too complex pretext tasks may not yield optimal results and pretext tasks with appropriate difficulty should be optimally designed for the target downstream task; 3) *Computation Cost:* different pretext tasks will lead to variable computation cost and pretext tasks should be designed to match the computation power; and 4) *Dataset Size:* for pretraining, larger datasets often need to design more complex pretext tasks in order to improve the learning ability.

C. SSL Backbone Design

DL-based SSL models in CV are typically built on a backbone, which often acts as a pretrained feature extractor. Early SSL backbones depended strongly on ImageNet-based pretrained CNNs. Furthermore, two traditional representatives are AlexNet [194] and ResNet [195]. The AlexNet model is consisted of eight layers, including five convolutional layers and three fully connected layers. The ResNet model adopts the residual block structure to ease the model training of deeper networks. Goyal et al. [196] compared AlexNet and ResNet-50 [32] models pretrained by the two self-supervised approaches (colorization and jigsaw) along three axes: data size, model capacity, and problem complexity. The results indicated that transfer performance increased log-linearly with the data size. Since 2020, SimCLR [40] and MoCo [54] have promoted the popularization of self-supervised backbones and spawned

TABLE V
COMMON BACKBONE NETWORKS OF SSL IN CV

Backbone	Architecture	Examples
CNN	AlexNet	BiGAN [29], Doersch et al. [30], Zhang et al. [31], Pathak et al. [33], Noroozi et al. [34], Larsson et al. [35], Gidaris et al. [38], InstDisc [39], BigBiGAN [46], Caron et al. [47], Hjelm et al. [51], Tian et al. [53], Zhuang et al. [58], etc.
	ResNet	Larsson et al. [35], ClusterFit [49], InstDisc [39], Bachman et al. [52], SimCLR v1 [40], MoCo v1 [54], SwAV [57], Zhuang et al. [58], BYOL [61], SimSiam [62], Barlow Twins [63], VICReg [65], etc.
ViT	ViT	MAE [41], DINO [66], Li et al. [67], BEIT [70], CAE [71], IBOT [72], MSN [73], SiameseIM [74], etc.

much new studies. In recent years, ViT has increasingly become a novel SSL backbone and replace classical CNNs. ViT-based models consist of three modules: linear projection of flattened patches, a transformer encoder, and a MLP head. Due to strong feature extraction ability, ViT has being widely used as backbones in CV. To now, the common backbone networks of SSL in CV are listed in Table V.

D. Loss Function Design

The loss function plays a crucial role in SSL, which is minimized in the feature space to train the model and obtain meaningful representations of the dataset (e.g., images). Common loss functions used in SSL training can be mainly classified into three categories: reconstruction loss, adversarial loss, and joint loss combining both reconstruction and adversarial losses. The reconstruction loss is used to measure the difference between the output of the model and the original input image, which can be calculated by using different functions, such as MAE* [70], MSE [33], and cross entropy [31]. For example, reconstruction loss is widely used in context prediction-based SSL tasks. In MIM-based SSL methods, loss functions are used to measure the difference between the reconstructed and original images in the feature space. Adversarial loss is often adopted in CL or GAN methods, where loss functions are utilized to minimize the distance between positive sample pairs while maximizing the distance between negative sample pairs. Common loss functions used in CL include contrastive loss, triplet loss, N-pair loss, InfoNCE loss, and logistic loss, such as NCE [48], InfoNCE [199], ProtoNCE [57], and NT-Xent [40]. Joint loss is often utilized while simultaneously considering multiple pretext tasks in order to improve the representation performance. Wang et al. [200] applied a joint loss function that combined the binary cross-entropy loss and Dice loss to reduce the influence of noise. Zhang et al. [201] proposed a self-supervised joint learning fault diagnosis method based on three-channel vibration images, in which the joint loss function combines the reconstruction loss and the adversarial loss. In summary, deigning loss functions of SSL methods in CV relies heavily on specific pretext tasks, which is outlined in Table VI.

TABLE VI
COMMON LOSS FUNCTION OF SSL IN CV

Loss function	Calculation	Examples
Reconstruction loss	MAE*	Xie et al. [69], Liu et al. [202], etc.
	MSE	Pathak et al. [33], Noroozi et al. [36], BYOL [61], etc.
	Cross entropy	Zhang et al. [31], CAE [71], etc.
Adversarial loss	NCE	InstDisc [39], SimCLR v1 [40], MoCo v1 [52], PCL [57], Oord et al. [20], etc.
Joint loss	/	Wang et al. [200], Zhang et al. [201], etc.

VI. DISCUSSION ON FUTURE TRENDS

As one class of novel DL methods without using labeled samples, SSL is often called "the future of AI." The SSL technique has been developed for more than ten years and many studies have been done in different CV fields, but it is not fully mature. With the increasing demands on AI, there are still many further efforts to promote the ability of SSL. In our opinions, there are several promising trends as follows in future.

A. Building Large-Scale Datasets With Diversity

With the increasing usage of AI in many domains, large-scale datasets continue to grow and become more available. In this case, SSL can make full use of these data without manual annotation. According to the literature, pretraining on large-scale unlabeled datasets has been proven to improve the downstream task performance on many CV tasks [203], such as object detection and classification. However, there are still two issues deserved to be investigated in future. The first one is how much an SSL pretraining model relies on the number of dataset and how many datasets are really needed for effective SSL. Exploring pretraining data efficiency is crucial for SSL as it can significantly reduce computational costs and guide economical image collection [204]. The other one is the unbalance of training dataset, which is often generated from the same images by data augment methods. The future trend is to build large-scale datasets with diversity for SSL and investigate how to use multimodal or cross-domain dataset to enhance the performance of SSL. For example, SSL4EO-S12 is a large-scale multimodal multitemporal dataset for SSL in earth observation [205].

B. Needing Stronger Model Structure and Representation Learning Ability

Despite many self-supervised pretrained models, most of them focus on a single dominant object and employ different views of the same image. In this case, it is difficult to distinguish two similar scenes containing several objects, such as autonomous driving scenarios. Meanwhile, it is also inefficient to deal with less structured images using the existing pretrained models, such as medical and satellite images. Therefore, more powerful model structures and representation learning methods need to be studied for SSL in future in order to better extract high-level semantic information in images or videos. Similar to large language models in NLP, on the one hand, large vision-language models [206] can be developed and used for SSL. On the other hand, GNN can

realize powerful representation learning on graphs, so GNN models pretrained by SSL on unlabeled datasets enable better transfer performance in downstream tasks [207], deserving to be further studied. In addition, self-attention mechanisms can be explored to enhance the representation ability of image features and the robustness [208].

C. Integrating With Prior Knowledge

In nature, SSL provides a universal framework for different CV applications. However, it is still difficult to learn robust representations in an unsupervised manner without semantics and guarantee of quality. At the same time, similar to other DL methods, the interpretability of SSL is also poor due to the lack of knowledge constraints. In order to improve the learning effectiveness and robustness, the promising solution is to integrate SSL with prior knowledge. Currently, knowledge-guided SSL is becoming a novel direction, and prior knowledge of the target application can significantly affect defining meaningful pretext tasks. For example, domain-expert knowledge can be applied to design more effective SSL pretext tasks and loss functions in the field of medical image processing. More importantly, meaningful representations can be learned from unlabeled dataset by using prior knowledge. Chen et al. [209] proposed a knowledge-based SSL framework to recognize biomedical microscopy images, which provided a better alternative to ImageNet-based pretraining. Yan et al. [210] proposed a novel domain knowledge-guided SSL approach for unsupervised change detection by fusing the domain knowledge of remote sensing indices.

D. Matching the Pretext and Downstream Tasks

In SSL, pretext tasks are designed, so that learned visual features can be well adapted to downstream tasks. Nevertheless, it still remains unclear what kind of pretext tasks perform the best for a given downstream task. It is very challenging to determine the most optimal SSL algorithm because there is no universal solution for common CV applications. The ideal selection of an SSL algorithm should make matching the pretext task and the downstream tasks according to the specific application scene. However, the situation is very complicated in actual applications. In order to deal with this issue, the following studies deserve to be carried out in future: 1) for common CV applications, multiple options for both pretext and downstream tasks can be discussed, so that the effect of different pretext tasks on different downstream tasks can be revealed; 2) for a given application, self-supervised feature learning on multiple pretext tasks can be investigated and compared; and 3) based on a large amount of investigations, a checklist of facilitating users to identify the most suitable pretext tasks under given circumstances should be pursued and summarized for future research. For example, Ding et al. [211] developed and compared three novel SSL tasks relevant to the downstream task of lung adenocarcinoma subtype classification.

E. Exploring Multimodal Representation Learning

Up to now, SSL methods, such as SimCLR [40], MoCo [54], and Simsiam [62], have achieved great success in CV by generating task-agnostic representations from unlabeled dataset. However, these conventional methods often rely on single-modality data, which may have limitations in capturing the diversity of information in complex scenes. To overcome the limitations, it is promising to turn to multimodal data to generate a great deal of complementary information. In recent years, therefore, some researchers have started focusing on multimodal SSL. For example, Chen and Bruzzone [212] proposed the combined utilization of contrastive loss at both image and pixel levels to enhance the suitability of multimodal fusion models for segmentation tasks. Jain et al. [213] explored the applicability of SSL with multimodal satellite imagery for downstream tasks. Fedorov et al. [214] proposed a novel self-supervised framework to extract multiple representations from multimodal neuroimaging data, leading to enhance group inferences.

Zong et al. [215] summarized three major challenges of SSL with multimodal data: 1) learning representations from multimodal data without labels; 2) fusion of different modalities; and 3) learning with unaligned data. In future, there are two feasible approaches to realize multimodal SSL. The first one is multimodal pretraining using a fusion framework. The other one is to integrate independently pretrained unimodal models.

VII. Conclusion

Due to lack of labeled datasets in many real-world applications, SSL methods have become a research hot spot in the field of CV. Just during the past several years, many remarkable breakthroughs have been achieved. This article presented a comprehensive review of SSL methods in CV and mainstream SSL methods were categorized. Then, task-related and field-related applications of SSL methods were summarized, respectively. Finally, key technologies and future trends were addressed. As a review paper, we hope that this article would provide a quick guide for future newcomers into this field.

References

[1] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 6, p. e1264, Nov. 2018, doi: 10.1002/widm.1264.

[2] W. Wang and Y. Yang, "Development of convolutional neural network and its application in image classification: A survey," *Opt. Eng.*, vol. 58, no. 4, p. 40901, Apr. 2019, doi: 10.1117/1.oe.58.4.040901.

[3] F. Shao et al., "Deep learning for weakly-supervised object detection and localization: A survey," *Neurocomputing*, vol. 496, pp. 192–207, Jul. 2022, doi: 10.1016/j.neucom.2022.01.095.

[4] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/JPROC.2023.3238524.

[5] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–35, Jul. 2020, doi: 10.1145/3329784.

[6] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, doi: 10.1109/TPAMI.2021.3059968.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.

[9] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[10] T. Li et al., "Small samples noise prediction of train electric traction system fan based on a multiple regression-fuzzy neural network," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106781, doi: 10.1016/j.engappai.2023.106781.

[11] T. Li et al., "A bearing fault diagnosis method under small sample conditions based on the fractional order Siamese deep residual shrinkage network," *Fractal Fractional*, vol. 8, no. 3, p. 134, Feb. 2024, doi: 10.3390/fractalfract8030134.

[12] S. Kar et al., "Self-supervised learning improves agricultural pest classification," in *Proc. 36th AAAI Conf. Artif. Intell. (AAAI)*, Vancouver, BC, Canada, Feb. 2022, pp. 1–5.

[13] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 392–408.

[14] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021, doi: 10.1109/TPAMI.2020.2992393.

[15] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023, doi: 10.1109/TKDE.2021.3090866.

[16] S. Albelwi, "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy*, vol. 24, no. 4, p. 551, Apr. 2022, doi: 10.3390/e24040551.

[17] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022, doi: 10.1109/MGRS.2022.3198244.

[18] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Comput. Sci.*, vol. 8, p. e1045, Jul. 2022, doi: 10.7717/peerj-cs.1045.

[19] H. Hojjati, T. K. K. Ho, and N. Armanfard, "Self-supervised anomaly detection in computer vision and beyond: A survey and outlook," *Neural Netw.*, vol. 172, Apr. 2024, Art. no. 106106, doi: 10.1016/j.neunet.2024.106106.

[20] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1747–1756.

[21] A. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Dec. 2016, pp. 1–9.

[22] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," 2014, *arXiv:1410.8516*.

[23] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," 2016, *arXiv:1605.08803*.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[25] A. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–10.

[26] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[27] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*.

[28] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," May 2016, *arXiv:1511.06434v2*.

[29] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," May 2016, *arXiv:1605.09782*.

[30] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[31] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.

[32] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 577–593.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[34] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 69–84.

[35] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6874–6883.

[36] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5899–5907.

[37] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6024–6033.

[38] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.

[39] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 1597–1607.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[42] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–11.

[43] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[44] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 1–10.

[45] A. Razavi, A. Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 1–11.

[46] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 1–11.

[47] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 132–149.

[48] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin, "Unsupervised pre-training of image features on non-curated data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2959–2968.

[49] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "ClusterFit: Improving generalization of visual representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6508–6517.

[50] MichaelU. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 9, May 2010, pp. 297–304.

[51] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Represent.*, May 2019, pp. 1–24.

[52] P. Bachman, R. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 1–11.

[53] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[54] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[55] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[56] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 22243–22255.

[57] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 9912–9924.

[58] C. Zhuang, A. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6002–6012.

[59] J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi, "Prototypical contrastive learning of unsupervised representations," 2020, *arXiv:2005.04966*.

[60] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[61] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[62] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[63] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 12310–12320.

[64] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 3015–3024.

[65] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2022, pp. 1–24.

[66] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[67] C. Li et al., "Efficient self-supervised vision transformers for representation learning," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2022, pp. 1–27.

[68] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[69] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[70] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[71] X. Chen et al., "Context autoencoder for self-supervised representation learning," *Int. J. Comput. Vis.*, vol. 132, no. 1, pp. 208–223, Aug. 2023, doi: 10.1007/s11263-023-01852-4.

[72] J. Zhou et al., "IBOT: Image BERT pre-training with online tokenizer," 2021, *arXiv:2111.07832*.

[73] M. Assran et al., "Masked Siamese networks for label-efficient learning," in *Proc. 17th Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 456–473.

[74] C. Tao et al., "Siamese image modeling for self-supervised vision representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2132–2141.

[75] K. Baek, M. Lee, and H. Shim, "PsyNet: Self-supervised approach to object localization using point symmetric transformation," in *Proc. AAAI Conf. Artif. Intell., Feb.*, 2020, vol. 34, no. 7, pp. 10451–10459.

[76] S. Liu, Z. Li, and J. Sun, "Self-EMD: Self-supervised object detection without ImageNet," 2020, *arXiv:2011.13677*.

[77] T. Dang, S. Kornblith, H. T. Nguyen, P. Chin, and M. Khademi, "A study on self-supervised object detection pretraining," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 86–99.

[78] C. Yang, Z. Wu, B. Zhou, and S. Lin, "Instance localization for self-supervised detection pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3987–3996.

[79] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 31–41.

[80] L. Hoyer, D. Dai, Y. Chen, A. Köring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11125–11135.

[81] X. Li et al., "Dense semantic contrast for self-supervised visual representation learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1368–1376.

[82] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 16291–16301.

[83] A. Ziegler and Y. M. Asano, "Self-supervised learning of object parts for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14502–14511.

[84] Z. Li et al., "Superpixel masking and inpainting for self-supervised anomaly detection," in *Proc. BMVC*, 2020, pp. 1–12.

[85] C. Huang, Q. Xu, Y. Wang, Y. Wang, and Y. Zhang, "Self-supervised masking for unsupervised anomaly detection and localization," *IEEE Trans. Multimedia*, vol. 25, pp. 4426–4438, 2022, doi: 10.1109/TMM.2022.3175611.

[86] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9664–9674.

[87] J. Long, Y. Yang, L. Hua, and Y. Ou, "Self-supervised augmented patches segmentation for anomaly detection," in *Proc. Asian Conf. Comput. Vis.*, Dec. 2022, pp. 1926–1941.

[88] S. Yoa, S. Lee, C. Kim, and H. J. Kim, "Self-supervised learning for anomaly detection with dynamic local augmentation," *IEEE Access*, vol. 9, pp. 147201–147211, 2021, doi: 10.1109/ACCESS.2021.3124525.

[89] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty detection via contrastive learning on distributionally shifted instances," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 11839–11852.

[90] H. Cho, J. Seol, and S.-G. Lee, "Masked contrastive learning for anomaly detection," 2021, *arXiv:2105.08793*.

[91] C. Zeng, W. Wang, A. Nguyen, J. Xiao, and Y. Yue, "Self-supervised learning for point cloud data: A survey," *Exp. Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121354, doi: 10.1016/j.eswa.2023.121354.

[92] J. Sauder and B. Sievers, "Self-supervised deep learning on point clouds by reconstructing space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 1–11.

[93] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, and V. G. Kim, "Self-supervised learning of point clouds via orientation estimation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Nov. 2020, pp. 1018–1028.

[94] I. Achituve, H. Maron, and G. Chechik, "Self-supervised learning for domain adaptation on point clouds," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 123–133.

[95] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner, "Unsupervised point cloud pre-training via occlusion completion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9782–9792.

[96] A. Sanghi, "Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 626–642.

[97] A. Janda, B. Wagstaff, E. G. Ng, and J. Kelly, "Contrastive learning for self-supervised pre-training of point cloud segmentation networks with image data," in *Proc. Conf. Robots Vis.*, 2023, pp. 145–152.

[98] B. Wang, Z. Tian, A. Ye, F. Wen, S. Du, and Y. Gao, "Generative variational-contrastive learning for self-supervised point cloud representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6154–6166, Sep. 2024, doi: 10.1109/TPAMI.2024.3378708.

[99] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "CrossPoint: Self-supervised cross-modal contrastive learning for 3D point cloud understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9892–9902.

[100] X. Sheng, Z. Q. Shen, and G. Xiao, "Contrastive predictive autoencoders for dynamic point cloud self-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 8, pp. 9802–9810.

[101] X. Li, L. Zhang, and Z. Zhu, "SnapshotNet: Self-supervised feature learning for point cloud data segmentation using minimal labeled data," *Comput. Vis. Image Understand.*, vol. 216, Feb. 2022, Art. no. 103339, doi: 10.1016/j.cviu.2021.103339.

[102] H. Yu, W. Li, D. Li, L. Wang, and Y. Wang, "Enhancing additive manufacturing with computer vision: A comprehensive review," *Int. J. Adv. Manuf. Technol.*, vol. 132, pp. 5211–5229, May 2024, doi: 10.1007/s00170-024-13689-3.

[103] C. F. Lui, A. Maged, and M. Xie, "A novel image feature based self-supervised learning model for effective quality inspection in additive manufacturing," *J. Intell. Manuf.*, vol. 34, no. 7, pp. 1–16, Oct. 2023, doi: 10.1007/s10845-023-02232-y.

[104] P. Fernandez-Zelaia, S. N. Dryepondt, A. K. Ziabari, and M. M. Kirka, "Self-supervised learning of spatiotemporal thermal signatures in additive manufacturing using reduced order physics models and transformers," *Comput. Mater. Sci.*, vol. 232, Jan. 2024, Art. no. 112603, doi: 10.1016/j.commatsci.2023.112603.

[105] R. Xu, R. Hao, and B. Huang, "Efficient surface defect detection using self-supervised learning strategy and segmentation network," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101566, doi: 10.1016/j.aei.2022.101566.

[106] M. Zabin, A. N. B. Kabir, M. K. Kabir, H.-J. Choi, and J. Uddin, "Contrastive self-supervised representation learning framework for metal surface defect detection," *J. Big Data*, vol. 10, no. 1, p. 145, Sep. 2023, doi: 10.1186/s40537-023-00827-z.

[107] Y. Min and Y. Li, "Self-supervised railway surface defect detection with defect removal variational autoencoders," *Energies*, vol. 15, no. 10, p. 3592, May 2022, doi: 10.3390/en15103592.

[108] N. Yao, Y. Zhao, S. G. Kong, and Y. Guo, "PCB defect detection with self-supervised learning of local image patches," *Measurement*, vol. 222, Nov. 2023, Art. no. 113611, doi: 10.1016/j.measurement.2023.113611.

[109] S. Sornapudi and R. Singh, "Self-supervised backbone framework for diverse agricultural vision tasks," 2024, *arXiv:2403.15248*.

[110] R. Güldenring and L. Nalpantidis, "Self-supervised contrastive learning on agricultural images," *Comput. Electron. Agricult.*, vol. 191, Dec. 2021, Art. no. 106510, doi: 10.1016/j.compag.2021.106510.

[111] T. Kim, H. Kim, K. Baik, and Y. Choi, "Instance-aware plant disease detection by utilizing saliency map and self-supervised pre-training," *Agriculture*, vol. 12, no. 8, p. 1084, Jul. 2022, doi: 10.3390/agriculture12081084.

[112] G. Yang, G. Chen, Y. He, Z. Yan, Y. Guo, and J. Ding, "Self-supervised collaborative multi-network for fine-grained visual categorization of tomato diseases," *IEEE Access*, vol. 8, pp. 211912–211923, 2020, doi: 10.1109/ACCESS.2020.3039345.

[113] U. Fang, J. Li, X. Lu, L. Gao, M. Ali, and Y. Xiang, "Self-supervised cross-iterative clustering for unlabeled plant disease images," *Neurocomputing*, vol. 456, pp. 36–48, Oct. 2021, doi: 10.1016/j.neucom.2021.05.066.

[114] M. M. Monowar, M. A. Hamid, F. A. Kateb, A. Q. Ohi, and M. F. Mridha, "Self-supervised clustering for leaf disease identification," *Agriculture*, vol. 12, no. 6, p. 814, Jun. 2022, doi: 10.3390/agriculture12060814.

[115] H. Liu, Y. Zhan, H. Xia, Q. Mao, and Y. Tan, "Self-supervised transformer-based pre-training method using latent semantic masking auto-encoder for pest and disease classification," *Comput. Electron. Agricult.*, vol. 203, Dec. 2022, Art. no. 107448, doi: 10.1016/j.compag.2022.107448.

[116] T. Choi, O. Would, A. Salazar-Gomez, and G. Cielniak, "Self-supervised representation learning for reliable robotic monitoring of fruit anomalies," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 2266–2272.

[117] V. Margapuri and M. Neilsen, "Classification of seeds using domain randomization on self-supervised learning frameworks," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 01–08.

[118] X. Lin et al., "Self-supervised leaf segmentation under complex lighting conditions," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109021, doi: 10.1016/j.patcog.2022.109021.

[119] A. Siddique, A. Tabb, and H. Medeiros, "Self-supervised learning for panoptic segmentation of multiple fruit flower species," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 12387–12394, Oct. 2022, doi: 10.1109/LRA.2022.3217000.

[120] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: A systematic review and implementation guidelines," *Npj Digit. Med.*, vol. 6, no. 1, p. 74, Apr. 2023, doi: 10.1038/s41746-023-00811-0.

[121] H. Yu and Q. Dai, "Self-supervised multi-task learning for medical image analysis," *Pattern Recognit.*, vol. 150, Jun. 2024, Art. no. 110327, doi: 10.1016/j.patcog.2024.110327.

[122] Z. Wang, Q. Zhang, Y. Wang, M. Zhu, and Q. Li, "A framework for immunofluorescence image augmentation and classification based on unsupervised attention mechanism," *J. Biophotonics*, vol. 16, no. 12, Dec. 2023, Art. no. e202300209, doi: 10.1002/jbio.202300209.

[123] H. Xie et al., "Synthesizing high-resolution magnetic resonance imaging using parallel cycle-consistent generative adversarial networks for fast magnetic resonance imaging," *Med. Phys.*, vol. 49, no. 1, pp. 357–369, Jan. 2022, doi: 10.1002/mp.15380.

[124] D. Cheng et al., "Self-supervised learning for modal transfer of brain imaging," *Frontiers Neurosci.*, vol. 16, Sep. 2022, Art. no. 920981, doi: 10.3389/fnins.2022.920981.

[125] A. M. L. Santilli et al., "Domain adaptation and self-supervised learning for surgical margin detection," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 5, pp. 861–869, May 2021, doi: 10.1007/s11548-021-02381-6.

[126] H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, and T. Dickscheid, "Improving cytoarchitectonic segmentation of human brain areas with self-supervised Siamese networks," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, Sep. 2018, pp. 663–671.

[127] W. Bai et al., "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2019, pp. 541–549.

[128] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539, doi: 10.1016/j.media.2019.101539.

[129] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 12546–12558.

[130] S. Azizi et al., "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3478–3488.

[131] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models," in *Proc. Med. Imag. Deep Learn.*, 2021, pp. 728–744.

[132] Y. N. T. Vu, R. Wang, N. Balachandar, C. Liu, A. Y. Ng, and P. Rajpurkar, "MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation," in *Proc. Mach. Learn. Healthcare Conf.*, 2021, pp. 755–769.

[133] A. Sriram et al., "COVID-19 prognosis via self-supervised representation learning and multi-image prediction," 2021, *arXiv:2101.04909*.

[134] H. Feng, Y. Z. Jia, R. J. Xu, M. Prasad, A. Anaissi, and A. Braytee, "Integration of self-supervised BYOL in semi-supervised medical image recognition," in *Proc. Int. Conf. Comput. Sci.*, 2024, pp. 163–170.

[135] X. Dong and M. L. Cappuccio, "Applications of computer vision in autonomous vehicles: Methods, challenges and future directions," 2023, *arXiv:2311.09093*.

[136] J. Zhao, J. Fang, Z. Ye, and L. Zhang, "Large scale autonomous driving scenarios clustering with self-supervised feature extraction," in *Proc. IEEE Intell. Vehicles Symp.*, Jul. 2021, pp. 473–480.

[137] C. Luo, X. Yang, and A. Yuille, "Self-supervised pillar motion learning for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3182–3191.

[138] L. Nunes, R. Marcuzzi, X. Chen, J. Behley, and C. Stachniss, "SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2116–2123, Apr. 2022, doi: 10.1109/LRA.2022.3142440.

[139] J. Seo, S. Sim, and I. Shim, "Learning off-road terrain traversability with self-supervisions only," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4617–4624, Aug. 2023, doi: 10.1109/LRA.2023.3284356.

[140] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[141] C. Fan, Z. Yin, F. Xu, A. Chai, and F. Zhang, "Joint soft–hard attention for self-supervised monocular depth estimation," *Sensors*, vol. 21, no. 21, p. 6956, Oct. 2021, doi: 10.3390/s21216956.

[142] I. Makarov, M. Bakhanova, S. Nikolenko, and O. Gerasimova, "Self-supervised recurrent depth estimation with attention mechanisms," *PeerJ Comput. Sci.*, vol. 8, p. e865, Jan. 2022, doi: 10.7717/peerj-cs.865.

[143] Y. Benkhoui, T. El-Korchi, and R. Ludwig, "An attention-based self-supervised approach to monocular depth estimation from UAV captured video sequences," in *Proc. IEEE 6th Int. Conf. Pattern Recognit. Artif. Intell. (PRAI)*, Aug. 2023, pp. 622–629.

[144] S. Tukra and S. Giannarou, "Stereo depth estimation via self-supervised contrastive representation learning," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, 2022, pp. 604–614.

[145] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon, "Full surround monodepth from multiple cameras," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5397–5404, Apr. 2022, doi: 10.1109/LRA.2022.3150884.

[146] W. Shi and R. R. Rajkumar, "Self-supervised pretraining for point cloud object detection in autonomous driving," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 4341–4348.

[147] A. Kumar, J. Kini, A. Mian, and M. Shah, "Self-supervised learning for multiple object tracking in 3D point clouds," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2022, pp. 3754–3761.

[148] G. Xie, Y. Li, H. Qu, and Z. Sun, "Masked autoencoder for pre-training on 3D point cloud object detection," *Mathematics*, vol. 10, no. 19, p. 3549, Sep. 2022, doi: 10.3390/math10193549.

[149] S. Li, S. Liu, Q. Zhao, and Q. Xia, "Quantized self-supervised local feature for real-time robot indirect VSLAM," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 3, pp. 1414–1424, Jun. 2022, doi: 10.1109/TMECH.2021.3085326.

[150] H. Xiu, Y. Liang, and H. Zeng, "Keypoint heatmap guided self-supervised monocular visual odometry," *J. Intell. Robotic Syst.*, vol. 105, no. 4, p. 78, Jul. 2022, doi: 10.1007/s10846-022-01685-2.

[151] G. Li, L. Yu, and S. Fei, "A deep-learning real-time visual SLAM system based on multi-task feature extraction network and self-supervised feature points," *Measurement*, vol. 168, Jan. 2021, Art. no. 108403, doi: 10.1016/j.measurement.2020.108403.

[152] Z. Xin, Z. Wang, Z. Yu, and B. Zheng, "ULL-SLAM: Underwater low-light enhancement for the front-end of visual SLAM," *Frontiers Mar. Sci.*, vol. 10, May 2023, Art. no. 1133881, doi: 10.3389/fmars.2023.1133881.

[153] P. Berg, M.-T. Pham, and N. Courty, "Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives," *Remote Sens.*, vol. 14, no. 16, p. 3995, Aug. 2022, doi: 10.3390/rs14163995.

[154] Z. Xue, X. Yu, A. Yu, B. Liu, P. Zhang, and S. Wu, "Self-supervised feature learning for multimodal remote sensing image land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533815, doi: 10.1109/TGRS.2022.3190466.

[155] D. Guo, Y. Xia, and X. Luo, "Self-supervised GANs with similarity loss for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2508–2521, 2021, doi: 10.1109/JSTARS.2021.3056883.

[156] S. Pang, H. Hu, Z. Zuo, J. Chen, and X. Hu, "Masked feature modeling for generative self-supervised representation learning of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8434–8449, 2024, doi: 10.1109/JSTARS.2024.3385420.

[157] J. Lu, G. He, H. Dou, Q. Gao, L. Fang, and Y. Deng, "Score-Seg: Leveraging score-based generative model for self-supervised semantic segmentation of remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8818–8833, 2023, doi: 10.1109/JSTARS.2023.3314866.

[158] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2020.3038420.

[159] Z. Zhao, Z. Luo, J. Li, C. Chen, and Y. Piao, "When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework," *Remote Sens.*, vol. 12, no. 20, p. 3276, Oct. 2020, doi: 10.3390/rs12203276.

[160] S. Vincenzi et al., "The color out of space: Learning self-supervised representations for Earth observation imagery," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3034–3041.

[161] W. Li, H. Chen, and Z. Shi, "Semantic segmentation of remote sensing images with self-supervised multitask representation learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6438–6450, 2021, doi: 10.1109/JSTARS.2021.3090418.

[162] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014, doi: 10.1109/TGRS.2022.3147513.

[163] L. Scheibenreif, M. Mommert, and D. Borth, "Contrastive self-supervised data fusion for satellite imagery," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 705–711, May 2022, doi: 10.5194/isprs-annals-v-3-2022-705-2022.

[164] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7797–7808, 2022, doi: 10.1109/JSTARS.2022.3204888.

[165] S. Mishra et al., "Object-aware cropping for self-supervised learning," 2021, *arXiv:2112.00319*.

[166] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep CNNs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2917–2931, Sep. 2020, doi: 10.1109/TCSVT.2019.2935128.

[167] S. Wang, R. Yao, Y. Zhang, Q. Jiang, and C. Zhang, "Data augmentation of random grid-hiding for video object segmentation," *Multimedia Tools Appl.*, vol. 78, no. 16, pp. 23029–23048, Apr. 2019, doi: 10.1007/s11042-019-7569-5.

[168] N. Ishida, Y. Nagatsu, and H. Hashimoto, "Unsupervised anomaly detection based on data augmentation and mixing," in *Proc. IECON 46th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2020, pp. 529–533.

[169] Q. M. Chung, T. D. Le, T. V. Dang, N. D. Vo, T. V. Nguyen, and K. Nguyen, "Data augmentation analysis in vehicle detection from aerial videos," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Oct. 2020, pp. 1–3.

[170] D. Wan, R. Lu, T. Xu, S. Shen, X. Lang, and Z. Ren, "Random interpolation resize: A free image data augmentation method for object detection in industry," *Exp. Syst. Appl.*, vol. 228, Oct. 2023, Art. no. 120355, doi: 10.1016/j.eswa.2023.120355.

[171] C. Li, H. Zhang, B. Yang, and J. Wang, "Image classification adversarial attack with improved resizing transformation and ensemble models," *PeerJ Comput. Sci.*, vol. 9, p. e1475, Jul. 2023, doi: 10.7717/peerj-cs.1475.

[172] C. Puttaruksa and P. Taeprasartsit, "Color data augmentation through learning color-mapping parameters between cameras," in *Proc. 15th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2018, pp. 1–6.

[173] Z. He et al., "Deconv-transformer (DecT): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture," *Inf. Sci.*, vol. 608, pp. 1093–1112, Aug. 2022, doi: 10.1016/j.ins.2022.06.091.

[174] J. Wang and S. Lee, "Data augmentation methods applying grayscale images for convolutional neural networks in machine vision," *Appl. Sci.*, vol. 11, no. 15, p. 6721, Jul. 2021, doi: 10.3390/app11156721.

[175] S. Khlamov, I. Tabakova, and T. Trunova, "Recognition of the astronomical images using the Sobel filter," in *Proc. Int. Conf. Syst., Signals Image Process.*, Jun. 2022, pp. 1–4.

[176] S. Agarwal and S. Agarwal, "Bone age assessment from lateral cephalograms using deep learning algorithms in the Indian population," *Indian J. Dental Res.*, vol. 33, no. 4, pp. 402–407, Oct. 2022, doi: 10.4103/ijdr.ijdr_955_21.

[177] V. de Vos, K. M. Timmins, I. C. van der Schaaf, Y. Ruigrok, B. K. Velthuis, and H. J. Kuijf, "Automatic cerebral vessel extraction in TOF-MRA using deep learning," *Proc. SPIE*, vol. 11596, pp. 651–660, Feb. 2021, doi: 10.1117/12.2581226.

[178] K. Sirinukunwattana et al., "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, Jan. 2017, doi: 10.1016/j.media.2016.08.008.

[179] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 341–346.

[180] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[181] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.

[182] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[183] Ö. Ö. Karadag and Ö. Erdas Çiçek, "Experimental assessment of the performance of data augmentation with generative adversarial networks in the image classification problem," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Oct. 2019, pp. 1–4.

[184] S. Motamed, P. Rogalla, and F. Khalvati, "Data augmentation using generative adversarial networks (GANs) for GAN-based detection of pneumonia and COVID-19 in chest X-ray images," *Informat. Med. Unlocked*, vol. 27, Jan. 2021, Art. no. 100779, doi: 10.1016/j.imu.2021.100779.

[185] W. You and X. Wang, "View enhanced jigsaw puzzle for self-supervised feature learning in 3D human action recognition," *IEEE Access*, vol. 10, pp. 36385–36396, 2022, doi: 10.1109/ACCESS.2022.3165040.

[186] S. Manna, S. Bhattacharya, and U. Pal, "Self-supervised representation learning for detection of ACL tear injury in knee MR videos," *Pattern Recognit. Lett.*, vol. 154, pp. 37–43, Feb. 2022, doi: 10.1016/j.patrec.2022.01.008.

[187] S. Yamaguchi, S. Kanai, T. Shioda, and S. Takeda, "Image enhanced rotation prediction for self-supervised learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 489–493.

[188] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, "Unsupervised representation learning by sorting sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 667–676.

[189] H. Jiang, G. Larsson, M. M. G. Shakhnarovich, and E. Learned-Miller, "Self-supervised relative depth learning for urban scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 19–35.

[190] S. Jenni and P. Favaro, "Self-supervised feature learning by learning to spot artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2733–2742.

[191] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 793–802.

[192] P. Chen, S. Liu, and J. Jia, "Jigsaw clustering for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11526–11535.

[193] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, "Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101746, doi: 10.1016/j.media.2020.101746.

[194] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, Dec. 2012, pp. 1–9.

[195] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[196] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6391–6400.

[197] M. Goldblum et al., "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Dec. 2023, pp. 29343–29371.

[198] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5410–5419.

[199] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[200] X. Wang, Y. Wang, X. Xu, F. Yan, and Z. Zeng, "Two-stage deep neural network with joint loss and multi-level representations for defect detection," *J. Electron. Imag.*, vol. 31, no. 6, Dec. 2022, Art. no. 063060, doi: 10.1117/1.jei.31.6.063060.

[201] W. Zhang, D. Chen, and Y. Kong, "Self-supervised joint learning fault diagnosis method based on three-channel vibration images," *Sensors*, vol. 21, no. 14, p. 4774, Jul. 2021, doi: 10.3390/s21144774.

[202] S. Liu, H. Fan, S. Lin, Q. Wang, N. Ding, and Y. Tang, "Adaptive learning attention network for underwater image enhancement," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5326–5333, Apr. 2022, doi: 10.1109/LRA.2022.3156176.

[203] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?" 2021, *arXiv:2112.10740*.

[204] S. G. Dhekane, H. Haresamudram, M. Thukral, and T. Plötz, "How much unlabeled data is really needed for effective self-supervised human activity recognition?" in *Proc. Int. Symp. Wearable Comput.*, Oct. 2023, pp. 66–70.

[205] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 3, pp. 98–106, Sep. 2023, doi: 10.1109/MGRS.2023.3281651.

[206] F. Zheng, J. Cao, W. Yu, Z. Chen, N. Xiao, and Y. Lu, "Exploring low-resource medical image classification with weakly supervised prompt learning," *Pattern Recognit.*, vol. 149, May 2024, Art. no. 110250, doi: 10.1016/j.patcog.2024.110250.

[207] X. Sun, Z. Wang, Z. Lu, and Z. Lu, "Self-supervised graph representations with generative adversarial learning," *Neurocomputing*, vol. 592, Aug. 2024, Art. no. 127786, doi: 10.1016/j.neucom.2024.127786.

[208] H. Chen et al., "Enhancing human activity recognition in smart homes with self-supervised learning and self-attention," *Sensors*, vol. 24, no. 3, p. 884, Jan. 2024, doi: 10.3390/s24030884.

[209] W. Chen, C. Li, D. Chen, and X. Luo, "A knowledge-based learning framework for self-supervised pre-training towards enhanced recognition of biomedical microscopy images," *Neural Netw.*, vol. 167, pp. 810–826, Oct. 2023, doi: 10.1016/j.neunet.2023.09.001.

[210] L. Yan, J. Yang, and J. Wang, "Domain knowledge-guided self-supervised change detection for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4167–4179, 2023, doi: 10.1109/JSTARS.2023.3270498.

[211] R. Ding, A. Yadav, E. Rodriguez, A. C. A. L. da Silva, and W. Hsu, "Tailoring pretext tasks to improve self-supervised learning in histopathologic subtype classification of lung adenocarcinomas," *Comput. Biol. Med.*, vol. 166, Nov. 2023, Art. no. 107484, doi: 10.1016/j.compbiomed.2023.107484.

[212] Y. Chen and L. Bruzzone, "Self-supervised SAR-optical data fusion and land-cover mapping using Sentinel-1/-2 images," 2021, *arXiv:2103.05543*.

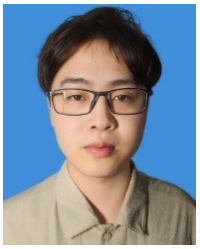
[213] P. Jain, B. Schoen-Phelan, and R. Ross, "Multi-modal self-supervised representation learning for Earth observation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3241–3244.

[214] A. Fedorov et al., "Self-supervised multimodal learning for group inferences from MRI data: Discovering disorder-relevant brain regions and multimodal links," *NeuroImage*, vol. 285, Jan. 2024, Art. no. 120485, doi: 10.1016/j.neuroimage.2023.120485.

[215] Y. Zong, O. M. Aodha, and T. Hospedales, "Self-supervised multimodal learning: A survey," 2023, *arXiv:2304.01008*.

Zhihua Chen received the Ph.D. degree in mechanical engineering from New Jersey Institute of Technology, Newark, NJ, USA, in 2002.

He is currently a Professor with the National Key Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing, China. His research interests include spherical robots, computer vision, image processing, and deep learning.



Bo Hu received the B.E. degree in automation from Central South University, Changsha, China, in 2021. He is currently pursuing the master's degree in robots with the National Key Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing, China.

His research interests include spherical robots, computer vision, image processing, and deep learning.

Zhongsheng Chen (Member, IEEE) received the Ph.D. degree in mechanical engineering from the National University of Defense Technology, Changsha, China, in 2004.

He is currently a Professor with the College of Automotive Engineering, Changzhou Institute of Technology, Changzhou, China. His research interests include prognostics and health management, industrial machine vision, defect detection, and deep learning.

Jiarui Zhang received the master's degree from Nanjing University of Science and Technology, Nanjing, China, in 2024, where he is currently pursuing the Ph.D. degree in computer vision and robots with the National Key Laboratory of Transient Physics.

His research interests include computer vision, object detection, and deep learning.