Self-Supervised Remote Sensing Feature Learning: Learning Paradigms, Challenges, and Future Works

Chao Tao©, Member, IEEE, Ji Qi©, Mingning Guo, Qing Zhu©, and Haifeng Li©, Member, IEEE

Abstract— Deep learning has achieved great success in learning features from massive remote sensing images (RSIs). To better understand the connection between three feature learning paradigms, which are unsupervised feature learning (USFL), supervised feature learning (SFL), and self-SFL (SSFL), this article analyzes and compares them from the perspective of feature learning signals and gives a unified feature learning framework. Under this unified framework, we analyze the advantages of SSFL over the other two learning paradigms in RSI understanding tasks and give a comprehensive review of existing SSFL works in RS, including the pretraining dataset, SSFL signals, and evaluation methods. We further analyze the effects of SSFL signals and pretraining data on the learned features to provide insights into RSI feature learning. Finally, we briefly discuss some open problems and possible research directions.

Index Terms— Deep learning, Earth observation, remote sensing, self-supervised feature learning (SSFL).

I. INTRODUCTION

Obtaining a feature representation being both discriminative and invariant is fundamental in remote sensing image (RSI) understanding [1], [2], [3]. The feature discrimination concern is closely related to the visual manifold hypothesis [4], that is, the real-world data samples presented in high-dimensional spaces are usually embedded into a lower dimensional manifold in human vision perception, where samples from different categories are naturally clustered. In other words, visual separability is highly dependent on the discriminative power of feature representation. The advantage of human vision lies in the invariance of object recognition under various environments. In the field of RSI understanding, the visual feature of objects is generally influenced by both external and internal factors [5], [6]. External factors are the variation of imaging condition (e.g., illumination, viewpoint, and scale) and imaging mechanism (e.g., optical and synthetic aperture radar (SAR) image), and the internal factors are the internal change over time like vegetation phenological shifts. These factors lead to significant changes in visual features, which make RSI understanding tasks very challenging. Thus, obtaining the feature representation insensitive to changes has received much attention, which is also the key to ensuring the generalization ability of the RSI understanding model for different regions, different times, and different imaging conditions.

Instead of traditional hand-crafted feature engineering, learning feature representation from massive data has attracted great attention during the past decade and exhibited very impressive performance in many RSI understanding tasks [7], [8], [9], [10], [11], [12]. The success of feature learning is mainly attributed to two factors. First, it uses deep hierarchical network architectures. Bengio et al. [13] discussed that the key to a good representation is to use deep architectures for feature learning, which carries two advantages: 1) deep architectures reuse features and 2) deep architectures help to learn abstract semantic features that keep invariant under most changes of the input. Second, it uses a large number of samples for deep network model training. Sun et al. [14] performed an empirical study using the JFT-300M dataset containing 300M samples for representation learning. They found that the performance of feature representation increases linearly with the number of training data for high-capacity models. In the remote sensing community, Long et al. [15] compared the performance of feature learning trained on existing remote sensing benchmark datasets. They argued that using a small number of samples may limit the performance of feature learning because a large number of remote sensing samples are more likely to represent the feature distribution of land cover in the real world.

Early studies on feature learning aimed at learning features in an unsupervised way using models like sparse coding [16], [17], [18], [19], [20], autoencoder [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], and deep belief networks [31]. However, due to the lack of ground truth and an effective feedback mechanism in the learning process, the learned features may not be both discriminative and invariant enough for RSI understanding tasks. With the development of convolutional neural networks (CNNs), many supervised feature
learning (SFL) methods were proposed and have proved to be promising in extracting high-level visual features with distinguishability and invariance and are successfully applied to RSI understanding tasks. Though SFL paradigms have witnessed great progress over unsupervised feature learning (USFL) paradigms, they require large-scale, high-quality labeled data, which are difficult to obtain as accurately annotating RSIs is tedious and requires rich experience and geographic knowledge. Moreover, the annotating approach used for RSIU tasks is extremely task-dependent. For example, scene classification tasks require image-level annotation, while semantic segmentation tasks require pixel-level annotation. In other words, different tasks need different level annotations, so great efforts need to be paid to constructing datasets in a task-dependent way.

Unlike machine vision which is “taught” by labeled data, human-like vision is not limited to a specific task or specific dataset, and human language-based labels are not the prerequisite for constructing the human visual system. Thus, a new feature learning paradigm, self-SFL (SSFL), was proposed in the field of natural language process (NLP) [32], [33], [34], [35] and computer vision (CV) [36], [37], [38]. It uses human-designed task-agnostic self-supervised learning signals to generate pseudolabels for massive unlabeled data, thereby replacing human labels to guide the model learning.

Up to now, several surveys on SSFL in CV have been done [39], [40], [41], [42]. To bridge the gap between the progress of SSFL in CV and remote sensing, Wang et al. [42] summarized some representative methods of SSFL and analyzed their application in remote sensing tasks. Here, we want to further contribute to the study on SSFL for remote sensing applications as follows.

1) For learning feature signals, we introduce a unified feature learning framework to understand the connection and difference between the SSFL paradigms and other feature learning paradigms.

2) Based on this unified framework, we propose a new taxonomy of SSFL in terms of data, SSFL signals, and evaluation methods and perform an extensive literature review of the remote sensing community in these three aspects. Moreover, we comprehensively compare the properties of training data and self-supervised learning signals to provide a deeper understanding of the necessary conditions for better self-supervised learning for RSI.

3) We discuss the limitations of existing remote sensing SSFL and suggest potential research directions including new ones that are not covered by previous surveys.

The rest of this survey is organized as follows. Section II introduces the mainstream feature learning paradigms in RS and clarifies the advantages of SSFL over other feature learning paradigms. Section III reviews the existing studies on SSFL in remote sensing, considering the training data, SSFL signals, and feature evaluation methods. Section IV analyzes how the SSFL signal and properties of pretrained data affect the performance of the learned features in downstream tasks. Section V gives a comparative study between three dominant learning paradigms: SFL, USFL, and SSFL. In Section VI, we suggest four potential research directions of SSFL in remote sensing. Some conclusions are drawn in Section VII.

II. Unified Feature Learning Paradigm for Feature Learning Signals

A. Definition

Instead of handcrafted feature engineering, learning feature representation from massive data has been the mainstream feature extraction method due to its impressive performance in many RSI understanding tasks. In terms of the learning paradigm, feature learning methods can be divided into three categories: USFL, SFL, and SSFL. To understand the intrinsic relationship between these learning paradigms, we analyze these paradigms from the perspective of feature learning signals and proposed a unified feature learning framework. Feature learning paradigm includes four parts: data, model, loss function, and optimizer, which can be mathematically described as

$$
\min_{\theta} \mathcal{L}(f_{\theta}(\cdot), D) \tag{1}
$$

where $f_{\theta}(\cdot)$ is the model used for feature learning and $\theta$ is the parameters needed to be learned in the model. $f_{\theta}(\cdot)$ can be a deep convolutional model, autoencoder model, Gaussian probability distribution model, and so on; $D$ denotes the samples used to train the model, which can be labeled or unlabeled datasets; $\mathcal{L}$ is the loss function, such as L2 loss and cross-entropy loss, which describes the metric used to approximate the ground truth or predefined optimization objective; $\min_{\theta}$ denotes the optimizer, such as stochastic gradient descent (SGD) and evolutionary algorithm (EA), used to find the optimal parameters of the model.

In the SFL paradigm, the training sample set $D$ is usually denoted as $D = \{(x_i, y_i)|x_i \in X, y_i \in Y\}_{i=0}^N$, where $x_i$ and $y_i$ are the $i$th data sample and its corresponding label. $y_i$ is regarded as the ground truth of $x_i$ to guide the feature learning process, but this is set by humans according to their criteria. That is, if we find a learning signal hidden in the data as the ground truth, we can perform feature learning without labeled samples. We call this learning signal the generic supervised learning signal and define a unified feature learning framework as follows:

$$
\min_{\theta} \mathcal{L}(f_{\theta}(\cdot), D, S) \tag{2}
$$

where $S$ denotes the generic supervised learning signal. In the following, we will explain how to define existing feature-learning paradigms under this unified feature-learning framework.

B. Unsupervised Feature Learning

From the perspective of learning signals, USFL uses the data intrinsic structure as the learning signal to guide the feature learning process. Specifically, given an unlabeled dataset $D = \{x_i|x_i \in X\}_{i=0}^N$, for each $x_i$, a pseudolabel $\tilde{y}_i$ is constructed...
using the data intrinsic structure as a generic supervised learning signal. Thus, USFL can be defined as follows:

$$\min_{\theta} \mathcal{L}\left(f_\theta(\cdot), \left\{ \left(x_i, \tilde{y}_i\right) \mid x_i \in X, \tilde{y}_i \in \tilde{Y} \right\}_{i=0}^N, S \right)$$  \hspace{1cm} (3)$$

where $S:X \rightarrow \tilde{Y}$. The commonly used USFL methods for remote sensing include the sparse coding model and autoencoder model. Specifically, the sparse coding model uses a complete visual dictionary learned from a large amount of unlabeled data to guide the reconstruction of the original image to obtain the corresponding sparse representation. The autoencoder is an encoder–decoder network that seeks to learn a compressed sparse representation of input by minimizing the reconstruction error between the input and output data. Although the two models use different algorithms, they both are based on the distribution law of data manifolds, that is, high-dimensional data of the same category tend to be concentrated near a low-dimensional manifold. However, due to the lack of ground truth or an effective feedback mechanism related to the specific RSI understanding task in the learning process, the learned features may not be both discriminative and invariable enough for RSI understanding tasks.

C. Supervised Feature Learning

In the SFL paradigm, the training samples $D$ are $D = \left\{ (x_i, y_i) \mid x_i \in X, y_i \in Y \right\}_{i=0}^N$, where $x_i$ denoting the $i$th data sample and $y_i$ denoting the corresponding data label. The data label $y_i$ can be regarded as the ground truth, which is set by human knowledge. Thus, the learning signal in the SFL paradigm is considered to be generated by human knowledge and SFL can be defined as follows:

$$\min_{\theta} \mathcal{L}\left(f_\theta(\cdot), \left\{ (x_i, y_i) \mid x_i \in X, y_i \in Y \right\}_{i=0}^N, S \right)$$  \hspace{1cm} (4)$$

where $S:X \rightarrow Y$. Deep CNN (DCNN) is a typical method of SFL, which has achieved great success in RSI understanding tasks due to its advantages in representation learning. However, this method requires a large number of high-quality labeled data, which are extremely expensive to acquire, due to the spatial–temporal heterogeneity of remote sensing data. This contradiction seriously restricts the application of DCNN in large-scale and complex RSI understanding tasks and brings two problems as follows.

1) From the perspective of training data, the successful application of DCNNs in RSI understanding highly depends on strong supervision, i.e., a large quantity of labeled training data. However, building a large high-quality training dataset is challenging: First, the representations of ground objects in RSIs vary with weather, climate, lighting, season, and satellite imaging condition. That is, the representation is affected by the temporal heterogeneity of RSIs. Each training sample only represents the characteristics of the objects at a specific time point, so building the RSI dataset for learning generic representations requires almost infinite labeled samples. Second, the distribution of ground objects varies over regions due to different climates and human activities [6], [43], [44], [45]. Such spatial heterogeneity leads to sample category imbalance inside the training set or between the training and test sets during supervised learning, which leads to “over-learning” or “under-learning” problems in application [46], [47].

2) From the perspective of the learning mechanism, existing supervised learning trains the model on a limited number of samples, resulting in the contradiction between the closed sample set and the dynamics and complexity of ground object features, which causes the performance collapse of the model [48], [49], [50]. Although this problem can be alleviated by increasing the number of labeled samples, the extremely high cost of obtaining high-quality data labels makes it difficult to model temporal heterogeneity. This is the inherent flaw of the supervised learning paradigm. Furthermore, supervised learning takes the semantic support provided by labels as the only learning signal. If labels are treated as a kind of prior knowledge, the model will be restricted in the given knowledge during the learning process. However, the large amount of remote sensing data contains far more information than that provided by sparse labels, theoretically. Therefore, over-reliance on manual labeling may lead to an “inductive bias” problem.

D. Self-Supervised Feature Learning

SSFL uses human-designed task-agnostic learning signals to generate pseudolabels for massive unlabeled data, thereby replacing human labels to guide the model learning. Specifically, given an unlabeled dataset $D = \{x_i \mid x_i \in X\}_{i=0}^N$, for each $x_i$ in $X$, a pseudolabel $\tilde{y}_i$ is generated by human-designed learning signals. Thus, SSFL can also be defined by the unified feature learning framework as follows:

$$\min_{\theta} \mathcal{L}\left(f_\theta(\cdot), \left\{ (x_i, \tilde{y}_i) \mid x_i \in X, \tilde{y}_i \in \tilde{Y} \right\}_{i=0}^N, S \right)$$  \hspace{1cm} (5)$$

where $S:X \rightarrow \tilde{Y}$. Although the USFL method and SSFL both use pseudolabels for feature learning, the former constructs pseudolabels using only the manifold hypothesis, but the latter use diverse pretext tasks (e.g., image inpainting, colorization, and jigsaw). Thus, many studies have shown that the feature learned in a self-supervised manner is more powerful than the one learned in an unsupervised manner [51].

Recently, SSFL has got significant advances in NLP [52], [53]. [54] and CV [55], [56], [57]. We believe SSFL is effective and robust for RSIs understanding tasks when the labeled data are insufficient. The reasons are as follows.

1) From the perspective of data: the rapidly evolving Earth observation system provides abundant remote sensing data [15], [58]. However, most of these data are unlabeled, so we cannot use them directly for SFL. Learning features from massive unlabeled data in a self-supervised manner can alleviate dependence on labeled samples as well as the sample-imbalance problem in SFL [59].

2) From the perspective of the feature learning mechanism: learning features in a label-free and task-independent
way may be closer to the human visual process than SFL. The human visual recognition system is not limited to a specific task or specific dataset. Therefore, feature learning via supervised data-dependent and task-dependent ways may limit the generalization ability of feature representation. In addition, human language-based labels are not the prerequisite for constructing the human visual system. For example, a person who has no remote sensing knowledge can quickly extract the key features for distinguishing different land covers by observing a certain number of RSIs, and these features are invariant with the changes in lighting, perspective, and scale.

III. PROGRESS OF SELF-SUPERVISION LEARNING ON REMOTE SENSING DATA

Based on the unified framework of SSFL (Section II-D), in this section, we review research on SSFL in the remote sensing community following the proposed taxonomy: data, SSFL signals, and evaluation methods. Section III-A describes the progress in constructing large remote sensing datasets that can be used for SSFL. Section III-B reviews how scholars construct SSFL signals to learn features from unlabeled remote sensing datasets. Section III-C lists the common methods to evaluate the features learned by SSFL signals.

A. Self-Supervised Learning Datasets

Studies [14], [38], [40], [60], [61] have shown that the models trained on large-scale and diverse datasets have two advantages: 1) the learned parameters provide a good starting point, so models can converge faster when they are trained on other tasks and 2) such models usually have learned rich features, which help to reduce the over-fitting risk when they are trained on new tasks. From the perspective of data volume and semantic diversity, this section reviews the RSI datasets suitable for SSFL.

1) Manually Constructed Labeled Remote Sensing Image Datasets: In the past decades, researchers in the RS field have released many high-quality RSI datasets, as shown in Table I. RSD46-WHU [62] is the first high-resolution RSI dataset with a sample size of more than 100,000, containing 117,000 labeled samples of 46 scene categories from Google Earth and Tianditu platforms. Later, Christie et al. [63] constructed the functional map of the world (fMoW) dataset, consisting of 132,716 sample pairs. Each sample pair contains multitemporal and multimodal images collected from many satellites of the DigitalGlobe series, so fMoW has up to 1,047,691 medium- and high-resolution images. More importantly, the fMoW dataset spatially covers more than 200 countries, temporally spans 15 years, and has 62 types of RS scenes in terms of semantic content. Therefore, compared with other small-scale datasets, the fMoW dataset provides a more comprehensive representation of the real-land cover, which allows models to learn diverse data features. Recently, Long et al. [15], [64] constructed the first million-scale high-resolution RSI scene classification dataset in the RS field, called million-aerial image dataset (AID). It contains 1,000,848 land cover and land use scenes of 51 categories. By collecting aerial images under various circumstances (i.e., illumination, viewpoint, scale, weather, season, and geo-location), million-AID can well represent the feature distribution of RS scenes in the real world.

For the low- and medium-resolution multispectral datasets, Sumbul et al. [65] constructed a scene classification dataset named BigEarthNet, containing 590,326 samples collected from Sentinel-2 satellite data and assigned one or more land cover labels to each sample. The SeasoNet [66] dataset is also constructed based on Sentinel-2 satellite data, which has nearly two million samples with pixel-level labels. In addition, Schmitt et al. [67] constructed a large-scale multimodality land cover classification dataset, SEN12MS, based on Sentinel-1 and Sentinel-2 satellite data, which contains 180,662 SAR-multispectral image pairs sampled from different regions and seasons, and thus, has a wide temporal and spatial coverage. The SEN12MS can be used for joint learning of multimodal RS data features.

Self-supervised learning does not require manual annotation, so the above-mentioned datasets can be directly used for SSFL after discarding labels. However, the size of these datasets is much smaller than that of the datasets used for CV, which is as large as hundreds of millions. Therefore, there is an urgent need to build larger datasets for SSFL.

2) Remote Sensing Image Dataset Constructed by Automated Sampling: By self-supervised learning, the feature learning process does not require manual labeling, which reduces the labeling cost for constructing ultralarge-scale RSI datasets.

Grid sampling is a traditional way to automatically collect RSI samples. As shown in Fig. 1, the grid sampling method meshes a large-scale RSI of size $H \times W$ into $(H/h) \times (W/w)$ nonoverlapped patches of size $h \times w$. This method can quickly and easily build a large-scale self-supervised learning dataset based on massive RSIs. However, since the spatial distribution of geographic elements is naturally unbalanced [44], [68], this method can hardly ensure data diversity.

To solve the problem, Manas et al. [69] collected Sentinel-2 images from urban areas globally to construct a large-scale dataset, called SeCo. This is based on the hypothesis that cities and their surrounding areas have the most diverse and relatively balanced land cover types. SeCo contains more than 1 million RSIs from over 200,000 urban areas in different seasons. Further, Wang et al. [70] constructed a self-supervised dataset, SSL4EO-S12, containing three million image samples, larger than all previous datasets. Compared with SeCo, SSL4EO-S12 covers a wider spatiotemporal range and contains multimodal remote sensing data including Sentinel-1 and Sentinel-2, supporting multimodal remote sensing feature learning. In addition, to enable the model to learn image-audio features, Heidler et al. [71] constructed the first large-scale RSI-audio dataset, called SoundingEarth, by two steps: 1) crawling surface audio data with geo-coordinates from an open source project called Radio Aporec::Maps1 and 2) collecting the corresponding RSIs from Google Earth platform based on the geo-coordinates of the audio. SoundingEarth contains

1https://archive.org/details/radio-aporec-maps
about 3500 hours of audio data and 50,545 RSIs with a spatial resolution of approximately 0.2 m/pixel.

In addition to semantic diversity, the class balance and image resolution of the self-supervised learning datasets are also crucial for the model to learn valuable images [72]. Therefore, Tao et al. proposed an automatic RSI sampling method guided by geographic data products. To guarantee semantic diversity and image resolution, RSIs with a resolution of 1m/pixel are collected from the Google Earth platform. Then, the label from a global land cover product FROM_GLC10 [44] is used to guide the sampling of natural geographical samples (e.g., woodlands and grasslands). In addition, this method uses the label from the open street map (OSM) to guide the sampling of man-made geographical elements (e.g., airports, parking lots, and schools). In this way, they first obtained a dataset containing more than three million RSIs, called the TOV-RS dataset, with a balanced number of natural and man-made samples. Furthermore, to ensure a balanced sampling of the subcategories of natural and man-made categories, a class-balanced resampling strategy is proposed to postprocess the obtained samples by their labels. Finally, they obtain a relatively class-balanced dataset, called the TOV-RS-balanced dataset. Table II summarizes the basic characteristics of the existing automated sampling-based datasets.

### B. Self-Supervised Feature Learning Signals

In SSFL, feature learning does not rely on manual labels but is guided by pseudolables, which are obtained automatically by mining the association information from massive unlabeled data with human-designed self-learning signals. Many studies showed that the choice of SSFL signal is crucial for the model learning ability of good features [36], [37], [55], [57], [73], [74]. As shown in Fig. 2, the existing SSFL signals can be classified into three categories: generative, predictive, and contrastive. In this section, the principles for designing these signals and the characteristics of the learned features are introduced.

#### 1) Generative Learning Signals

Generative learning signals train the model to reconstruct an original input from a partially corrupted one for feature learning. It assumes that the model can recover the missing information if the contextual information features are well-learned. The construction process of this learning signal is as follows.

**Step 1:** Corrupting the origin data $x$ by adding random noise, masks, or downsampling $x$ to obtain a destroyed version $\tilde{x}$.

**Step 2:** A model $f(\cdot)$ with an encoder–decoder architecture learns features by minimizing the objective function $\| f(\tilde{x}) - x \|^2_2$. 

Fig. 2. Categories of the SSFL signal in the remote sensing field: generative, predictive, and contrastive.

The generative learning signals can be further divided into learning signals based on spatial missing content generation and that based on temporal missing content generation.

a) Learning signals based on spatial missing content generation:

1) Image-denoising: Since noise leads to image content missing, early studies construct this learning signal using image-denoising tasks for learning abstract and robust features [21], [75]. These studies are commonly based on the hypothesis that the model must learn stable spatial features to recover clear images from the noise images. Based on the hypothesis, image denoising-based SSFL methods have been used to learn features from various RSIs [76], [77], [78], [79], [80]. For example, Zhang et al. [81] employs a stacked autoencoder to learn spatial features from both SAR and high-resolution optical RSIs via the denoising task, and thus, benefit downstream cross-modal change detection tasks.

2) Inpainting: Pathak et al. [82] further developed a similar learning signal using a more challenging inpainting task, which corrupts images by patch-level masks instead of pixel-level noise. The motivation is that the model should capture higher level context features to predict a reasonable hypothesis for the missing part(s) of the input image. Similar work can be found in [82], [83], [84], [85], and [86]. Singh et al. [87] used this kind of SSFL signal to learn spatial features from high-resolution RSIs. To increase the inpainting task difficulty, they introduced an adversarial training framework to select texturally complex regions for masking instead of masking randomly selected regions. Their experiments showed that increasing the difficulty of inpainting tasks can improve the model learning ability of better features for downstream semantic segmentation tasks. Similarly, the masked autoencoder (MAE) method makes the inpainting task more challenging by masking more than 75% of the image [88], thus stimulating the learned transformer model [89], [90] to capture key features of images. Most recently, MAE-based SSFL methods have received much attention for their power of generic feature learning [91], [92], [93], [94], [95]. Sun et al. [96] applied the MAE method to SSFL for a three million unlabeled RSI dataset. Experiments show that the learned features generalize well to various downstream tasks, including scene classification, target recognition, semantic segmentation, and change detection.

3) Super-resolution: The super-resolution-based SSFL signal follows the above-mentioned idea of image missing content generation to recover high-resolution images from blurred low-resolution images for capturing spatial features of the various object profiles [97], [98], [99]. For paired hyperspectral and multispectral images, Yao et al. [100] proposed a two-branch super-resolution model for learning coupled spatial–spectral features based on the hypothesis of content consistency between the two modal data. However, the generalization per-
formance of the features learned by the learning signal based on the super-resolution tasks in RSI understanding tasks remains unexplored.

b) Learning signals based on temporal missing content generation: Constructing such a learning signal for SSFL by temporal missing content generation tasks is based on the assumption that there exist correlations between temporal variation patterns and semantics [101], [102], [103], [104], [105], [106], [107], [108]. For example, as shown in Fig. 3, a model with an encoder–decoder structure should exploit the learned temporal pattern of land cover to reconstruct the missing frame of the input time-series data. Yuan and Lin [106] applied this idea to time-series Sentinel-2 images to learn spectral–temporal features related to crop semantics, as the phenological change pattern of crops on time-series RSIs can reflect the crop type. They add random noise to the original time-series images and construct a transformer model to learn features by recovering clean versions. To increase the difficulty of temporal missing content generation tasks, Yuan et al. [108] proposed a patch-level generative SSFL method that randomly masks some frames from the time-series Sentinel-2 images to reconstruct the missing image frame. The model can understand what the object is in order to accurately predict the spatial relations between the parts of the object.

2) Jigsaw puzzle: Based on the same hypothesis [109], the jigsaw puzzles task [110], [111], [112], [113], [114], [115], [116] is also used to construct the learning signal for SSFL. For example, Noroozi and Favaro [110] constructed a jigsaw puzzle prediction task by randomly shuffling tiles of an image, as shown in Fig. 4(b). The SSFL model is then trained to learn semantic context features by selecting the correct one from the candidate tile order schemes.

There are two key challenges [109], [110], [111] in learning the desired spatial context features for the learning signals constructed in the ways described earlier: 1) shortcuts—texture and color continuity between image tiles are simple cues that can serve as shortcuts for models to solve the above-mentioned prediction tasks. Unfortunately, shortcuts undermine the model’s motivation to learn desired features, reducing the performance of the learned features in downstream tasks. A possible solution to this problem is to use harder versions of the spatial relation prediction task or integrate multiple tasks to construct learning signals. For example, Kim et al. [112] created a complex composite task by integrating the jigsaw puzzle with colorization and inpainting tasks. In this composite task, the model has to solve a more difficult version jigsaw puzzle, in which one image tile is dropped and the rest are decolored. Then, the model is trained to reconstruct the missing image tile and predict the color of the rest ones. 2) Task ambiguities—ambiguity is common in jigsaw puzzles when image tiles lack valid content or are highly similar to each other [110], [111]. In these cases, SSFL models can only solve the tasks by guessing or using other shortcuts, rather than by learning meaningful spatial contextual features. This problem is especially common for RSIs of homogenized scenes, such as seas, deserts, and forests. The above points were confirmed by the experiments of [74], which found that jigsaw puzzles cannot achieve satisfactory performance in the downstream task of high-resolution RSI scene classification.

In addition to using intrimage spatial relationships, interimage spatial relationships can also be exploited to construct the learning signal.

1) Image rotation prediction: Using the image rotation prediction task to construct the signal [117], [118], [119], [120], [121] is based on the assumption that an SSFL model that can accurately predict the rotation angle of the input image should understand the concept of objects in the image [such as the position, type, and pose of the aircraft in Fig. 4(c)]. The construction process of this learning signal is as follows [Fig. 4(c)].

Step 1: Given a predefined set of discrete rotation transformations $G = \{g(\gamma)\}_{\gamma=0}^{K}$, the rotated image $\bar{x}$ is obtained by applying a random rotation transfor-
Fig. 4. Illustration of three types of spatial relationship prediction tasks: (a) relative position prediction, (b) jigsaw puzzle, and (c) image rotation angle prediction. Here, (a) and (b) aim at predicting in-image spatial relationships and (c) aims at predicting inter-image spatial relationships.

**Intra-image spatial relationship prediction**

- **Relative position prediction**
  - Input image
  - $X = (\_\_), Y = 2$

- **Jigsaw puzzles**
  - Input image
  - Candidate set of order scheme
  - $Y = (5, 7, 2, 3, 4, 6, 1, 8, 9)$

**Inter-image spatial relationship prediction**

- **Image rotation prediction**
  - A set of predefined discrete rotation transformations with set size $K = 4$
  - Rotating $0^\circ$ $360^\circ/K$ degrees:
    - $X = (\_\_), Y = 0$
    - $X = (\_\_), Y = 1$
    - $X = (\_\_), Y = 2$
    - $X = (\_\_), Y = 3$

Fig. 5. Illustration of spectral feature prediction.

**Learning signals based on spectral feature prediction:**

Grayscale image

- Encoder
- Decoder
- Reconstruction loss

Multispectral/hyperspectral image

- Representation
- Prediction

Feature learned from this SSFL signal has a better performance in object detection tasks on SAR RSIs than traditional CNNs that focus only on rotation-invariant features. Recent studies [131], [132] also confirmed that learning rotation-equivariant features in a self-supervised manner reduces the dependence of the object detection task on the labeled sample volume. However, this learning signal is not favored in rotation-agnostic RSIs like domed buildings, oceans, and desserts.

b) Learning signals based on spectral feature prediction:

Using this self-supervised learning signal for feature learning...
is based on the assumption of a strong correlation between semantics and its corresponding spectra [133], [134], [135], [136], [137]. For example, vegetation preferentially reflects more near-infrared and green light than other wavelengths of light, so it appears green. And the seawater regions are blue because water preferentially absorbs red spectra. Therefore, the model should understand the semantics of the image to correctly predict the corresponding spectrum. The construction process of this learning signal is as follows (Fig. 5).

**Step 1:** Obtaining gray-scale image \(\tilde{x}\) from an input multispectral/hyperspectral image (HSI) \(x\).

**Step 2:** Training a model \(f(\cdot)\) with encoder–decoder structure by minimizing the objective function \(||f(\tilde{x}) - x||^2_2\).

For RSI with only RGB channels, the above-mentioned process can be regarded as a colorization task, i.e., predicting the RGB color channels using the gray-scale channels. Zhang et al. [133] first applied this idea to construct the learning signal for SSFL. They also considered the shortcuts and task ambiguities problems. But for the learning signals based on spectral feature prediction, these two problems are caused by reasons different from those mentioned in Section III-B2.a. They suggest constructing the colorization task in laboratory space instead of RGB color space because the gray-scale channel \(L\) and the color channels \(ab\) are independent. Thus, to complete the task, the model should understand image semantics, rather than use the interchannel correlation as a shortcut. Empirical experiments [138] show that the learned features using the laboratory space are more beneficial for downstream RSI scene classification tasks than using RGB space. Task ambiguity may occur in the colorization task because, for both natural and RSIs, different objects share the same spectrum and one object may have different spectra. In such cases, the model completes the colorization task by guessing, rather than understanding the semantics of the image. Therefore, Zhang et al. [133] transformed the color value regression problem in colorization tasks into a classification problem to select the most suitable color scheme. Also to reduce task ambiguity, Larsson et al. [134] constructed a learning signal using spectral histogram prediction tasks instead of spectral value regression. For multispectral RSIs, Vincenzi et al. [139] suggested training the model by predicting color channels based on multispectral channels, not including RGB instead of gray-scale channels, to fully exploit the diverse features of each band. For HSIs with hundreds of bands, it may be too difficult to predict the spectral values of all bands directly. Therefore, Yang et al. [140] constructed the learning signal using the spectral order prediction task, rather than spectral information prediction tasks. In this way, the model still needs to understand the semantics of HSI to correctly predict image spectral order, as the fine-grained spectral band combination order usually reflects the type of objects in the HSI.

**3) Contrastive Learning Signals:** Extensive experimental work in psychology has shown that infants learn perceptual categories primarily through observation, rather than linguistic supervision [141]. In this way, they can summarize different manifestations of the same object (i.e., invariance) and distinguish objects by different manifestations (i.e., distinguishablity). Contrast learning signals are designed to mimic this process, which brings different augmented views (positive sample pairs) of the same image closer and separate views (negative sample pairs) of different images, to learn both invariant and distinguishable visual features [141], [142]. The contrastive learning signals are constructed by the following two steps.

**Step 1:** Given an unlabeled training set \(X = \{x_1, x_2, \ldots, x_n\}\), each sample \(x_i\) is augmented by \(T(\cdot)\) to create a pair of positive samples \((x_1^n, x_2^n)\). In contrast, any two augmented views of different samples are treated as a negative sample pair \((x_1^j, x_2^j)\). Here, \(T(\cdot)\) is a stochastic set of augmentations, including random crop, random flip, color distortion, adding noise, and other common digital image processing operations.

**Step 2:** Training a model \(f(\cdot)\) to discriminate the positive and negative samples by embedding them to a proper feature space using the loss function defined as (6), shown at the bottom of the next page, where \(\text{sim}(\cdot, \cdot)\) indicates the similarity of the two sample feature vectors, generally using the cosine similarity. According to the type of data used, contrastive learning signals can be divided into two categories: single-modal contrastive learning signals and multimodal contrastive learning signals.

**a) Single-modal contrastive learning signal:** Single-modal means that positive and negative samples used to construct this learning signal are collected from the same modality. Empirical experiments demonstrate that the way of constructing positive and negative samples has a significant effect on the performance of learned features on downstream tasks [143], [144], [145]. Thus, according to the construction ways of positive and negative samples, single-modal contrastive learning signals can be further classified into instance-level, category-level, and time-series-level signals, as shown in Fig. 6.

**Instance-Level Contrastive Learning Signal:** This [146], [147], [148], [149], [150], [151], [152], [153], [154] takes different augmented views of the same image as positive samples and the views of any other images as negative samples. As a milestone work, Chen et al. [155] proposed a simple yet powerful contrastive learning framework, called SimCLR. For constructing positive sample pairs, they claimed that the combination of multiple augmentation methods significantly outperformed a single-data augmentation method. The reason is that the invariance of the features learned by this learning signal depends on the diversity of data augmentations applied to positive samples [144], [154]. For negative sample construction, they found that a sufficiently large number of negative samples are crucial for the model to learn distinguishable features with high generalization. Their experiments quantitatively demonstrated that larger batch size (i.e., equal to the number of negative samples) and longer training cycles resulted in higher performance of the learned features. To learn features from more negative samples, He et al. [156] proposed MoCo, which uses a dynamic queue to preserve the features of historical negative samples to obtain negative samples with a number far beyond the training batch size limit (i.e., generally equivalent to the number of negative samples). Kalantidis et al. [157] further argued that increasing
the number of negative samples alone does not guarantee performance improvement of contrastive learning. Therefore, they proposed to obtain hard negative samples by mixing the features of negative samples to learn more distinguishable features. From the perspective of model optimization, contrast learning between negative samples can avoid the trivial solution, i.e., the outputs of the model are always constant values. However, some novel methods [158], [159], [160], [161], such as bootstrap your own latent (BYOL) [159] and SimSiam [158], can perform SSFL with only positive sample pairs constructed and the learned features perform surprisingly well in various downstream vision tasks. These methods design the Siamese model with an online and an offline encoder in the same architecture and update the parameters of the two encoders separately. This avoids model performance collapse resulting from the absence of negative samples [162].

In the field of RS, Tao et al. [74] first applied the instance-level contrastive self-supervised method (i.e., SimCLR) to a high-resolution RSI scene interpretation task. They found that the self-supervised learning features outperformed the supervised learning features in scene classification tasks using only 10% of the labeled samples required by the supervised method. On the basis of the first law of geography, Jean et al. [163] assumed that geographically closer regions are more likely to have the same land cover or land use type. To construct the instance-level contrastive learning signal based on this assumption, they took two patches sampled from adjacent locations from a large-scale RSI as a positive sample pair and another patch sampled from distant locations as a negative sample. Following the above-mentioned paradigm, similar works are proposed for learning features from unlabeled RSIs for scene classification [164], [165], [166]. In addi-

\[
\mathcal{L} = -\mathbb{E}_x \left[ \log \frac{\text{sim}(f(x_i^1)^T f(x_i^2))}{\text{sim}(f(x_i^1)^T f(x_i^2)) + \sum_{j=1}^{n-1} \text{sim}(f(x_i^1)^T f(x_j^2))} \right]
\]  

(6)
tion, considering the gap between instance-level contrastive learning and pixel-level segmentation, Li et al. [167] constructed a global style and local matching contrastive learning network (GLCNet) to extend instance-level contrastive learning to super-pixel-level contrastive learning. In this network, different global view representations of the same image are taken as positive samples. Local views are randomly cropped from the positive sample pairs as additional positive samples for local detail feature learning. For the same purpose, some recent research [168], [169] pays attention to the construction of fine contrast-level learning signals, e.g., pixel-level. For example, Muhtar et al. [170] introduced a pixel-level contrastive learning branch called index contrast, which first tracks the spatial index of each identical pixel across two views of the same image, and then takes the corresponding pixels of different views as additional positive samples. Experiments showed that the method outperformed the traditional instance-level contrast learning in land use classification tasks for four high-resolution RSIs because it can learn pixel-level features that are more suitable for semantic segmentation tasks. Similarly, scholars [171] in CV are concerned about the gap between object-level recognition and this learning signal that aims at discriminating images. This is an important but unexplored issue for contrastive learning using RSIs that often cover complex scenes.

**Category-Level Contrastive Learning Signal:** Instance-level contrastive learning signals have been widely used for RSI feature learning [172], [173], [174], [175], [176], [177], [178], [179], [180], [181], [182], [183], [184], but images with the same semantic content may be treated as negative sample pairs, which may mislead feature learning. This is called the false-negative sample problem in contrastive learning [185], [186]. To address this problem, category-level contrastive learning signal only takes images of different “categories” as negative samples [187], [188], [189], [190], [191], rather than discriminating all images even if some of them have similar semantics. As no labels are available as category priors, this signal is constructed through the following steps [Fig. 6(b)].

**Step 1:** Samples are clustered by unsupervised approaches (e.g., k-means) to obtain pseudolabels.

**Step 2:** According to pseudolabels, samples in the same category are treated as positive samples, and those in different categories are treated as negative samples. By doing so, the false-negative sample problem is avoided.

As a milestone work, Caron et al. [192] proposed DeepCluster train the SSFL model by iterating the above-mentioned two steps. Studies [193], [194], [195] have shown that the category-level contrastive learning signal has significant advantages over the instance-level contrastive learning signal in learning distinguishable and invariant features, resulting in better performance in various downstream vision tasks. However, the performance of these methods is heavily dependent on the clustering results. DeepCluster often suffers from the problem of clustering degradation [196]. That is, due to the poor features used for clustering at the early stage of training or the improper setting of parameters of k-means, all samples may be classified into one cluster, and thus, mislead feature learning. To avoid clustering degradation, recent studies [192], [194] add constraints to the clustering process to make the sample size of each category as balanced as possible. Besides, during training, the above-mentioned two-step methods need to frequently traverse the entire dataset for offline clustering, which limits the application to large-scale datasets. To solve this issue, Caron et al. [197] proposed an online clustering-based method, SwAV, that combines the two processes of clustering and cluster assignments prediction into a classification task. Further studies [198], [199], [200], [201], [202] found that joint category-level and instance-level contrastive learning (i.e., multigranular contrastive learning) has higher generalizability than single-granular contrastive learning on diverse downstream tasks. The main reason is that various downstream tasks often require multigranular features. For example, in the object detection task dataset, DOTA, one needs coarse-grained features to distinguish between bridges and aircraft, and fine-grained features to distinguish subcategories of aircraft, such as Boeing 737 and Boeing 747. Therefore, multigranular contrastive learning has become a focus of the studies relating to general-purpose SSFL. Representative approaches of multigranular contrastive learning [203], [204], [205], [206], [207] include prototypical contrastive learning (PCL) [200] and Mugs [202].

**Time-Series-Level Contrastive Learning Signal:** This aims at learning temporal-invariant features of RSIs [208], [209], [210], [211], [212], [213], [214]. It takes RSIs of different time phases in the same region as positive samples and RSIs of different regions as negative samples to construct the learning signal for learning temporal-invariant features. The motivation of this learning signal is that multitemporal RSIs have inherent temporal self-similarity. In other words, ground objects with close geospatial distance and temporal phase should be similar. As a representative method, Ayush et al. [209] proposed the geography-aware contrastive learning method to construct positive samples using spatially aligned images over time instead of different augmented views of the same image. They argued that this learning signal makes the learned features more invariant to subtle temporal changes (e.g., due to imaging conditions), resulting in higher distinguishability for spatial variation, thus benefiting target detection and land cover classification tasks. In addition, considering the seasonal visual differences of some ground objects (e.g., vegetation and cropland), Manas et al. [69] proposed the seasonal contrast framework (SeCo) to learn the temporal-invariant features. SeCo collects Sentinel-2 images from the four seasons globally and considers the images of the same region in different seasons as positive samples and images from different regions as negative samples. The seasonal-invariant features are learned by the model by distinguishing between positive and negative sample representations. However, as the temporal variation of surface coverage, RSIs of the same area at different time phases may contain completely different land use or land cover types, but they still are treated as positive samples by the above method, which would reduce the distinguishability of the learned features. To construct positive samples that always have the same semantic content while representing various spatiotemporal distributions, Huang et al. simulate RSIs as different spatiotemporal visual styles and used them...
as positive samples by introducing an optimal transmission mechanism [215]. In this way, the SSFL model can obtain features with both distinguishability and spatiotemporal invariance, and thus, better generalize to unseen RSI scene classification datasets.

b) Multimodal contrastive learning: We can get multimodal data for the same scene in the same region. These different modal data have great visual differences, but semantically they are the different view representations of the same scene, so they contain embedded invariant features for the same scene. Furthermore, due to the different imaging mechanisms, different modal data of the same scene have complementary features to each other. Huang et al. [216] theoretically and experimentally demonstrated that the more modal data used, the more likely the deep learning model to gain high-quality complementary features and achieve better performance in classification tasks. Therefore, the multimodal contrastive learning method, which takes multimodal data of the same scene as positive samples and data of different scenes as negative ones, has attracted great attention [217], [218], [219], [220], [221], [222], [223], [224], [225]. As shown in Fig. 7, the method has been applied to learn the cross-modal features of different types of RSIs (e.g., SAR-optical RSIs contrastive learning [226], [227], [228], [229], [230], [231], [232], [233], [234]) or the combinations of RSIs and other types of data (e.g., RSI-text and RSI-audio contrastive learning [71], [235], [236]).

1) SAR-optical contrastive learning: Optical RSIs have rich texture details, but the image quality is susceptible to cloud and rain interference. In contrast, SAR satellites provide consistent quality observations in a variety of weather and lighting conditions. Therefore, combining SAR and optical RSIs is an important issue for multimodal RSI contrastive learning. Chen and Bruzzone [230] took pairs of heterogeneous images (SAR-optical) from the same location as positive samples and images from different locations as negative samples. Considering the different data structures of SAR and optical images, the method first constructs two independent encoders to extract features from the two types of images separately and then uses the multimodal contrastive learning signal to train the encoders for SSFL. The performance of the learned features was assessed using the heterogeneous RSI change detection tasks for two SAR-optical image datasets (including the Sentinel-1/2 dataset and the Sentinel-1/Landsat-8 dataset). Moreover, Chen and Bruzzone [231] compared the performance of multimodal (SAR-optical) and single-modal (SAR or optical) SSFL models in three land-cover classification tasks and found that the multimodal contrastive learning approach outperforms single-modal contrastive learning approaches.

2) RSI-text contrastive learning: The research of contrastive learning based on image-text data develops rapidly in the field of CV [237], [238], [239], [240], because a massive amount of paired natural images and corresponding caption text can be automatically crawled from the Internet at a low cost. As a millstone work, contrastive language image pretraining (CLIP) [38] constructed a large-scale Internet crawler dataset containing 400 million image-text sample pairs for multimodal SSFL. CLIP constructs this learning signal by taking paired images and text as positive samples, and unpaired images and text as negative samples. During the pretraining, CLIP feeds image and text samples into the image encoder and text encoder for independent encoding, which generates uniformly structured feature vectors. Then, the contrastive learning signal is used to drive the encoders to learn cross-modal features. By the above-mentioned pretraining, CLIP can directly generate textual descriptions that accurately describe the semantics of the input images. Based on predefined mapping rules, these descriptions can be associated with the categories of the downstream dataset for classification or recognition. Without using any labeled sample for fine-tuning, the multimodal CLIP matches the accuracy of the original ResNet50 supervised trained with 1.28 million labeled images on the ImageNet classification task. However, there are few relevant studies in the field of remote sensing since textual descriptions of RSIs are difficult to obtain. For the text-based RSI retrieval tasks, Mikrukov et al. [236] proposed an RSI-text contrastive learning method for learning cross-modal representations. Specifically, they employed distinct encoders for RSI and text and trained them by a multiobjective contrastive signal to ensure consistency in cross-modal representation and dissimilarity in representation between different texts (or images). Similarly, Yuan et al. [241] used a triplet contrastive loss that makes paired RSI-text samples closer and unpaired samples farther in the learning cross-modal representation space. Furthermore, Zhang et al. [242] combined contrast loss with image and text reconstruction loss to decouple modal-specific features from the learned cross-modal representations to reduce redundancy and improve the efficiency of text-based image retrieval. However, the scalability of the above-mentioned approach is limited by the need for manual annotation of the RSI-text pairs. Therefore, how to obtain geo-tagged textual information of RSIs from crowd-sourced data, such as OpenStreetMap and Twitter, at a low-cost needs investigation.

3) RSI-audio contrastive learning: Multimodal contrastive learning based on image-audio has also attracted the attention of scholars [218], [219], [243], [244]. In an open-source project called Radio Aporee:::Maps,² Heidler et al. [71] constructed a multimodal dataset called SoundingEarth containing 50 545 RSI-audio sample pairs. Specifically, they collected live sounds recorded by volunteers around the world in residential areas, parks, lakes, and wilderness scenes with corresponding high-resolution RSIs from the Google Earth platform. On the basis of SoundingEarth, they used an approach similar to RSI-text multimodal SSFL for RSI-audio feature learning. Experiments demonstrate that introducing

²https://archive.org/details/radio-aporee-maps
audio data into feature learning is also useful for downstream vision tasks, such as remote sensing scene classification and land cover classification.

C. Evaluation Methods for Self-Supervised Feature Learning

The representation capability of the features learned from massive remote sensing data in a self-supervised manner should be assessed accurately and objectively. Specifically, do these learned features have strong distinguishability in remote sensing interpretation tasks? Are they invariant for RSI s of different regions, time phases, and resolutions? The performance of SSFL can be evaluated qualitatively and quantitatively.

1) Qualitative Evaluation Methods: The qualitative evaluation method visualizes the learned feature to evaluate the quality. There are three commonly used methods, including kernel visualization [245], feature map visualization [246], [247], [248], and t-distributed stochastic neighbor embedding (t-SNE) [249] unsupervised clustering visualization. Feature map visualization got a wide application, which leverages technologies, such as deconvolution and class activation visualization, to visualize the activation feature map of the input image obtained by the SSFL model. On this basis, we can observe which regions the model pays attention to when understanding an input image, and are these regions consistent with those attracting human attention?

The literature [198], using the Feature Map Visualization method, assessed the quality of the features obtained from natural image datasets by self-supervised learning based on the vision transformation (ViT) framework and compares those features with those obtained by supervised methods. The experimental results showed that the features learned by the ViT framework in a self-supervised manner without using any category labels can focus on the most relevant regions of the image’s semantic meaning, and their interpretability is even better than the features got by the SFL method. In addition, those features form more separated clusters, indicating that those features have strong category distinguishability. The literature [122] used a similar visual analysis method to assess the quality of the features obtained by the joint SFL method and the SSFL method from a remote sensing scene classification dataset. The results showed that combining two
The fine-tuning approach to test the representation capability of the features learned by the SSFL method on ImageNet, which used 12 publicly available datasets for the downstream tasks of remote sensing scene classification, object detection, and semantic segmentation. The results (Fig. 8) show that in all three types of tasks, the features obtained by the SSFL method perform better than those obtained by the supervised learning method. However, this comparison approach cannot intuitively accomplish a comprehensive comparison between different self-supervised feature methods. For example, SimCLR performs better than MoCov2 for the scene classification task but worse for the target recognition task. Theoretically, self-supervised learning is a task-independent approach, and the generality of the learned features is also very important. Thus, future research could be focused on evaluating the combined performance of the features on multiple tasks.

The literature [70] assessed the transferability of self-supervised learned features by the EuroSAT [250] and BigEarthNet [65] datasets using the linear probe and fine-tuning. They performed SSFL using the EuroSAT dataset after removing the original annotation and transferred the learned features to the downstream BigEarthNet scene classification dataset. The results show that regardless of the feature transfer method, the self-supervised learning features have good transferability. Notably, the fine-tuned transfer method using only 10% of the downstream task labels can get a result similar to that using all downstream task labels.

IV. KEY FACTORS INFLUENCING SELF-SUPERVISED FEATURE LEARNING

The past two years saw the rapid development of SSFL methods and their application in the field of remote sensing. But the key factors influencing SSFL remain unclear. Here, we analyze how the SSFL signal and properties of pretraining data affect the performance of the learned features in downstream tasks.

A. Self-Supervised Feature Learning Signals

For the downstream tasks with very limited labeled data, optimizing the model based on a good starting point can reduce the risk of overfitting. In this case, the choice of SSFL signal is crucial, as it determines what features can be learned by pretraining and whether the features are relevant to downstream tasks. This experiment analyzes how the SSFL signal affects the performance of the learned features in downstream tasks. Based on the categories of SSFL signals, as shown in Fig. 2, we first select three main major categories of SSFL signals (i.e., generative, predictive, and contrastive) for comparison. Then, we compare two subcategories of contrastive signals: single-modal contrast and multimodal contrast.

1) Generative versus Predictive and Contrastive:

a) SSFL methods: To cover three types of mainstream self-supervised learning signals (i.e., generative, predictive, and contrastive), we choose five representative SSFL methods as follows.

1) Generative Signal:

a) Inpainting—Context encoder [82] is a generative SSFL method that learns features by inpainting,
that is, recovering the manually masked parts of the input image.

2) Predictive signal:
   a) Image rotation Prediction—RotNet [117] is a predictive SSFL method that learns features by predicting the rotation angle of the input image.
   b) Spectral feature prediction—Colorization [133] is a predictive SSFL method that predicts the color channel based on the gray-scale version of the input image.

3) Contrastive Signal:
   a) Instance-level contrast—SimCLR [155] is a staple instance-level contrastive SSFL method that learns representations by maximizing the similarity between augmented views of each sample.
   b) Category-Level Contrast—DeepClusterV2 [197] is a typical category-level contrastive SSFL method designed to maximize the feature similarity between the samples of each cluster grouped by k-means.

b) Pretraining datasets: We pretrain the above-mentioned three SSFL methods separately using two large-scale RSI datasets, Million-AID [15] and TOV-RS [72]. As described in Section III-A, Million-AID is a manually collected and labeled dataset, containing 1,000,848 RSI samples of 51 categories with good diversity. The TOV-RS dataset is an automatically collected unlabeled dataset containing three million samples. We randomly selected one million samples from each dataset for pretraining, without using labels.

c) Downstream datasets: To evaluate the performance of the features learned by these SSFL methods, we adopt three types of RSI understanding tasks, which are scene classification, object detection, and semantic segmentation, using different datasets.

1) Dataset for Scene Classification Tasks.
   a) RSD46-WHU [62], [251], a large-scale scene classification dataset containing 117,000 samples of 46 categories. These images were collected from Google Earth and Tianditu.
   b) EuroSAT [250], a scene classification dataset containing 27,000 Sentinel-2 images of ten categories.

2) Dataset for Object Detection Tasks:
   a) DOTA v1.0 [9], [252], a large-scale object detection dataset containing 188,282 objects of 15 categories. It consists of RGB images and grayscale images collected from Google Earth, CycloMedia, GF-2, and JL-1 satellites.
   b) Learning, Vision and Remote Sensing Laboratory (LEVIR) [253], an object detection dataset consisting of over 22,000 Google Earth images with a resolution of 1.0–0.2 m/pixels. It has three categories: airplane, ship, and oil tank.

3) Dataset for Semantic Segmentation Tasks:
   a) Potsdam [254], one of the most popular semantic segmentation datasets in the RS field. It contains 38 UAV images with a resolution of 0.05 m/pixels and a size of 6000 × 6000. This dataset is annotated in six classes.
   b) Deep Globe Land Cover Classification Dataset (DGLCC) [255], a land cover classification dataset containing 1146 images annotated in seven classes. These images are collected by the DeepGlobe satellites. The images have a size of 2448 × 2448 and a resolution of 0.5 m/pixels. These images are mainly located in Thailand, Indonesia, and India, covering a total area of 1716.9 km².

Details of the training and testing set of the above-mentioned datasets are shown in Table III.

d) Implementation details: We pretrain the ResNet50 [256] model for 400 epochs using the three SSFL methods on two pretraining datasets separately and then evaluate the six pretrained models by different downstream tasks. We follow the default hyperparameter settings of the official repository of the three SSFL methods and use 2 NVIDIA A100 GPUs for the experiments, on which the batch size is 256.

For evaluation, the pretrained models are applied to six downstream task datasets by fine-tuning them on the training set of each dataset (Table III). The performance in scene classification and semantic segmentation is evaluated by Cohen’s
TABLE III

| Dataset       | RSD+6-WHU | EuroSAT | DOTA | LEVIR | Potsdam | DGLCC |
|---------------|-----------|---------|------|-------|---------|-------|
| Classes       | 46        | 10      | 15   | 3     | 6       | 7     |
| Training set size | 800       | 800     | 400  | 400   | 200     | 200   |
| Testing set size | 10000     | 10000   | 5000 | 5000  | 2500    | 2500  |

kappa coefficient (kappa), and that in object detection is evaluated by the mean average precision (mAP).

e) Experiment results: The following are our findings.

1) Contrastive learning is an optimal choice because the learned features are superior in most downstream tasks. As shown in Fig. 9(a), when pretrained by Million AID, the performance scores of instance-level contrastive learning are much higher than other types of SSFL methods in all six downstream tasks. When pretrained by TOV-RS [Fig. 9(b)], instance-level contrastive learning also achieves the highest scores in scene classification and semantic segmentation tasks. Therefore, the contrastive learning signal is a promising solution for multiple RSI understanding tasks or uncertain downstream tasks.

2) The choice of SSFL signal should consider the downstream tasks. For the two object detection tasks, as shown in Fig. 9(b), the image rotation prediction method achieves higher scores than the instance-level contrastive learning method. This suggests that the rotation-equivariant features learned by image rotation prediction signals are important for object detection tasks. In contrast, the instance-level contrastive learning aiming at distinguishing RSI scenes may not learn the object-level distinguishable feature. This confirms the gap between instance-level contrastive learning and object detection tasks mentioned in Section III-B3.a.

2) Single-Modal Contrast Versus Multimodal Contrast:

a) SSFL methods: For single-modal contrast, we use the typical version of SimCLR to pretrain optical and SAR RSI data, respectively. For multimodal contrast, we adopt RandomSensorDrop [226] as an additional data augmentation strategy to pretrain paired SAR-optical RSI data by SimCLR. By randomly constructing the positive sample pair based on a single SAR/optical image or a pair of optical-SAR images in each training batch, the model can learn both intramodal and crossmodal representations.

b) Pretraining datasets: For SFL, we adopt the largest publicly available multimodal dataset, SSL4EO-S12 [70]. By sampling both Sentinel-1 and Sentinel-2 imagery in different seasons over more than 250,000 areas near cities worldwide, SSL4EO-S12 contains one million paired SAR-optical RSIs and can support both single-modal contrast learning and multimodal contrast learning.

c) Downstream datasets: For performance evaluation, we adopt a representative multimodal dataset, BigEarthNet-MM [257], as the downstream task dataset. BigEarthNet-MM contains 590,325 paired SAR and optical images with multilabels for land cover classification tasks. The SAR image containing two bands and the optical image containing 13 bands were collected from Sentinel-1 and Sentinel-2, respectively. The sample sizes used for training, validation, and testing are 269,695, 125,866, and 123,723 respectively.

d) Implementation details: First, we pretrained the ResNet50 model on the SAR, optical, and SAR-optical versions of SSL4EO-S12 to obtain three pretrained models respectively, where the first two models are pretrained by the single modal contrast method and the third model is pretrained by the multimodal contrast method. Then, we applied each of the three pretrained models to the SAR, optical, and SAR-optical versions of BigEarthNet-MM, respectively, for downstream evaluation. Specifically, instead of using the entire training set, we fine-tune the model using only 1% of randomly selected samples from the training set and report the average precision (AP) on the test set to evaluate the performance.

e) Experiment results: The following are our findings.

1) Multimodal contrast learning outperforms single-modal contrast learning, even in modality-limited downstream tasks. As shown in Table IV, the model pretrained by the SAR-optical contrast SSFL method consistently outperforms the other two models pretrained by an single-modal contrast SSFL using an arbitrary modal test dataset. The advantages of multimodal contrast learning may arise from two reasons: 1) Multimodal collaborative learning can learn higher quality and more complete latent spatial features than single-modal pretraining [216] and 2) SAR-optical contrast learning models can benefit downstream SAR data tasks by transferring useful information learned from optical pretraining data, and vice versa [70].

2) Optical data seems to provide richer information than SAR data for learning more general features by the contrastive SSFL method. For downstream tasks with arbitrary modal data, the model pretrained on optical data consistently outperforms the model.
pretrained on SAR data (rows 1 and 2 of Table IV). In the downstream task using both optical and SAR, the optical-pretrained model achieved a 1.52% AP improvement over the SAR-pretrained model. Even in the downstream task using SAR data, the optical pretrained model surprisingly outperforms the SAR pretrained model by 0.51% in AP.

B. Effects of the Pretraining Datasets on the Performance of the Self-Supervised Learned Features in Downstream Tasks

Spatial resolution and data volume are the two basic and important properties of a dataset. The former determines the richness of the spatial information that the model learns from RSIs. The latter determines the diversity of the learned features. Thus, the following experiments analyze how these two properties affect the performance of the learned features in downstream tasks.

1) Study on Spatial Resolution:

a) Experiment setup: We use instance-level contrast learning for feature learning, as its superiority has been demonstrated in the above-mentioned experiments. To investigate the effects of spatial resolution, we use one medium-low resolution dataset (SeCo) and two high-resolution datasets (TOV-RS and Million-AID) for pretraining. SeCo has a spatial resolution of up to 10 m/pixel, and the TOV-RS and Million-AID have resolutions of up to 1 and 0.5 m, respectively. To evaluate the performance of the features learned by the model pretrained on different datasets, we use the six downstream datasets described in Section IV-A1 for the experiment.

b) Experiment results: Results for the data of different spatial resolutions are shown in Fig. 10. The following are our findings.

1) Spatial resolution of the pretraining dataset is critical for SSFL. For all three kinds of downstream tasks, especially the object detection task, the performance of the model pretrained on the two high-resolution datasets (TOV-RS and Million-AID) is significantly higher than that of the model trained on the medium-low resolution dataset (SeCo). The reason may be that the high-resolution RSIs dataset can provide more textural and geometric details of the ground objects than the low-resolution dataset, and thus, the learned features are more distinguishable for various objects and scenes.

2) The impact of the pretraining data resolution on SFL may be greater than that of the domain gap between the pretraining and downstream datasets. For example, the EuroSAT dataset and the SeCo are constructed from the Sentinel-2 data, which have a significant domain difference from TOV-RS and Million-AID which are constructed based on Google Earth data. However, as shown in Fig. 10, the models pretrained on TOV-RS and Million-AID achieve almost the same performance scores as the model pretrained on SeCo. For EuroSAT, the kappa of three models pretrained on SeCo, TOV-RS, and Million-AID are 0.8108, 0.8089, and 0.8111, respectively. This suggests that it is unnecessary to sacrifice the resolution of the pretraining data for reducing the domain gap between the pretraining data and the downstream task data.

In summary, we recommend using RSI datasets with as high resolution as possible for SSFL pretraining.

2) Study on Data Size:
a) Experiment setup: To analyze how pretraining data size affects the performance of learned features in downstream tasks, we randomly sample subsets of 30K, 50K, 100K, 300K, and 1M images from each of Million-AID and TOV-RS. The other settings are the same as those in Section IV-B1.

b) Experiment results: Results for different pretraining data sizes are shown in Fig. 11. The following are our findings.

1) Larger pretraining data size brings higher learned features’ performance in most downstream tasks. In Fig. 11(a) and (c), as the pretraining data size grows from 30K to 1M, the performance of the learned features in both scene classification and semantic segmentation tasks improves accordingly. The model pretrained on 1M Million-AID samples outperforms the model pretrained on 30K samples, with kappa increases of 5% and 8% for RSD46-WHU and Potsdam, respectively. This phenomenon is not surprising, as deep learning methods are data-hungry [14]. However, for two object detection tasks, as shown in Fig. 11(b), the performance of the model shows irregular fluctuations as the pretraining data size increases. This confirms the findings in Section IV-A1 that the features learned by instance-level contrastive learning that aims at discriminating RSI scenes are unsuitable for object detection tasks.

2) Little benefit over 300 000 pretrained samples. The performance of the learned features in these tasks seems to saturate before 1M. As shown in Fig. 11(a) and (c), when pretrained on Million-AID, the performance difference between 300 000 and 1 000 000 samples is smaller than 1% in both scene classification and object detection tasks. That is, 70% of the one-million image dataset contributes less than 1% of the accuracy improvement, which can be sacrificed for a remarkable efficiency improvement.

V. SUPERVISED VERSUS UNSUPERVISED VERSUS SELF-SUPERVISED

In this section, we perform a comparative study between three dominant learning paradigms: SFL, USFL, and SFL. Insights distilled from our comparative study can help understand the current progress in self-supervised learning.

A. Experiment Settings

1) Feature Learning Methods: To cover all three feature learning paradigms, we choose five representative methods as follows.

1) Supervised Feature Learning: We pretrained the model on a manually labeled dataset and used the cross-entropy loss function to measure the difference between the model predictions and labels for learning features.

2) Unsupervised Feature Learning: Sparse coding model [258] and autoencoder [259] are two classic USFL methods. The sparse coding model uses a complete visual dictionary learned from unlabeled data to guide the reconstruction of the original image to obtain the corresponding sparse representation. The autoencoder is an encoder–decoder network that learns a compressed sparse representation of input by minimizing the reconstruction error between the input and output data.

3) Self-Supervised Feature Learning: We pretrained the model by using two representative SSFL methods, RotNet based on image rotation prediction signal [117] and SimCLR based on instance-level contrast signal [155].

2) Pretraining and Downstream Fine-Tuning: We pretrain the above-mentioned five feature learning methods using Million-AID [15]. For evaluation, we apply those pretrained models to four downstream RSI understanding tasks, which are RSD46 and EuroSAT, as mentioned in Section IV-A1.

3) Implementation Details: For SFL, we directly use the pretrained ResNet50 model weights released by [260], since the sample labels of Million-AID are not publicly available. For USFL, we pretrain the sparse coding model and autoencoder on million-AID based on scikit-learn3 and PyTorch-lightning4 respectively. For SSFL, we pretrain the ResNet50 model on Million-AID using two SSFL methods based on the same settings, as described in Section IV-A1. For comparison, we add the same task adapter, a simple fully connected layer as the classifier, to the above pretrained models. By fine-tuning the task adopters on the training set of each two downstream datasets (Table III), we compare the three feature learning paradigms and report the performance scores of different pretrained models on the test set (Table V).

B. Experiment Results

Results for different pretraining data sizes are shown in Table V. The following are our findings.

1) Unsupervised learning struggles to achieve satisfactory performance. As shown in Table V, the performance of both unsupervised models lags significantly behind that of the supervised pretrained model in all downstream tasks. For difficult tasks like RSD46, in which the model needs to assign exactly one of the 46 category labels to each RSI, the sparse coding model and the autoencoder are 85.6% and 78.3% lower in kappa than the supervised pretrained model, respectively. Such a gap may be due to that unsupervised learning has no ground truth or effective feedback mechanisms related to the specific RSI understanding.

| Learning paradigms | Methods | RSD46 | EuroSAT |
|--------------------|---------|-------|---------|
| Supervised         | Sparse coding model | 8.83 | 34.80 |
|                    | AutoEncoder | 13.31 | 36.57 |
| Self-supervised    | Cross-entropy loss | 19.11 | 42.83 |
|                    | Image rotation prediction | 62.14 | 81.11 |

3Scikit-learn 1.2.2.
4https://github.com/Lightning-AI
task during the learning process. As a result, the learned features may be neither discriminative nor invariant enough for the RSI understanding task.

2) **The self-supervised learning can achieve comparable results to supervised learning with well-designed learning signals.** The pretrained model using the instance-level contrast method achieves the best results in both two downstream tasks (Table V). For example, the instance-level contrast pretraining model has a kappa value of 1.17% and 0.65% higher than the second-best results obtained by the supervised pretraining model on RSD46 and EuroSAT, respectively. However, the image rotation prediction pretraining model still under-performs the supervised pretraining model on both two downstream tasks. This demonstrates that well-designed self-supervised signals can replace manual annotations to learn good representations directly from large-scale data and finally benefit various downstream RSI understanding tasks. The reason behind this may be that a well-designed self-supervised learning signal provides an effective feedback mechanism that enables the model to learn generic features required by various downstream tasks, such as the discriminative and invariant features learned using instance-level contrast.

**VI. FUTURE WORK**

**A. Theoretical Foundation of the Internal Relationship Between Pretext Tasks and the Performance of SSFL**

SSFL methods learn features by the learning signals constructed by multiple pretext tasks, which is different from the SFL methods using only the class priors from data labels as the feature learning signal. Current research [36], [37], [55], [57], [73], [74], [83] have shown that the choice of learning signals is crucial for the performance of the learned features in downstream tasks, and the learning signal constructed by a poor pretext task may hardly learn the features that can promote the downstream tasks. However, the intrinsic relationship between the learning signal and the feature representation capability is still unclear. What kind of features can be learned from remote sensing data by existing self-supervised learning methods? Are the features learned by different SSFL signals different? If so, is it possible to combine different SSFL signals to achieve complementary feature learning? In addition, does design an SSFL signal relevant to the remote sensing interpretation task can improve the feature representation capability? To answer the above-mentioned questions, it is necessary to carry out related theoretical studies to improve the understanding of the current SSFL paradigms and their intrinsic mechanisms, as well as to provide theoretical guidance for designing better self-supervised pretext tasks.

**B. Transfer Self-Supervised Learned Features to Downstream Tasks**

SSFL signals are designed with various motivations, so the correlation between the learned features and the downstream target tasks may be different. If the feature is regarded as a kind of knowledge, the feature that is strongly associated with the target task may be a kind of “special knowledge,” and the feature that is weakly associated with the target task may be a kind of “general knowledge.” It means that they have different effects on downstream tasks. Thus, using a unified feature transfer strategy, either linear probe or fine-tuning, for the features learned by different SSFL methods may result in the ineffective transfer or even negative transfer, and consequently weaken the generalization performance of the model. Therefore, further research should be done on self-supervised learning feature transfer methods. For example: 1) propose criteria to measure the correlation between self-supervised learned features and downstream tasks and then design feature transfer methods according to their relationship. 2) Develop end-to-end transfer methods from SSFL to supervised downstream tasks, allowing the network to learn features adaptively to fit the downstream tasks.

**C. Continual Self-Supervised Feature Learning Model From Multimodal Remote Sensing Data**

With the rapid development of the Global Earth Observation System of Systems (GEOSS), a large number of remote sensing data become available. To process the constantly growing unlabeled remote sensing data, it is necessary to achieve self-growth feature representation capability through continual self-supervised learning from streaming remote sensing data. In addition, the remote sensing data may be acquired by multiple sensors and in different modalities (e.g., hyperspectral, multispectral, SAR, and satellite video), so constructing models for each modality will increase the training cost and may not be able to learn complementary features. Although a lot of methods related to multimodal SSFL have been developed in the field of CV [38], [218], [222], most of these methods are based on the assumption that the data from different modalities are precisely aligned. So, they are seldom used in the field of remote sensing. The precise spatial–temporal alignment between different modalities of data is a technical problem that has not been solved [261]. Therefore, for the unaligned data without annotation, how to construct continual feature learning paradigms for multimodal data and learn the intrinsic relationship between different modalities is worth researching in the future.

**D. Benchmark for Remote Sensing SSFL Evaluation**

As described in Section III-C, the quality of the self-supervised learned remote sensing features is generally assessed by downstream RSI understanding tasks. However, there is still a lack of a standardized benchmark for SSFL evaluation in the field of RS. In addition, it is important to measure the generality of self-supervised learned features on different downstream tasks. However, the evaluation indicators for different tasks are different, for example, scene classification uses overall accuracy, and object detection generally uses mAP. So, it is difficult to directly aggregate them statistically (such as averaging) to form a comprehensive evaluation indicator. Therefore, the construction of a general evaluation benchmark and comprehensive evaluation indicators for remote sensing SSFL is important.
VII. CONCLUSION

In this survey, we provided a unified feature learning framework to link different remote sensing feature learning paradigms. Based on this framework, we analyzed self-supervised remote sensing feature learning from three perspectives: training data, learning signal, and evaluation metrics. Under such a taxonomy, we provided a systematic review of existing research across these categories. Moreover, we performed a comprehensive study to analyze the impacts of training data selection and the choice of SSFL signals on self-supervised remote sensing feature learning. We also compare the differences in the generalization performance of features obtained by self-supervised, unsupervised, and traditional supervised learning paradigms in various downstream tasks. This study can help to foster the development of SSFL in the remote sensing community. Finally, we discuss some problems that should be solved in future research to improve the development of self-supervised remote sensing feature learning.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful suggestions. They also thank the Open Foundation of Key Laboratory of Ecological Environment Protection of Space Information Application, Henan, China, to support this research with Key Technologies for Water Resources Security in Henan Yellow River Water Supply Zone in Response to Drought and Water Depletion (22-FW-07-0107).

REFERENCES

[1] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” IEEE Geosci. Remote Sens. Mag., vol. 4, no. 2, pp. 22–40, Jun. 2016.
[2] X. X. Zhu et al., “Deep learning in remote sensing: A comprehensive review and list of resources,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2017.
[3] Q. Yuan et al., “Deep learning in environmental remote sensing: Achievements and challenges,” Remote Sens. Environ., vol. 241, May 2020, Art. no. 111716.
[4] C. Fefferman, S. Mitter, and H. Narayanan, “Testing the manifold hypothesis,” J. Amer. Math. Soc., vol. 29, no. 4, pp. 983–1049, Feb. 2016.
[5] H. Cui, G. Zhang, T. Wang, X. Li, and J. Qi, “Combined model color-correction method utilizing external low-frequency reference signals for large-scale optical satellite image mosaics,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 6, pp. 4993–5007, Jun. 2021.
[6] H. Wang, C. Tao, J. Qi, R. Xiao, and H. Li, “Avoiding negative transfer for semantic segmentation of remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4413215.
[7] G. Xia et al., “AID: A benchmark data set for performance evaluation of aerial scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[8] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[9] G. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3974–3983.
[10] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, “A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data,” Remote Sens. Environ., vol. 270, Mar. 2022, Art. no. 112830.
[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” Tech. Rep., 2018.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019, pp. 4171–4186.

[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” Tech. Rep., 2019.

[35] T. B. Brown et al., “Language models are few-shot learners,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 1877–1901.

[36] A. Kolesnikov, X. Zhai, and L. Beyer, “Revisiting self-supervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1920–1929.

[37] A. Newell and J. Deng, “How useful is self-supervised pretraining for visual tasks?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 7343–7352.

[38] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., vol. 139, 2021, pp. 8748–8763.

[39] X. Liu et al., “Self-supervised learning: Generative or contrastive,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 1, pp. 857–876, Jan. 2023.

[40] 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.

[41] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, “Self-supervised representation learning: Introduction, advances, and challenges,” IEEE Signal Process. Mag., vol. 39, no. 3, pp. 42–62, May 2022.

[42] 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.

[43] P. Gong et al., “Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data,” Int. J. Remote Sens., vol. 34, no. 7, pp. 2607–2654, Apr. 2013.

[44] P. Gong et al., “Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017,” Sci. Bull., vol. 64, no. 6, pp. 370–373, Mar. 2019.

[45] H. Liu, P. Gong, J. Wang, N. Clinton, Y. Bai, and S. Liang, “Annual dynamics of global land cover and its long-term changes from 1982 to 2015,” Earth Syst. Sci. Data, vol. 12, no. 2, pp. 1217–1243, Jun. 2020.

[46] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” Inf. Sci., vol. 250, pp. 113–141, Nov. 2013.

[47] D. Tuia, C. Persello, and L. Bruzzone, “Domain adaptation for the classification of remote sensing data: An overview of recent advances,” IEEE Geosci. Remote Sens. Mag., vol. 4, no. 2, pp. 41–57, Jun. 2016.

[48] Z. Shen et al., “Towards out-of-distribution generalization: A survey,” arXiv:2108.13024.

[49] L. Zhang and X. Gao, “Transfer adaptation learning: A decade survey,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 21, 2022, doi: 10.1109/TNNLS.2022.3183326.

[50] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. Change Loy, “Domain generalization: A survey,” arXiv:2103.02593.

[51] L. Schmarje, M. Santarossa, S. Schröder, and R. Koch, “A survey on semi- and unsupervised learning for image classification,” IEEE Access, vol. 9, pp. 82146–82168, 2021.

[52] M. Kayser et al., “E-VIL: A dataset and benchmark for natural language explanations in vision-language tasks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 1244–1254.

[53] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in Proc. EMNLP Workshop: BlackboxNLP: Analyzing Interpreting Neural Nets (NLP), 2018, pp. 353–355.

[54] A. Wang et al., “SuperGLUE: A sticker benchmark for general-purpose language understanding systems,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 32, Red Hook, NY, USA: Curran Associates, 2019, pp. 1–30.

[55] 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. 6390–6399.
D. Muhtar, X. Zhang, and P. Xiao, “Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4411511.

J. Xie, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, “Unsupervised object-level representation learning from scene images,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 34, Red Hook, NY, USA: Curran Associates, 2021, pp. 28864–28876.

C. Agastya, S. Ghebremusse, I. Anderson, C. Reed, H. Vahabi, and A. Todeschini, “Self-supervised contrastive learning for irrigation detection in satellite imagery,” 2021, arXiv:2108.05484.

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.

B. Ren, Y. Zhao, B. Hou, J. Chanussot, and L. Jiao, “A mutual information-based self-supervised learning model for PolSAR land cover classification,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 11, pp. 9224–9237, Nov. 2021.

S. Seneviratne, K. A. Nice, J. S. Wijnands, M. Stevenson, and P. Duan, Z. Xie, X. Kang, and S. Li, “Self-supervised learning-based object-level representation learning from scene images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 1182–1191.

Y. Xu, H. Sun, J. Chen, L. Lei, K. Ji, and G. Kuang, “Adversarial self-supervised learning for robust SAR target recognition,” Remote Sens., vol. 13, no. 20, pp. 4158, Oct. 2021.

X. Zheng, B. Kellenberger, R. Gong, I. Hajnsek, and D. Tuia, “Self-supervised pretrained and controlled augmentation improve rare wildlife recognition in UAV images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, pp. 732–741.

P. Duan, Z. Xie, X. Kang, and S. Li, “Self-supervised learning-based oil spill detection of hyperspectral images,” Sci. China Technolog. Sci., vol. 65, no. 4, pp. 793–801, Apr. 2022.

V. Marosucci, S. Scardapane, and N. Komodakis, “MARE: Self-supervised multi-attention RESu-Net for semantic segmentation in remote sensing,” Remote Sens., vol. 13, no. 16, p. 3275, Aug. 2021.

H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, “Contrastive self-supervised learning with smoothed representation for remote sensing,” IEEE Trans. Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

C. Patel, S. Sharma, V. J. Pasquarella, and V. Gulshan, “Evaluating self and semi-supervised methods for remote sensing segmentation tasks,” 2021, arXiv:2111.10079.

L. Scheibenreif, J. Hanna, M. Mommet, and D. Borth, “Self-supervised vision transformers for land-cover segmentation and classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 1422–1431.

D. Wang, C. Zhang, and M. Han, “FIAD net: A fast SAR ship detection network based on feature integration attention and self-supervised learning,” Int. J. Remote Sens., vol. 43, no. 4, pp. 1485–1513, Feb. 2022.

M. Yang et al., “Partially view-aligned representation learning with noise-robust contrastive loss,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1134–1143.

Z. Zhang, X. Wang, X. Mei, C. Tao, and H. Li, “FALSE: False negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

M. A. Bautista, A. Sanakoyeu, E. Tikhoncheva, and B. Ommer, “CliqueCNN: Deep unsupervised exemplar learning,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 29, Red Hook, NY, USA: Curran Associates, 2016, pp. 478–487.

J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 478–487.

T. Li, Y. Cai, Y. Zhang, Z. Cai, and X. Liu, “Deep mutual information subspace clustering network for hyperspectral images,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
Chao Tao (Member, IEEE) received the B.S. degree from the School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree from the Institution of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, in 2012.

He is currently a Processor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. He is also a Communication Evaluation Expert with the National Natural Science Foundation of China. He has authored more than 30 peer-reviewed articles in international journals from multiple domains, such as remote sensing and computer vision. His research interests include computer vision, machine learning, deep learning, and their applications in remote sensing.

Dr. Tao has frequently served as a Reviewer for several international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTAR), Photogrammetric Engineering and Remote Sensing (PERS), and ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS JPRS).

Ji Qi received the B.E. degree from the School of Geosciences and Info-Physics, Central South University, Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing and computer vision, machine learning, deep learning, and their applications in remote sensing.

Mingning Guo received the B.E. degree from the School of Geosciences and Info-Physics, Central South University, Changsha, China, in 2022, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing and computer vision, machine learning, deep learning, and their applications in remote sensing.

Qing Zhu received the Ph.D. degree from Northern Jiaotong University, Beijing, China, in 1995.

He was a Post-Doctoral Fellow with Wuhan University, Wuhan, China, from 1995 to 1997, where he was a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, from 1998 to 2013. He is currently a Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. His research interests include the digital elevation model, 3-D geographic information science (GIS), virtual geographic environments, and smart cities.

Haifeng Li (Member, IEEE) received the master’s degree in transportation engineering from the South China University of Technology, Guangzhou, China, in 2005, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009.

He is currently a Professor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. He was a Research Associate with the Department of Land Surveying and Geoinformatics, The Hong Kong Polytechnic University, Hong Kong, in 2011, and a Visiting Scholar with the University of Illinois at Urbana-Champaign, Urbana, IL, USA, from 2013 to 2014. He has authored over 30 journal articles. His research interests include geo/remote sensing big data, machine/deep learning, and artificial/brain-inspired intelligence.

Dr. Li is a reviewer of many journals.