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 feature learning paradigms (e.g., unsupervised feature learning (USFL), supervised feature learning (SFL), and self-supervised feature learning (SSFL)), this paper 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 RSIs understanding tasks and give a comprehensive review of the existing SSFL work in RS, including the pre-training dataset, self-supervised feature learning signals, and the evaluation methods. We further analyze the effect of SSFL signals and pre-training data on the learned features to provide insights for improving the RSI feature learning. Finally, we briefly discuss some open problems and possible research directions.

Index Terms—Deep learning, self-supervised feature learning, remote sensing, earth observation

I. INTRODUCTION

OBTAINING a feature representation being both discriminative and invariable is fundamental in remote sensing image (RSI) understanding [1]–[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 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 makes remote sensing image 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]–[12]. The success of feature learning is mainly attributed to two factors. First, it uses deep hierarchical network architectures. In 2012, 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; 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 is 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]–[20], autoencoder [21]–[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 invariable enough for RSI understanding tasks. With the development of convolutional neural networks (CNNs), many supervised feature learning 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 supervised feature learning (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-supervised feature learning (SSFL), was proposed in the field of natural language process (NLP) [32]–[35] and computer vision (CV) [36]–[38]. It uses human-designed task-agnostic self-supervised learning signals to generate pseudo-labels for massive unlabeled data, thereby replacing human labels to guide the model learning.

Up to now, several surveys on SSFL in computer vision have been done [39]–[42]. To bridge the gap between the progress of SSFL in computer vision 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, self-supervised feature learning 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 clarify 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 pre-training data affect the performance of the learned features in downstream tasks. In Section V, we suggest four potential research directions of SSFL in remote sensing. Some conclusions are drawn in Section VI.

II. A 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: unsupervised feature learning (USFL), supervised feature learning (SFL), and self-supervised feature learning

(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:

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

where $f_\theta(\cdot)$ is the model used for feature learning and θ is the parameters needed to be learned in the model. $f_\theta(\cdot)$ can be a deep convolutional model, auto-encoder model, Gaussian probability distribution model, etc.; 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 ith 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:

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

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 pseudo label $\widetilde{y_i}$ is constructed using the data intrinsic structure as a generic supervised learning signal. Thus, USFL can be defined as follows:

$$\min_\theta \mathcal{L}(f_\theta(\cdot), \{(x_i, \widetilde{Y_i}) | x_i \in X, \widetilde{Y_i} \in \widetilde{Y}\}_{i=0}^N, S). \tag{3}$$

where $S : X \to \widetilde{Y}$. 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 $D = \{(x_i, y_i) | x_i \in X, y_i \in Y\}_{i=0}^{N}$, where x_i denoting the ith 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:

$$\min_{\theta} \mathcal{L}(f_\theta(\cdot), \{(x_i, y_i) | x_i \in X, y_i \in Y\}_{i=0}^{N}, S). \quad (4)$$

where $S : X \to Y$. The deep convolutional neural network (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 remote sensing image 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]–[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]–[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 pseudo-labels for massive unlabeled data, thereby replacing human labels to guide the model learning. Specifically, given an unlabeled dataset $D = \{x_i | x_i \in X\}_{i=0}^{N}$, for each x_i in X, a pseudo label $\widetilde{y_i}$ is generated by human-designed learning signals. Thus, SSFL can also be defined by the unified feature learning framework as follows:

$$\min_{\theta} \mathcal{L}(f_\theta(\cdot), \{(x_i, \widetilde{y_i}) | x_i \in X, \widetilde{y_i} \in \widetilde{Y}\}_{i=0}^{N}, S). \quad (5)$$

where $S : X \to \widetilde{Y}$. Although the USFL method and SSFL both use pseudo-labels for feature learning, the former constructs pseudo-labels using only the manifold hypothesis, but the latter use diverse pretext tasks (e.g., image inpainting, colorization, 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 natural language processing (NLP) [52]–[54] and computer vision (CV) [55]–[57]. We believe SSFL is effective and robust for RSIs understanding tasks when the labeled data are insufficient. The reasons are:

- 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].
- 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-SUPERVISED LEARNING ON REMOTE SENSING DATA

Based on the unified framework of SSFL (Sec. II-D), in this section, we review research on SSFL in the remote sensing

TABLE I
THE MANUALLY CONSTRUCTED LARGE-SCALE REMOTE SENSING IMAGE DATASETS SUITABLE FOR SELF-SUPERVISED LEARNING.

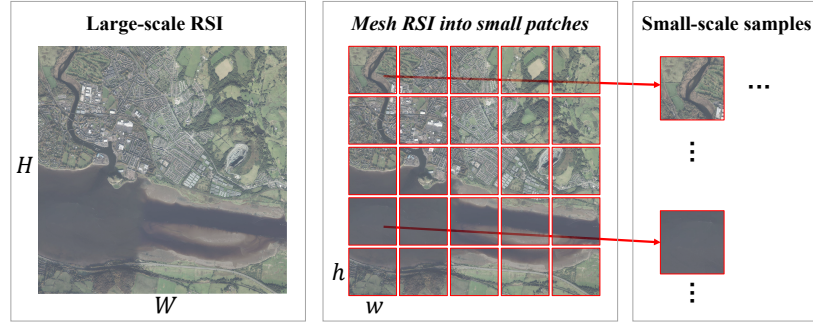| Dataset	Year	Categories	#Num.	Image size	Resolution (m)	Image Type	Image source
RSD46-WHU	2017	46	117,000	256	$0.5 \sim 2$	RGB	Google Earth, Tianditu
fMoW	2018	63	1,047,691	-	-	Multispectral	Digital Globe
Million-AID	2021	51	1,000,848	$110 \sim 31,672$	$0.5 \sim 153$	RGB	Google Earth
BigEarthNet	2019	19	590,326	$20 \sim 120$	$10 \sim 60$	Multispectral	Sentinel-2
SEN12MS	2019	33	180,662	256	10	SAR-Multispectral	Sentinel-1, Sentinel-2
SeasoNet	2022	33	1,759,830	$20 \sim 120$	$10 \sim 60$	Multispectral	Sentinel-2

#Num. is the number of samples

Fig. 1. Illustration of the traditional grid sampling method.

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; 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 multi-temporal and multi-modal 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-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, 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 multi-modal RS data features.

Self-supervised learning does not require manual annotation, so the above 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 computer vision, which is as large as hundreds of millions. Therefore, there is an urgent need to build larger datasets for SSFL.

TABLE II
REMOTE SENSING IMAGE DATASET CONSTRUCTED BY AUTOMATED SAMPLING.

Dataset	Year	#Num.	Image size	Resolution (m)	Image Type	Image source
SoundingEarth	2021	50,545	1024	0.2	RGB - Audio	Google Earth
SeCo	2021	1,000,000	-	$10 \sim 60$	Multispectral	Sentinel-2
TOV-RS-Balanced	2022	500,000	600	$1 \sim 20$	RGB	Google Earth
SSL4EO-S12	2022	3,012,948	$20 \sim 120$	$10 \sim 60$	SAR - Multispectral	Sentinel-1, Sentinel-2

#Num. is the number of samples

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 ultra-large-scale RSI datasets.

Grid sampling is a traditional way to automatically collect RSI samples. As shown in Figure 1, the grid sampling method meshes a large-scale RSI of size $H \times W$ into $\frac{H}{h} \times \frac{W}{w}$ non-overlapped 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 2,000 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 spatio-temporal range, and contains multi-modal 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 Aporee:::Maps[1]; 2) Collecting the corresponding RSIs from Google Earth platform based on the geo-coordinates of the audio. SoundingEarth contains about 3,500 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 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 3 million RSIs, called the TOV-RS dataset, with a balanced number of natural and man-made samples. Further, to ensure a balanced sampling of the subcategories of natural and man-made categories, a class-balanced resampling strategy is proposed to post-process 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 pseudo labels, 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 Figure 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 down-sampling x to obtain a destroyed version $\widetilde{x}$.

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

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:

- **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
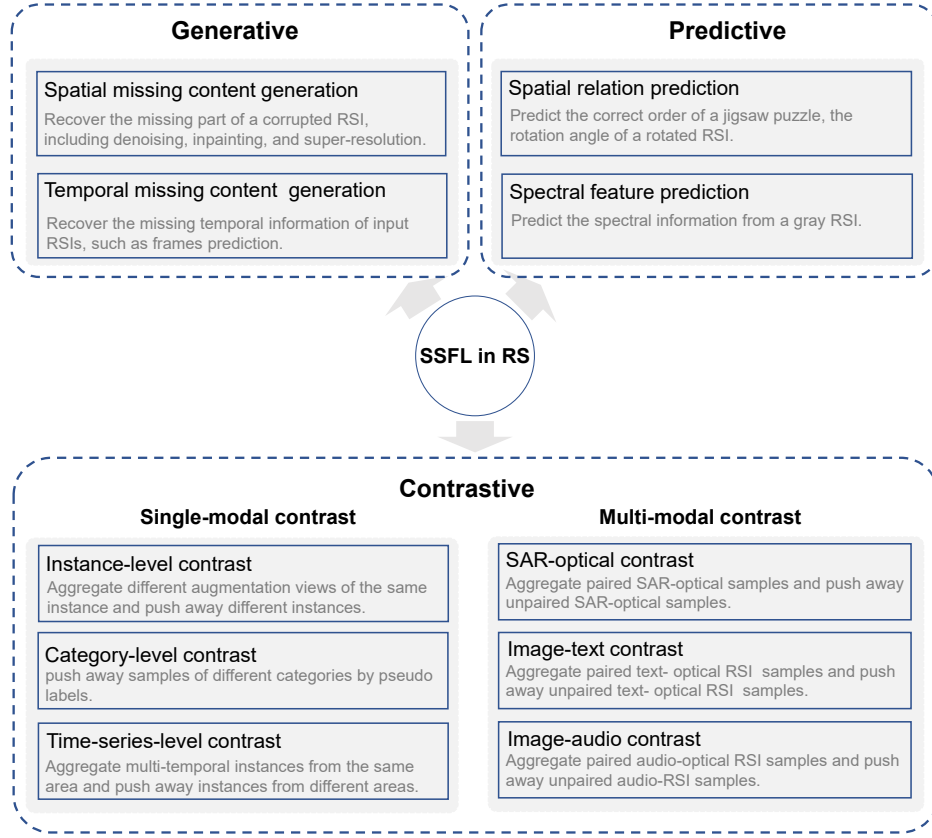
---

[1] https://archive.org/details/radio-aporee-maps

Generative

Spatial missing content generation
Recover the missing part of a corrupted RSI, including denoising, inpainting, and super-resolution.

Temporal missing content generation
Recover the missing temporal information of input RSIs, such as frames prediction.

Predictive

Spatial relation prediction
Predict the correct order of a jigsaw puzzle, the rotation angle of a rotated RSI.

Spectral feature prediction
Predict the spectral information from a gray RSI.

SSFL in RS

Contrastive

Single-modal contrast

Instance-level contrast
Aggregate different augmentation views of the same instance and push away different instances.

Category-level contrast
push away samples of different categories by pseudo labels.

Time-series-level contrast
Aggregate multi-temporal instances from the same area and push away instances from different areas.

Multi-modal contrast

SAR-optical contrast
Aggregate paired SAR-optical samples and push away unpaired SAR-optical samples.

Image-text contrast
Aggregate paired text- optical RSI samples and push away unpaired text- optical RSI samples.

Image-audio contrast
Aggregate paired audio-optical RSI samples and push away unpaired audio-RSI samples.

Fig. 2. Categories of the self-supervised feature learning signal in the remote sensing field: generative, predictive, and contrastive.

RSIs [76]–[80]. For example, Zhang et al. [81] employ a stacked auto-encoder 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.

- **Inpainting**: Pathak at el. further developed a similar learning signal using a more challenging inpainting task [82], 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]–[85]. Singh et al. [86] 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 auto-encoder (MAE) method makes the inpainting task more challenging by masking more than 75% of the image [87], thus stimulating the learned transformer model [88], [89] to capture key features of images. Most recently, MAE-based SSFL methods have received much attention for their power of generic feature learning [90]–[94]. Sun
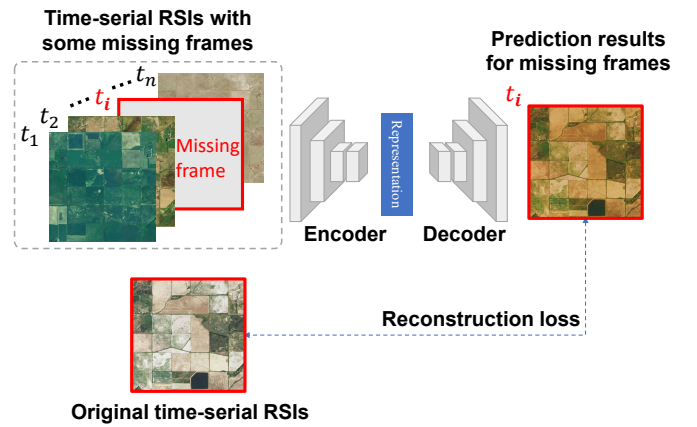

Fig. 3. Illustration of the temporal missing content generation.

et al. [95] applied the MAE method to SSFL for a 3 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.

- **Super-resolution**: The super-resolution-based SSFL signal follows the above 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 [96]–[98]. For paired hyper-
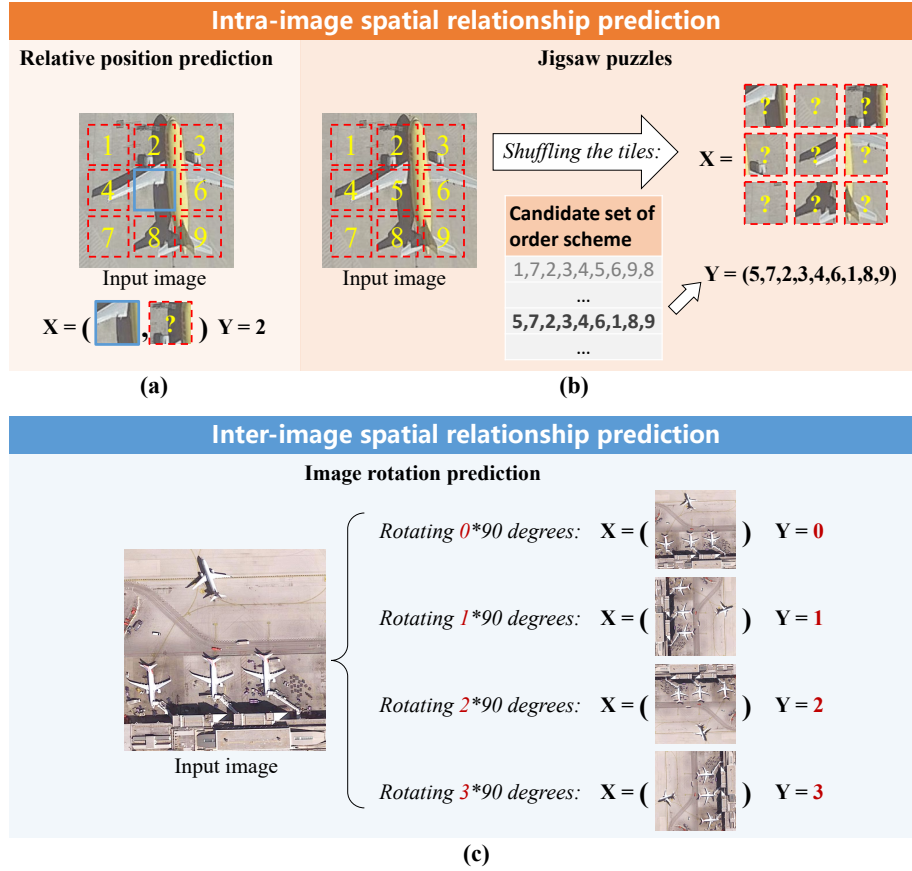
Fig. 4. Illustration of three types of spatial relationship prediction tasks: (a) relative position prediction, (b) jigsaw puzzle, and (c) image rotation angle prediction. (a) and (b) aim at predicting intra-image spatial relationships. (c) aims at predicting inter-image spatial relationships.

spectral and multi-spectral images, Vedaldi et al. [99] 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 performance 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 [100]–[107]. For example, as shown in Figure 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 [105] 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. [107] proposed a patch-level generative SSFL method that randomly masks some frames from the time-series Sentinel-2 images completely rather than adding only pixel-level noise. Using the same datasets as [105], their experiments show that the method can further improve the performance of the pre-trained transformer in crop classification.

2) Predictive learning signals: Unlike the generative learning signals that deal with pixel-level details, the predictive learning signal focuses on learning semantics context features. Such learning signals can be divided into two categories: learning signals based on spatial relation prediction and that based on spectral feature prediction.

a) Learning signals based on spatial relation prediction: The application of this kind of signal in SSFL is based on the assumption that the spatial relation information between object parts is correlated with object semantics. For example, an airplane is composed of many parts including wings, a fuselage, tailplane in a fixed spatial combination. This type of signal is constructed in the following common ways:

- **Relative position prediction**: Doersch et al., [108] forced the SSFL model to predict the relative positions of two image tiles cropping from the input image. As shown in Figure 4(a), one tile is in the middle and the other tile is randomly located at any position around the former. Their method is based on a hypothesis that the model can understand what the object is in order to accurately predict the spatial relations between the parts of the object.

- **Jigsaw puzzle**: Based on the same hypothesis [108], the

jigsaw puzzles task [109]–[115] is also used to construct the learning signal for SSFL. For example, Noroozi and Favaro [109] constructed a jigsaw puzzle prediction task by randomly shuffling tiles of an image as shown in Figure 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 [108]–[110] in learning the desired spatial context features for the learning signals constructed in the ways described above: i) Shortcuts: Texture and color continuity between image tiles is simple cues that can serve as shortcuts for models to solve the above 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. [111] 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. ii) Task ambiguities. Ambiguity is common in jigsaw puzzles when image tiles lack valid content or are highly similar to each other [109], [110]. 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 sea, 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 intra-image spatial relationships, inter-image spatial relationships can also be exploited to construct the learning signal:

- **Image rotation prediction**: Using the image rotation prediction task to construct the signal [116]–[120] 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 Figure 4 (c)). The construction process of this learning signal is as follows (Figure 4(c)):
  Step 1: Given a predefined set of discrete rotation transformations $G = \{g(\cdot|y)\}_{y=0}^{K}$, the rotated image $\widetilde{x}$ is obtained by applying a random rotation transformation $g(\cdot|y)$ from G to the input image x. whose corresponding label is y. $g(\cdot|y)$ means rotating the image by $y*90$ degrees.
  Step 2: Using $(\widetilde{x}, y)$ as the supervision for training the SSFL model.
  In the above steps, the definition of G affects the performance of the learned features in downstream tasks. Empirical experiments of [116] found that a small number of rotation angles (e.g., K=2) may lead to insufficient
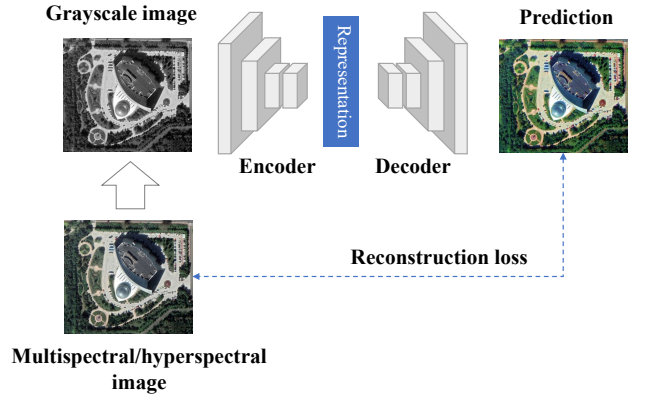

Fig. 5. Illustration of spectral feature prediction.

monitoring information, while a larger K (e.g., K=8) leads to task ambiguity. Zhao et al. [121] performed SSFL on high-resolution optical RSIs by this learning signal and visualized the feature maps of scenes like airports and sea surfaces. The results show that unlike using supervised learning only, the model learned via image rotation prediction will pay more attention to key objects in images which helps to improve the accuracy of RSI scene classification. Some work used this learning signal to exploit high-resolution [122], [123] or hyperspectral RSI data [124], [125] for SSFL, thus benefiting downstream interpretation tasks based on the corresponding data. Further studies [126], [127] have found that rotation-equivariant features [128] varying accordingly with the rotation of the objects (e.g., aircraft, ship, buildings) could be learned by this signal. And Zhang et al. [129] suggested that the rotation-equivariant feature learned from this SSFL signal has better performance in object detection tasks on SAR RSIs than traditional CNNs that focus only on rotation-invariant features. Recent studies [130], [131] 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 the domed buildings, oceans, and dessert.

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 [132]–[136]. 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 (Figure 5):

Step 1: Obtaining gray-scale image $\widetilde{x}$ from an input multi-spectral/hyperspectral image x.

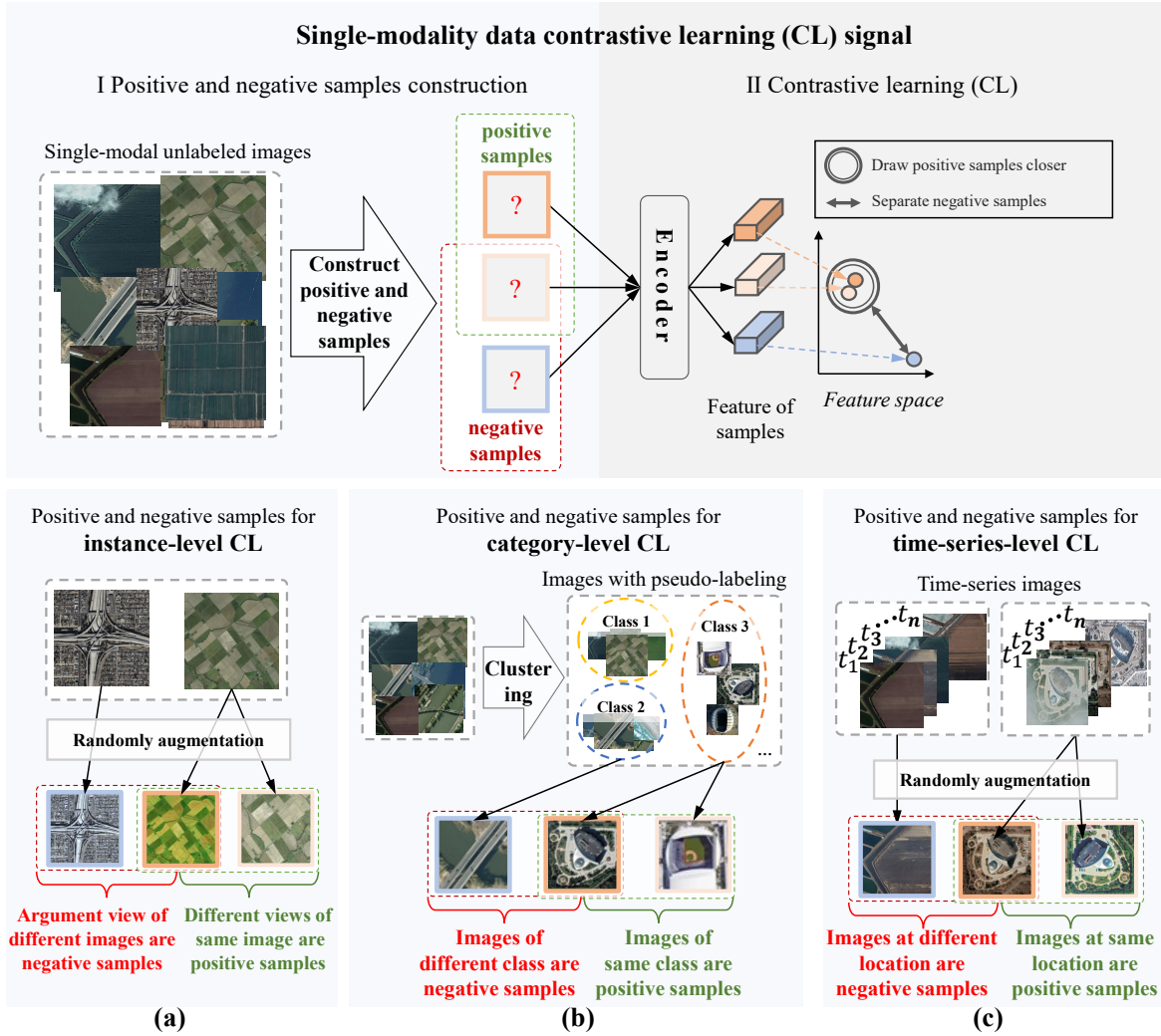Step 2: Training a model $f(\cdot)$ with encoder-decoder struc-

Fig. 6. Illustration of three kinds of single-modal contrastive learning signals: instance-level, category-level, and time-series-level.

ture by minimizing the object function $||f(\tilde{x}) - x||_2^2$.

For RSI with only RGB channels, the above process can be regarded as a colorization task, i.e., predicting the RGB color channels using the gray-scale channels. Zhang et al. [132] 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-B2a. They suggest constructing the colorization task in Lab 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 inter-channel correlation as a shortcut. Empirical experiments [137] show that the learned features using Lab 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 remote sensing images, 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 un-

derstanding the semantics of the image. Therefore, Zhang et al. [132] 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. [133] constructed a learning signal using spectral histogram prediction tasks instead of spectral value regression. For multispectral RSIs, Vincenzi et al. [138] suggested training the model by predicting color channels based on multi-spectral channels not including RGB instead of gray-scale channels, to fully exploit the diverse features of each band.

3) Contrastive learning signals: Extensive experimental work in psychology has shown that infants learn perceptual categories primarily through observation rather than linguistic supervision [139]. In this way, they can summarize different manifestations of the same object (i.e., invariance) and distinguish objects by different manifestations (i.e., distinguishability). 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 [139], [140]. The

contrastive learning signals are constructed by the following two steps:

Step 1: Given an unlabeled training set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, each sample $\boldsymbol{x}_i$ is augmented by $T(\cdot)$ to create a pair of positive samples $(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2)$. In contrast, any two augmented views of different samples are treated as a negative sample pair $(\boldsymbol{x}_i^1, \boldsymbol{x}_j^2)$. 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 Eq. (6), 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 multi-modal 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 [141]–[143]. 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 Figure 6.

Instance-level contrastive learning signal [144]–[152] 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. [153] 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 [142], [152]. For negative sample construction, they found that a sufficiently large number of negative samples is 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. [154] 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. [155] 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 [156]–[159], such as BYOL [157] and SimSiam [156], 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 [160].

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. [161] 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 the negative sample. Following the above paradigm, similar works are proposed for learning features from unlabeled RSIs for scene classification [162], [163]. In addition, considering the gap between instance-level contrastive learning and pixel-level segmentation, Li et al. [164] constructed a global style and local matching contrastive learning network (GLCNet) to extend instance-level contrast 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 [165], [166] pays attention to the construction of fine contrast-level learning signals, e.g., pixel-level. For example, Muhtar et al. [167] 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

$$\mathcal{L} = -\mathrm{E}_X \left[\log \frac{\text{sim}\left(f\left(\boldsymbol{x}_i^1\right)^T f\left(\boldsymbol{x}_i^2\right) \right)}{\text{sim}\left(f\left(\boldsymbol{x}_i^1\right)^T f\left(\boldsymbol{x}_i^2\right) \right) + \sum_{j=1}^{n-1} \text{sim}\left(f\left(\boldsymbol{x}_i^1\right)^T f\left(\boldsymbol{x}_j^2\right) \right)} \right] \tag{6}$$

suitable for semantic segmentation tasks. Similarly, scholars [168] in computer vision 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 [169]–[181], 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 contrast learning [182]. To address this problem, category-level contrastive learning signal only takes images of different "categories" as negative samples [183]–[187] 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 (Figure 6(b)):

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

Step 2: According to pseudo-labels, 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. [188] proposed Deep-Cluster train the SSFL model by iterating the above two steps. Studies [189]–[191] 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 [192]. 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 [188], [190] add constraints to the clustering process to make the sample size of each category as balanced as possible. Besides, during training, the above 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., [193] proposed an online clustering-based method, SwAV, that combines the two processes of clustering and cluster assignments prediction into a classification task. Further studies [194]–[198] found that joint category-level and instance-level contrastive learning (i.e., multi-granular contrastive learning) has higher generalizability than single-granular contrastive learning on diverse downstream tasks. The main reason is that various downstream tasks often require multi-granular 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 sub-categories of aircraft, such as Boeing 737 and Boeing 747. Therefore, multi-granular contrastive learning has become a focus of the studies relating to general-purpose SSFL. Representative approaches of multi-granular contrastive

learning [199]–[203] include PCL [196] and Mugs [198].

Time-series-level contrastive learning signal aims at learning temporal-invariant features of RSIs [204]–[210]. 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 multi-temporal 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. [205] 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, 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 spatio-temporal distributions, Huang et al. simulate RSIs as different spatio-temporal visual styles and used them as positive samples by introducing an optimal transmission mechanism [211]. In this way, the SSFL model can obtain features with both distinguishability and spatio-temporal invariance and thus better generalize to unseen RSI scene classification datasets.

*b) **Multi-modal contrastive learning**:* We can get multi-modal 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. [212] 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 multi-modal contrastive learning method, which takes multi-model data of the same scene as positive samples and data of different scenes as negative ones, has attracted great attention [213]–[219]. As shown in Figure 7, the method has been applied to learn the cross-modal features of different types of RSIs (e.g., SAR-optical RSIs contrastive learning [220]–[228]) or
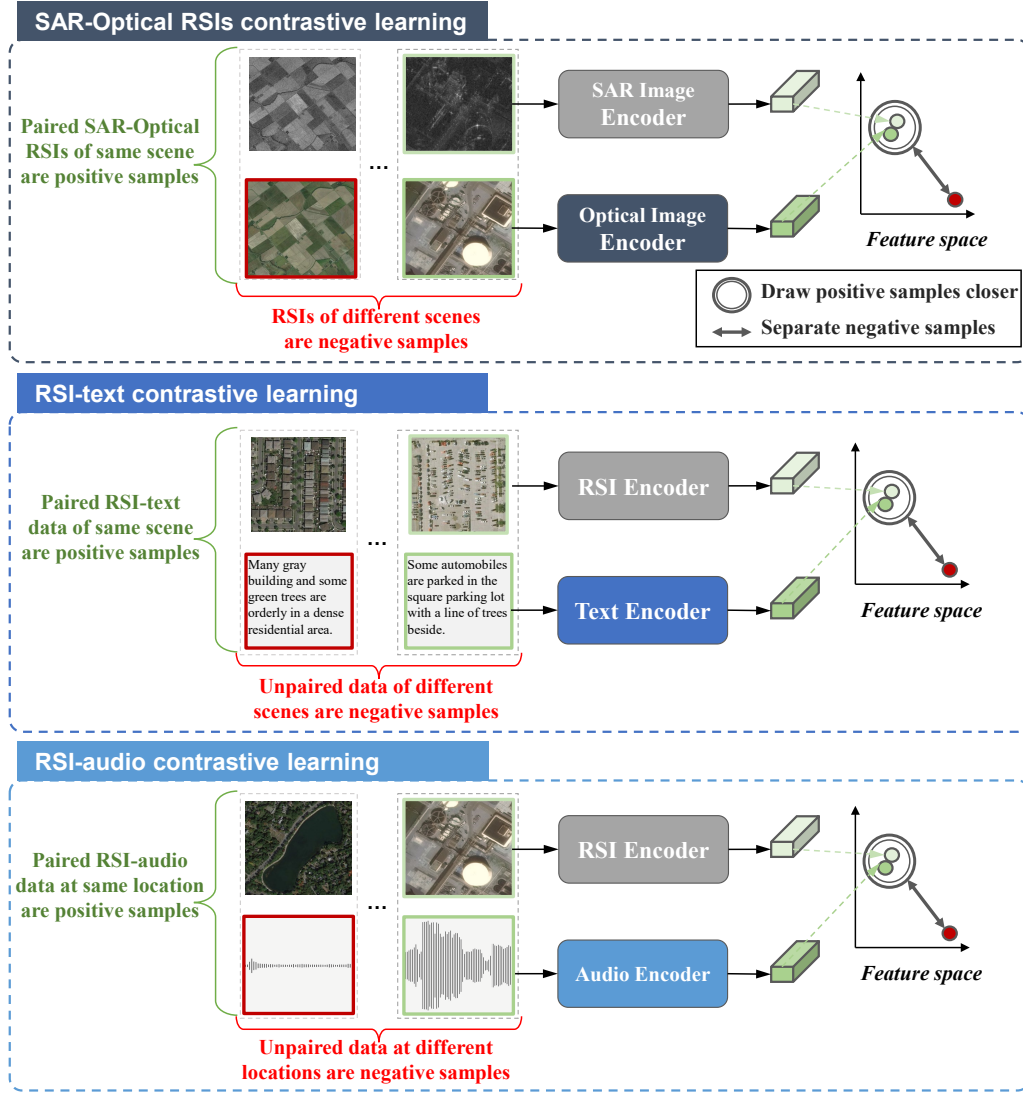
Fig. 7. Illustration of three kinds of multi-modal data contrastive learning signals: SAR-optical, RSI-text, and RSI-audio.

the combinations of RSIs and other types of data (e.g., RSI-text, and RSI-audio contrastive learning [71], [229], [230]).

- **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 multi-modal RSI contrastive learning. Chen and Bruzzone [224] 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 multi-modal 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 [225] compared the performance of multi-modal (SAR-optical) and single-modal (SAR or optical) SSFL models in three land-cover classification tasks and found that the multi-modal contrastive learning approach outperforms single-modal contrastive learning approaches.

- **RSI-text contrastive learning**: The research of contrastive learning based on image-text data develops rapidly in the field of computer vision [231]–[234], 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 Pre-training (CLIP) [38] constructed a large-scale Internet crawler dataset containing 400 million image-text sample pairs for multi-modal 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 pre-training, CLIP feeds image and text samples into the image encoder and text encoder for independent encod-

ing, 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 pre-training, 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 finetuning, the multi-modal 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. So far, the only work is done by Mikriukov et al [230], which used two publicly available RSI captions datasets[2], UC Merced Land Use and RSICD. The results of both two datasets showed that the model trained by the learning signal can accurately retrieve the corresponding RSIs based on the input text caption. However, the text descriptions of RSIs in these two datasets are obtained by manual annotation, so this paradigm is hard to be expended. Therefore, it is worth studying how to obtain geo-tagged textual information of RSIs from crowed-sourced data such as OpenStreetMap and Twitter at a low cost.

- **RSI-audio contrastive learning**: Multi-modal contrastive learning based on image-audio has also attracted the attention of scholars [214], [215], [235], [236]. In an open-source project called Radio Aporee:::Maps[3], Heidler et al. [71] constructed a multi-modal 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 multi-modal SSFL for RSI-audio feature learning. Experiments demonstrate that introducing 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 remote sensing images 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

[2]https://github.com/201528014227051/RSICD_optimal
[3]https://archive.org/details/radio-aporee-maps

Kernel Visualization [237], Feature Map Visualization [238]–[240], and T-SNE [241] 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 attentions?

The literature [194], 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 supervised feature learning method. In addition, those features form more separated clusters, indicating that those features have strong category distinguishability. The literature [121] used a similar visual analysis method to assess the quality of the features obtained by the jointly supervised feature learning method and the self-supervised feature learning method from a remote sensing scene classification dataset. The results showed that combining two feature learning paradigms can better capture the details that represent the semantic features in the images with complex backgrounds, and achieve higher accuracy in classifying some difficult objects than the baseline model does.

2) Quantitative evaluation methods: At present, the self-supervised learned features are usually assessed quantitatively by downstream tasks. Specifically, the features obtained by SSFL are used as pre-trained model parameters and are transferred to downstream tasks (e.g., remote sensing scene classification, semantic segmentation, and target detection). Then their performance in downstream tasks is evaluated and the results are taken as the assessment results of the features. The commonly used transfer methods are linear probes and fine-tuning.

The process of the linear probe method is: 1) the network parameters obtained by self-supervised learning are fixed and a linear classifier is added to the last layer of the network. 2) The linear classifier is trained using downstream labeled data to evaluate the performance of self-supervised learning features. However, due to the simple classification structure, this method can only evaluate the performance of self-supervised learning features in image-level classification tasks (e.g., remote sensing scene classification). The literature [74] used the linear probe approach to compare the representation ability of the models pre-trained by SSFL methods for three popular RSIs scene classification datasets. The experiments showed that the features learned by the instance-level contrastive SSFL method have better performance in scene classification tasks than the methods based on image inpainting or predicting relative position.

The fine-tuning approach uses the model parameters pre-trained by SSFL methods as an initialization of the task-
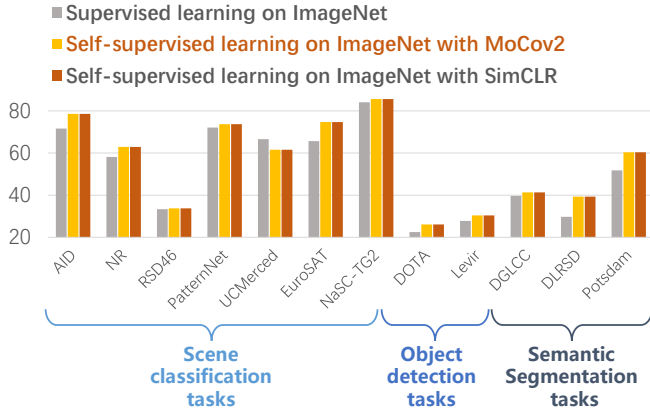
Fig. 8. Quantitative evaluation of self-supervised and supervised learned features on multiple downstream tasks by fine-tuning approach. To evaluate the performance of the features learned by different methods, we adopt three types of RSI understanding tasks, including scene classification, object detection, and semantic segmentation. Scene classification tasks used AID, NR, RSD46, PatternNet, UC Merced, EuroSAT, and NaSC-TG2. Object detection tasks used DOTA and LEVIR. Semantic segmentation tasks include DGLCC, DLRSD, and Potsdam.

specific model's backbone, then transfers the learned features to various RSIU tasks by adding task-specific adapters following the backbone. Since the features learned by the SSFL method can be adapted to different downstream tasks by simply modifying the task-specific adapters, this feature transfer approach enables a more comprehensive feature representation capability assessment. Tao et al. used the fine-tuning approach to test the representation capability of the features learned by the SSFL method on ImageNet [72], which used 12 publicly available datasets for the downstream tasks of remote sensing scene classification, object detection, and semantic segmentation. The results (Figure 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 [242] 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. THE KEY FACTORS INFLUENCING SELF-SUPERVISED FEATURE LEARNING

In the past two years, SSFL methods have developed rapidly in the field of remote sensing, but there lacks a comprehensive study on the key factors that would influence SSFL. Therefore, we analyze how the SSFL signal and properties of pre-training data affect the performance of the learned features in downstream tasks in Section IV-A and IV-B, respectively.

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. Therefore, the choice of SSFL signal is crucial, as it determines what features can be learned by pre-training 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.

1) Experiment Setup: Corresponding to the generative, predictive, and contrastive learning signals, we choose the following three SSFL methods for comparison:

i. Inpainting [82]. The model learns features by recovering the manually masked parts.

ii. Image rotation prediction [116]. The model learns features by predicting the rotation angle of the input images.

iii. Instance-level contrastive learning [153]. First, the argument views of the same sample are regarded as positive instances and that of different samples in a training batch are regarded as negative instances. Then, the model learns features by enhancing the similarity between positive instances and the difference between negative instances.

a) Pre-training datasets: We pre-train the above three SSFL methods using two large-scale RSI datasets, Million-AID [15], [64] and TOV-RS [72], separately. 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. We use these samples without labels for pre-training. TOV-RS dataset is an automatically collected unlabeled dataset containing 3 million samples.

b) 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.

i. Dataset for scene classification tasks:
- RSD46-WHU [62], [243], a large-scale scene classification dataset containing 117,000 samples of 46 categories. The images are collected from Google Earth and Tianditu.
- EuroSAT [242], a scene classification dataset containing 27,000 Sentinel-2 images of 10 categories.

ii. Dataset for object detection tasks:
- DOTA v1.0 [9], [244], 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 satellite.

TABLE III
INFORMATION OF THE TRAINING AND TESTING SETS OF THE SIX DATASETS FOR THE DOWNSTREAM TASKS IN THE EXPERIMENTS.

Dataset	RSD46-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

- LEVIR [245], is 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.

iii. Dataset for semantic segmentation tasks:

- Potsdam [246], is 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 6000x6000. This dataset is annotated in 6 classes.
- DGLCC [247], a land cover classification dataset containing 1,146 images annotated in 7 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 km^2.

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

c) Implementation details: We pre-train the ResNet50 [248] model for 400 epochs using the three SSFL methods on two pre-training datasets separately and then evaluate the six pre-trained 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 pre-trained 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 assessed by Cohen's kappa coefficient (kappa), and that in object detection is assessed by the mean average precision (mAP).

2) Experiment results: Below are our findings:

- **Contrastive learning is an optimal choice because the learned features are superior in most downstream tasks**. As shown in Figure 9(a), when pre-trained 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 pre-trained by TOV-RS (Figure 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.
- **The choice of SSFL signal should consider the downstream tasks**. For the two object detection tasks, as shown in Figure 9(b), the image rotation prediction method achieves higher scores than the instance-
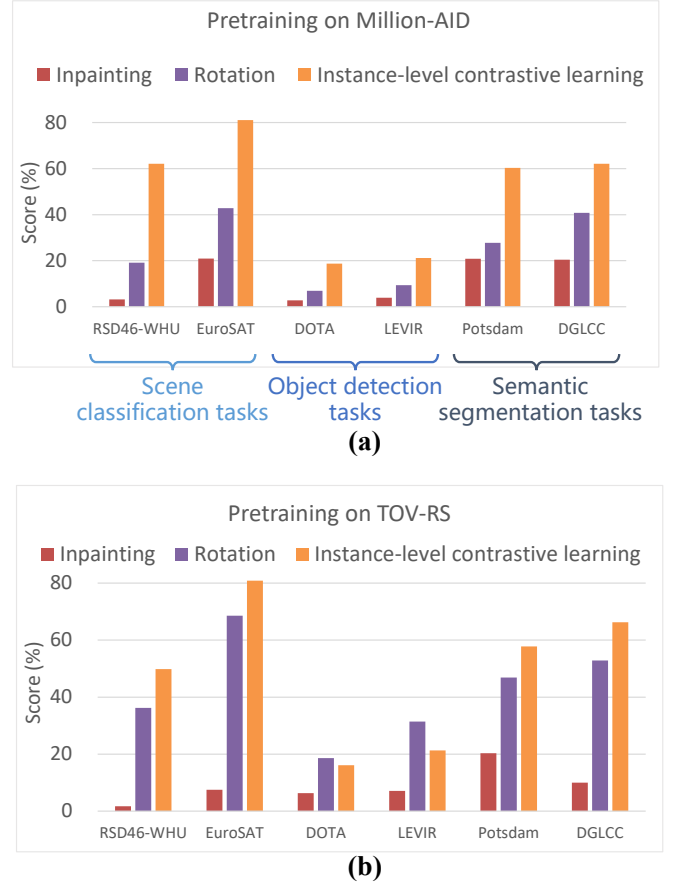

Fig. 9. Performance of the ResNet50 model pre-trained by three self-supervised feature learning signals using (a) Million-AID and (b) TOV-RS. The performance using RSD46-WHU, EuroSAT, Potsdam, and DGLCC is assessed by the Kappa value, and that using DOTA and LEVIR is evaluated by the mAP value.

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-B3a.

B. Effects of the pre-training 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
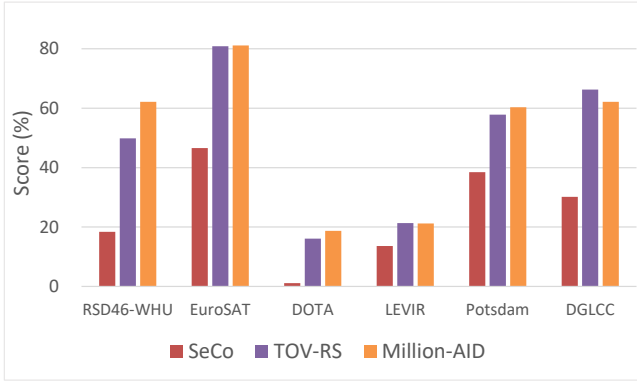
Fig. 10. Results for the data of different spatial resolutions. RSD46-WHU, EuroSAT, Potsdam, and DGLCC are assessed by Kappa. DOTA and LEVIR are assessed by mAP.

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 contrastive learning for feature learning, as its superiority has been demonstrated in the above 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 pre-training. SeCo has a spatial resolution of up to 10 m/pixel, and the TOV-RS and Million-AID have resolutions of up to 1 m and 0.5 m, respectively. To evaluate the performance of the features learned by the model pre-trained by different datasets, we use the six downstream datasets described in Section IV-A1 for downstream tasks.

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

- **Spatial resolution of the pretraining dataset is critical for SSFL**. For all three kinds of downstream tasks, the performance of the model pre-trained using the two high-resolution datasets (TOV-RS and Million-AID) is significantly higher than that of the model trained by the medium-low resolution dataset (SeCo). The reason may be that the high-resolution RSIs dataset can additionally provide more textural and geometric details of the ground objects than the low-resolution dataset, thus the learned features are more distinguishable for various objects and scenes.
- **The impact of the pre-training data resolution on self-supervised feature learning may be greater than that of the domain gap between the pre-training and downstream datasets**. For example, the EuroSAT dataset and the SeCo are constructed by the RSIs from 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 Figure 10, models pre-trained on TOV-RS and Million-AID achieve much higher performance scores than the model pre-
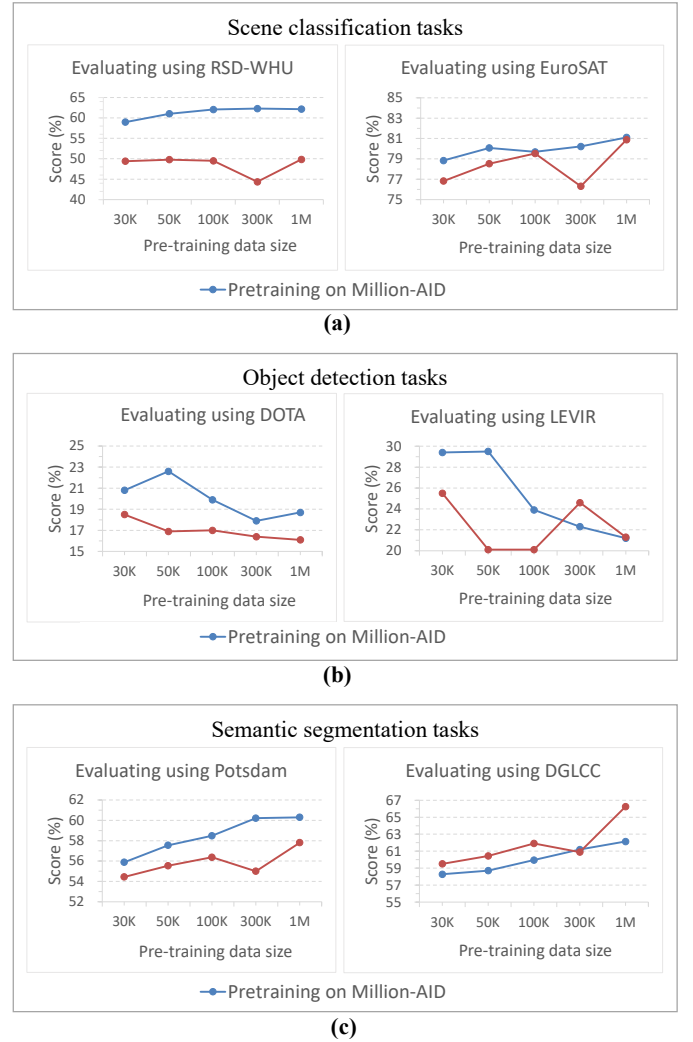

Fig. 11. To evaluate the performance of the features learned from pre-training data of different sizes, we use three types of downstream tasks, including (a) scene classification tasks, (b) object detection tasks, and (c) semantic segmentation tasks.

trained on SeCo. For EuroSAT, the kappa of the TOV-RS pre-training and Million-AID pre-training models are 0.8089 and 0.8111, respectively, while the kappa of the SeCo pre-training model was only 0.4658. This suggests that it is unnecessary to sacrifice the resolution of the pre-training data for reducing the domain gap between the pre-training 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 pre-training 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 pre-training data sizes are shown in Figure 11. Below are our findings:

- Enlarging the pre-training data size improves the learned features' performance in most downstream tasks. In Fig-

ure 11 (a) and (c), as the pre-training 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 pre-trained on 1M Million-AID samples outperforms the model pre-trained on 30K Million-AID 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 Figure 11 (b), the performance of the model shows irregular fluctuations as the pre-training data size increases. This confirms the findings in Section IV-A2 that the features learned by instance-level contrastive learning that aims at discriminating RSI scenes are unsuitable for object detection tasks.

- Little benefit over 300,000 pre-trained samples. The performance of the learned features in these tasks seems to saturate before 1M. For example, as shown in Figure 11(a) and (c), when pre-trained 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. 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 supervised feature learning 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 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 self-supervised feature learning 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 huge amount 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, multi-spectral, SAR, 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 multi-modal SSFL have been developed in the field of computer vision [38], [214], [218], most of these methods are based on the assumption that the data from different modalities are precisely aligned. So, they are seldomly 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 [249]. Therefore, for the unaligned data without annotation, how to construct continual feature learning paradigms for multi-modal data and learn the intrinsic relationship between different modalities are worth researching in the future.

D. The 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 mean average precision. So it is difficult to directly aggregate them statistically (such as averaging) to form a comprehensive evaluation indicator. Therefore, the construction of a general

VI. Conclusions

In this survey, we provide 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 comparative study to analyze the impacts of training data selection and the choice of SSFL signals on self-supervised remote sensing feature learning. 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

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 Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, p. 22–40, 2016.

[2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, p. 8–36, 2017.

[3] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, and J. Wang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote sensing of environment*, vol. 241, p. 111716, 2020.

[4] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *Journal of the American Mathematical Society*, vol. 29, no. 4, p. 983–1049, 2016.

[5] H. Cui, G. Zhang, T.-Y. 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 Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, p. 4993–5007, 2021.

[6] H. Wang, C. Tao, J. Qi, R. Xiao, and H. Li, "Avoiding negative transfer for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–15, 2022.

[7] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, p. 3965–3981, 2017.

[8] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, p. 1865–1883, 2017.

[9] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, p. 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 Sensing of Environment*, vol. 270, p. 112830, 2022.

[11] C. Tao, W. Lu, J. Qi, and H. Wang, "Spatial information considered network for scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 6, p. 984–988, 2021.

[12] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of image classification algorithms based on convolutional neural networks," *Remote Sensing*, vol. 13, no. 22, p. 4712, 2021.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, p. 1798–1828, 2013.

[14] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017, p. 843–852.

[15] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 4205–4230, 2021.

[16] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, p. 439–451, 2014.

[17] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, p. 173–176, 2011.

[18] D. Ratha, A. Bhattacharya, and A. C. Frery, "Unsupervised classification of polsar data using a scattering similarity measure derived from a geodesic distance," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, p. 151–155, 2018.

[19] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, p. 1349–1362, 2016.

[20] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International Journal of Remote Sensing*, vol. 33, no. 8, p. 2395–2412, 2012.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, p. 3371–3408, 2010.

[22] T. Chao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyprspectral imagery classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, 2015.

[23] Y. Tao, M. Xu, F. Zhang, B. Du, and L. Zhang, "Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, p. 6805–6823, 2017.

[24] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, p. 2693–2705, 2017.

[25] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, p. 391–406, 2018.

[26] S. Ozkan, B. Kaya, and G. B. Akar, "Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, p. 482–496, 2019.

[27] F. F. Tomenotti, L. Luppino, M. Hansen, G. Moser, and S. Anfinsen, "Heterogeneous change detection with self-supervised deep canonically correlated autoencoders," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. Waikoloa, HI, USA: IEEE, 2020, p. 680–683.

[28] Q. Jin, Y. Ma, F. Fan, J. Huang, X. Mei, and J. Ma, "Adversarial autoencoder network for hyperspectral unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–15, 2021.

[29] R. C. Sharma and K. Hara, "Self-supervised learning of satellite-derived vegetation indices for clustering and visualization of vegetation types," *Journal of Imaging*, vol. 7, no. 2, p. 30, 2021.

[30] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, p. 1340–1372, 2022.

[31] X. Q. Lu, X. T. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, p. 5148–5157, 2017.

[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, p. 4171–4186.

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

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, vol. 33. Curran Associates, Inc., 2020, p. 1877–1901.

[36] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 1920–1929.

[37] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, p. 7343–7352.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th international conference on machine learning*, vol. 139. PMLR, 2021, p. 8748–8763.

[39] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, p. 1–1, 2021.

[40] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, p. 4037–4058, 2021.

[41] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, p. 42–62, 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," *arXiv preprint arXiv:2206.13188*, 2022.

[43] P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, S. Liu, C. Li, X. Li, W. Fu, C. Liu, Y. Xu, X. Wang, Q. Cheng, L. Hu, W. Yao, H. Zhang, P. Zhu, Z. Zhao, H. Zhang, Y. Zheng, L. Ji, Y. Zhang, H. Chen, A. Yan, J. Guo, L. Yu, L. Wang, X. Liu, T. Shi, M. Zhu, Y. Chen, G. Yang, P. Tang, B. Xu, C. Giri, N. Clinton, Z. Zhu, J. Chen, and J. Chen, "Finer resolution observation and monitoring of global land cover: first mapping results with landsat tm and etm+ data," *International Journal of Remote Sensing*, vol. 34, no. 7, p. 2607–2654, 2013.

[44] P. Gong, H. Liu, M. Zhang, C. Li, J. Wang, H. Huang, N. Clinton, L. Ji, W. Li, Y. Bai, B. Chen, B. Xu, Z. Zhu, C. Yuan, H. Ping Suen, J. Guo, N. Xu, W. Li, Y. Zhao, J. Yang, C. Yu, X. Wang, H. Fu, L. Yu, I. Dronova, F. Hui, X. Cheng, X. Shi, F. Xiao, Q. Liu, and L. Song, "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," *Science Bulletin*, vol. 64, no. 6, p. 370–373, 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 System Science Data*, vol. 12, no. 2, p. 1217–1243, 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," *Information Sciences*, vol. 250, p. 113–141, 2013.

[47] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, p. 41–57, 2016.

[48] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[49] L. Zhang and X. Gao, "Transfer adaptation learning: A decade survey," *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–22, 2022.

[50] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization in vision: A survey," *arXiv preprint arXiv:2103.02503*, 2021.

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

[52] M. Kayser, O.-M. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, and T. Lukasiewicz, "E-vil: A dataset and benchmark for natural language explanations in vision-language tasks," 2021, p. 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 *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, p. 353–355.

[54] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[55] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, 2019, p. 6390–6399.

[56] A. Islam, C.-F. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, "A broad study on the transferability of visual representations with contrastive learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, p. 8825–8835.

[57] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2020.

[58] W. Steffen, K. Richardson, J. Rockström, H. J. Schellnhuber, O. P. Dube, S. Dutreuil, T. M. Lenton, and J. Lubchenco, "The emergence and evolution of earth system science," *Nature Reviews Earth and Environment*, vol. 1, no. 11, p. 54–63, 2020.

[59] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma, "Self-supervised learning is more robust to dataset imbalance," 2021.

[60] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, and P. Bojanowski, "Self-supervised pretraining of visual features in the wild," *arXiv preprint arXiv:2103.01988*, 2021.

[61] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[62] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, "High-resolution remote sensing image retrieval based on cnns from a dimensional perspective," *Remote Sensing*, vol. 9, no. 7, p. 725, 2017.

[63] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, p. 6172–6180.

[64] Y. Long, G.-S. Xia, L. Zhang, G. Cheng, and D. Li, "Aerial scene parsing: From tile-level scene classification to pixel-wise semantic labeling," *arXiv preprint arXiv:2201.01953*, 2022.

[65] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. Yokohama, Japan: IEEE, 2019, p. 5901–5904.

[66] D. Koßmann, V. Brack, and T. Wilhelm, "Seasonet: A seasonal scene classification, segmentation and retrieval dataset for satellite imagery over germany," *arXiv preprint arXiv.2207.09507*, 2022.

[67] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, p. 153–160, 2019.

[68] C. Jun, Y. Ban, and S. Li, "Open access to earth land-cover map," *Nature*, vol. 514, no. 7523, p. 434–434, 2014.

[69] O. Manas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, p. 9394–9403.

[70] Y. Wang, N. A. A. A. Braham, C. M. Albrecht, Z. Xiong, C. Liu, and X. X. Zhu, "Ssl4eo-s12: A large-scale multimodal multitemporal dataset for self-supervised learning in earth observation," 2022.

[71] K. Heidler, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, "Self-supervised audiovisual representation learning for remote sensing data," *arXiv preprint arXiv:2108.00688*, 2021.

[72] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "Tov: The original vision model for optical remote sensing image understanding via self-supervised learning," *arXiv preprint arXiv:2204.04716*, 2022.

[73] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?" 2021, p. 5414–5423.

[74] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[75] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning - ICML '08*. Helsinki, Finland: ACM Press, 2008, p. 1096–1103.

[76] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Speckle2void: Deep self-supervised sar despeckling with blind-spot convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–17, 2022.

[77] Y. Yuan, J. Guan, and J. Sun, "Blind sar image despeckling using self-supervised dense dilated convolutional neural network," *arXiv preprint arXiv:1908.01608*, 2019.

[78] R. Imamura, T. Itasaka, and M. Okuda, "Self-supervised hyperspectral image restoration using separable image prior," *arXiv preprint arXiv:1907.00651*, 2019.

[79] Y. Qian, H. Zhu, L. Chen, and J. Zhou, "Hyperspectral image restoration with self-supervised learning: A two-stage training approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–17, 2022.

[80] X. Wang, Z. Luo, W. Li, X. Hu, L. Zhang, and Y. Zhong, "A self-supervised denoising network for satellite-airborne-ground hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–16, 2022.

[81] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, p. 24–41, 2016.

[82] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, p. 2536–2544.

[83] W. Li, H. Chen, and Z. Shi, "Semantic segmentation of remote sensing images with self-supervised multitask representation learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 6438–6450, 2021.

[84] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 1–14, 2017.

[85] Z. Xue, X. Yu, A. Yu, B. Liu, P. Zhang, and S. Wu, "Self-supervised feature learning for multimodal remote sensing image land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–15, 2022.

[86] S. Singh, A. Batra, G. Pang, L. Torresani, S. Basu, M. Paluri, and C. V. Jawahar, "Self-supervised feature learning for semantic segmentation of overhead imagery," vol. 1, 2018, p. 4.

[87] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2022, p. 16000–16009.

[88] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[89] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, p. 9992–10002.

[90] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.

[91] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, "Convmae: Masked convolution meets masked autoencoders," *arXiv preprint arXiv.2205.03892*, 2022.

[92] Y. Shi, N. Siddharth, P. Torr, and A. R. Kosiorek, "Adversarial masking for self-supervised learning," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, p. 20026–20040.

[93] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," 2022, p. 14668–14678.

[94] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," 2022, p. 9653–9663.

[95] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, p. 1–1, 2022.

[96] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, p. 105–114.

[97] H. Liu, J. Liu, S. Hou, T. Tao, and J. Han, "Perception consistency ultrasound image super-resolution via self-supervised cyclegan," *Neural Computing and Applications*, 2021.

[98] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, "Self-supervised multi-image super-resolution for push-frame satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, 2021, p. 1121–1131.

[99] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12374. Cham: Springer International Publishing, 2020, p. 208–224.

[100] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *Proceedings of the 32nd international conference on international conference on machine learning - volume 37*. Lille, France: JMLR.org, 2015, p. 843–852.

[101] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Proceedings of the 30th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 64–72.

[102] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2018.

[103] V. Kumar, V. Tripathi, and B. Pant, "Unsupervised learning of visual representations via rotation and future frame prediction for video retrieval," in *Advances in Computing and Data Sciences*, vol. 1440. Cham: Springer International Publishing, 2021, p. 701–710.

[104] B. Peng, Q. Huang, J. Vongkusolkit, S. Gao, D. B. Wright, Z. N. Fang, and Y. Qiang, "Urban flood mapping with bitemporal multispectral imagery via a self-supervised learning framework," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 2001–2016, 2021.

[105] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 474–487, 2021.

[106] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," *arXiv preprint arXiv.2205.09113*, 2022.

[107] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z.-G. Zhou, "Sits-former: A pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102651, 2022.

[108] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, p. 1422–1430.

[109] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 9910. Cham: Springer International Publishing, 2016, p. 69–84.

[110] R. S. Cruz, B. Fernando, A. Cherian, and S. Gould, "Visual permutation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, p. 3100–3114, 2019.

[111] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, p. 793–802.

[112] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, p. 9359–9367.

[113] C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, and A. L. Yuille, "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, p. 1910–1919.

[114] F. Haghighi, M. R. Hosseinzadeh Taher, Z. Zhou, M. B. Gotway, and J. Liang, "Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, p. 137–147.

[115] P. Chen, S. Liu, and J. Jia, "Jigsaw clustering for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, 2021, p. 11521–11530.

[116] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations (ICLR)*, 2018.

[117] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning by rotation feature decoupling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, p. 10356–10366.

[118] T. Chen, X. Zhai, M. Ritter, M. Lucic, and N. Houlsby, "Self-supervised gans via auxiliary rotation loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, p. 12146–12155.

[119] K. Yun, J. Park, and J. Cho, "Robust human pose estimation for rotation via self-supervised learning," *IEEE Access*, vol. 8, p. 32502–32517, 2020.

[120] X. Li, X. Hu, X. Qi, L. Yu, W. Zhao, P.-A. Heng, and L. Xing, "Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, p. 2284–2294, 2021.

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

[122] B. Dang and Y. Li, "Msresnet: Multiscale residual network via self-supervised learning for water-body detection in remote sensing imagery," *Remote Sensing*, vol. 13, no. 16, p. 3122, 2021.

[123] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–13, 2022.

[124] M. E. Paoletti, J. M. Haut, S. K. Roy, and E. M. T. Hendrix, "Rotation equivariant convolutional neural networks for hyperspectral image classification," *IEEE Access*, vol. 8, p. 179575–179591, 2020.

[125] J. Yue, L. Fang, H. Rahmani, and P. Ghamisi, "Self-supervised learning with adaptive distillation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–13, 2022.

[126] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljacic, "Equivariant self-supervised learning: Encouraging equivariance in representations," 2022.

[127] J. Lee, B. Kim, and M. Cho, "Self-supervised equivariant learning for oriented keypoint detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 4847–4857.

[128] G.-J. Qi, L. Zhang, C. W. Chen, and Q. Tian, "Avt: Unsupervised learning of transformation equivariant representations by autoencoding variational transformations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, 2019, p. 8129–8138.

[129] S. Zhang, Z. Wen, Z. Liu, and Q. Pan, "Rotation awareness based self-supervised learning for sar target recognition," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. Yokohama, Japan: IEEE, 2019, p. 1378–1381.

[130] Z. Wen, Z. Liu, S. Zhang, and Q. Pan, "Rotation awareness based self-supervised learning for sar target recognition with limited training samples," *IEEE Transactions on Image Processing*, vol. 30, p. 7266–7279, 2021.

[131] Y. Xu, Z. Cui, W. Guo, Z. Zhang, and W. Yu, "Self-supervised auto-encoding multi-transformations for airplane classification," in *IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium*. Brussels, Belgium: IEEE, 2021, p. 2365–2368.

[132] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 9907. Cham: Springer International Publishing, 2016, p. 649–666.

[133] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 9908. Cham: Springer International Publishing, 2016, p. 577–593.

[134] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," 2017, p. 1058–1067.

[135] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, p. 840–849.

[136] V. Stojnić and V. Risojević, "Analysis of color space quantization in split-brain autoencoder for remote sensing image classification," in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. Belgrade: IEEE, 2018, p. 1–4.

[137] V. Stojnic and V. Risojevic, "Evaluation of split-brain autoencoders for high-resolution remote sensing scene classification," in *2018 International Symposium ELMAR*. Zadar: IEEE, 2018, p. 67–70.

[138] S. Vincenzi, A. Porrello, P. Buzzega, M. Cipriano, P. Fronte, R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara, "The color out of space: learning self-supervised representations for earth observation imagery," in *2020 25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy: IEEE, 2021, p. 3034–3041.

[139] E. Orhan, V. Gupta, and B. M. Lake, "Self-supervised learning through the eyes of a child," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, p. 9960–9971.

[140] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 11, p. 2, 2021.

[141] W. Huang, M. Yi, and X. Zhao, "Towards the generalization of contrastive self-supervised learning," *arXiv preprint arXiv:2111.00743*, 2022.

[142] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 6827–6839.

[143] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, p. 9929–9939.

[144] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. San Diego, CA, USA: IEEE, 2005, p. 539–546.

[145] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 9, p. 207–244, 2009.

[146] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, 2015, p. 815–823.

[147] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[148] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, p. 3733–3742.

[149] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2019.

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

[151] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, p. 22243–22255.

[152] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," vol. 32. Curran Associates, Inc., 2019, p. 15509–15519.

[153] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 119. PMLR, 2020, p. 1597–1607.

[154] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, p. 9726–9735.

[155] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 21798–21809.

[156] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, 2021, p. 15745–15753.

[157] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent a new approach to self-supervised learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 21271–21284.

[158] V. Marsocci and S. Scardapane, "Continual barlow twins: continual self-supervised learning for remote sensing semantic segmentation," *arXiv preprint arXiv:2205.11319*, 2022.

[159] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, p. 12310–12320.

[160] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, p. 10268–10278.

[161] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 3967–3974, 2019.

[162] H. Jung and T. Jeon, "Self-supervised learning with randomised layers for remote sensing," *Electronics Letters*, vol. 57, no. 6, p. 249–251, 2021.

[163] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, p. 2598–2610, 2021.

[164] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–14, 2022.

[165] M. Yang, L. Jiao, F. Liu, B. Hou, S. Yang, Y. Zhang, and J. Wang, "Coarse-to-fine contrastive self-supervised feature learning for land-cover classification in sar images with limited labeled data," *IEEE Transactions on Image Processing*, vol. 1, p. 1–1, 2022.

[166] M. Zhu, J. Fan, Q. Yang, and T. Chen, "Sc-eadnet: A self-supervised contrastive efficient asymmetric dilated network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–17, 2022.

[167] D. Muhtar, X. Zhang, and P. Xiao, "Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–11, 2022.

[168] J. Xie, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Unsupervised object-level representation learning from scene images," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, p. 28864–28876.

[169] C. Agastya, S. Ghebremusse, I. Anderson, C. Reed, H. Vahabi, and A. Todeschini, "Self-supervised contrastive learning for irrigation detection in satellite imagery," *arXiv preprint arXiv:2108.05484*, 2021.

[170] D. Guo, Y. Xia, and X. Luo, "Self-supervised gans with similarity loss for remote sensing image scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 2508–2521, 2021.

[171] 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 Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, p. 9224–9237, 2021.

[172] S. Seneviratne, K. A. Nice, J. S. Wijnands, M. Stevenson, and J. Thompson, "Self-supervision. remote sensing and abstraction: Representation learning across 3 million locations," in *2021 Digital Image Computing: Techniques and Applications (DICTA)*. Gold Coast, Australia: IEEE, 2021, p. 01–08.

[173] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, 2021, p. 1182–1191.

[174] Y. Xu, H. Sun, J. Chen, L. Lei, K. Ji, and G. Kuang, "Adversarial self-supervised learning for robust sar target recognition," *Remote Sensing*, vol. 13, no. 20, p. 4158, 2021.

[175] X. Zheng, B. Kellenberger, R. Gong, I. Hajnsek, and D. Tuia, "Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in uav images," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Montreal, BC, Canada: IEEE, 2021, p. 732–741.

[176] P. Duan, Z. Xie, X. Kang, and S. Li, "Self-supervised learning-based oil spill detection of hyperspectral images," *Science China Technological Sciences*, vol. 65, no. 4, p. 793–801, 2022.

[177] V. Marsocci, S. Scardapane, and N. Komodakis, "Mare: Self-supervised multi-attention resu-net for semantic segmentation in remote sensing," *Remote Sensing*, vol. 13, no. 16, p. 3275, 2021.

[178] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, "Contrastive self-supervised learning with smoothed representation for remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[179] C. Patel, S. Sharma, V. J. Pasquarella, and V. Gulshan, "Evaluating self and semi-supervised methods for remote sensing segmentation tasks," *arXiv preprint arXiv:2111.10079*, 2022.

[180] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," 2022, p. 1422–1431.

[181] D. Wang, C. Zhang, and M. Han, "Fiad net: a fast sar ship detection network based on feature integration attention and self-supervised learning," *International Journal of Remote Sensing*, vol. 43, no. 4, p. 1485–1513, 2022.

[182] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, p. 1134–1143.

[183] M. A. Bautista, A. Sanakoyeu, E. Tikhoncheva, and B. Ommer, "Cliquecnn: Deep unsupervised exemplar learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[184] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," 2016, p. 5147–5156.

[185] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning*. PMLR, 2016, p. 478–487.

[186] T. Li, Y. Cai, Y. Zhang, Z. Cai, and X. Liu, "Deep mutual information subspace clustering network for hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[187] H. Ren, X. Yu, X. Wang, S. Liu, L. Zou, and X. Wang, "Siamese subspace classification network for few-shot sar automatic target recognition," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia: IEEE, 2022, p. 2634–2637.

[188] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11218. Cham: Springer International Publishing, 2018, p. 139–156.

[189] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in

Proceedings of the European Conference on Computer Vision (ECCV), vol. 12355. Cham: Springer International Publishing, 2020, p. 268–285.

[190] A. Ym, R. C, and V. A, "Self-labelling via simultaneous clustering and representation learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[191] C. Zhuang, A. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, 2019, p. 6001–6011.

[192] K. Walter, M. J. Gibson, and A. Sowmya, "Self-supervised remote sensing image retrieval," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. Waikoloa, HI, USA: IEEE, 2020, p. 1683–1686.

[193] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[194] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, p. 9630–9640.

[195] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, p. 8547–8555.

[196] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," 2021.

[197] X. Wang, Z. Liu, and S. X. Yu, "Unsupervised feature learning by cross-level instance-group discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, 2021, p. 12581–12590.

[198] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, and S. Yan, "Mugs: A multi-granular self-supervised learning framework," *arXiv preprint arXiv:2203.14415*, 2022.

[199] Z. Cao, X. Li, Y. Feng, S. Chen, C. Xia, and L. Zhao, "Contrastnet: Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral imagery classification," *Neurocomputing*, vol. 460, p. 71–83, 2021.

[200] X. Hu, T. Li, T. Zhou, and Y. Peng, "Deep spatial-spectral subspace clustering for hyperspectral images based on contrastive learning," *Remote Sensing*, vol. 13, no. 21, p. 4418, 2021.

[201] L. E. C. La Rosa, D. A. B. Oliveira, and P. Ghamisi, "Learning crop type mapping from regional label proportions in large-scale sar and optical imagery," *arXiv preprint arXiv:2208.11607*, 2022.

[202] K. Li, Y. Qin, Q. Ling, Y. Wang, Z. Lin, and W. An, "Self-supervised deep subspace clustering for hyperspectral images with adaptive self-expressive coefficient matrix initialization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 3215–3227, 2021.

[203] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for earth observation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–11, 2022.

[204] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," 2022, p. 8203–8215.

[205] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, p. 10161–10170.

[206] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "Fccdn: Feature constraint network for vhr image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, p. 101–119, 2022.

[207] H. Chen, Y. Zao, L. Liu, S. Chen, and Z. Shi, "Semantic decoupled representation learning for remote sensing image change detection," *arXiv preprint arXiv:2201.05778*, 2022.

[208] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, "Self-supervised representation learning for remote sensing image change detection based on temporal prediction," *Remote Sensing*, vol. 12, no. 11, p. 1868, 2020.

[209] M. Leenstra, D. Marcos, F. Bovolo, and D. Tuia, "Self-supervised pre-training enhances change detection in sentinel-2 imagery," in *Pattern Recognition. ICPR International Workshops and Challenges*, vol. 12667. Cham: Springer International Publishing, 2021, p. 578–590.

[210] J. Pöppelbaum, G. S. Chadha, and A. Schwung, "Contrastive learning based self-supervised time-series analysis," *Applied Soft Computing*, vol. 117, p. 108397, 2022.

[211] H. Huang, Z. Mou, Y. Li, Q. Li, J. Chen, and H. Li, "Spatial-temporal invariant contrastive learning for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[212] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably))," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, p. 10944–10956.

[213] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, p. 24206–24221.

[214] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017, p. 609–617.

[215] H. Lan, Y. Liu, and L. Lin, "Audio-visual contrastive learning for self-supervised action recognition," *arXiv preprint arXiv:2204.13386*, 2022.

[216] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021, p. 2592–2607.

[217] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, p. 9876–9886.

[218] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12356. Cham: Springer International Publishing, 2020, p. 776–794.

[219] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," 2017, p. 1851–1858.

[220] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Self-supervised vision transformers for joint sar-optical representation learning," *arXiv preprint arXiv:2204.05381*, 2022.

[221] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–10, 2022.

[222] A. Montanaro, D. Valsesia, G. Fracastoro, and E. Magli, "Semi-supervised learning for joint sar and multispectral land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[223] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and sar images," *arXiv preprint arXiv:2205.02049*, 2022.

[224] Y. Chen and L. Bruzzone, "Self-supervised change detection in multiview remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–12, 2022.

[225] Y. Chen and L. Bruzzone, "Self-supervised sar-optical data fusion of sentinel-1/-2 images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–11, 2022.

[226] K. Cha, J. Seo, and Y. Choi, "Contrastive multiview coding with electro-optics for sar semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[227] P. Jain, B. Schoen-Phelan, and R. Ross, "Multi-modal self-supervised representation learning for earth observation," in *IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium*. Brussels, Belgium: IEEE, 2021, p. 3241–3244.

[228] Y. Chen and L. Bruzzone, "Self-supervised remote sensing images change detection at pixel-level," *arXiv preprint arXiv:2105.08501*, 2021.

[229] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "An unsupervised cross-modal hashing method robust to noisy training image-text correspondences in remote sensing," *arXiv preprint arXiv:2202.13117*, 2022.

[230] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing," *arXiv preprint arXiv:2201.08125*, 2022.

[231] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, p. 4904–4916.

[232] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[233] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[234] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

[235] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, p. 721–725.

[236] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 9905. Cham: Springer International Publishing, 2016, p. 801–816.

[237] L. Jing, X. Yang, J. Liu, and Y. Tian, "Self-supervised spatiotemporal feature learning via video rotation prediction," *arXiv preprint arXiv.1811.11387*, 2019.

[238] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, 2020.

[239] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, 2020, p. 111–119.

[240] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 8689. Cham: Springer International Publishing, 2014, p. 818–833.

[241] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, p. 2579–2605, 2008.

[242] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, p. 2217–2226, 2019.

[243] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, p. 2486–2498, 2017.

[244] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2021.

[245] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, p. 1100–1111, 2018.

[246] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The isprs benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I–3, no. 1, p. 293–298, 2012.

[247] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Salt Lake City, UT: IEEE, 2018, p. 172–17209.

[248] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, p. 770–778.

[249] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," 2022, p. 19679–19688.