CMID: A Unified Self-Supervised Learning Framework for Remote Sensing Image Understanding

Dilxat Muhtar, Xueliang Zhang, *Member, IEEE*, Pengfeng Xiao, *Senior Member, IEEE*, Zhenshi Li, and Feng Gu

***Abstract*— Self-supervised learning (SSL) has gained widespread attention in the remote sensing (RS) and Earth observation (EO) communities owing to its ability to learn task-agnostic representations without human-annotated labels. Nevertheless, most existing RS SSL methods are limited to learning either global semantic separable or local spatial perceptible representations. We argue that this learning strategy is suboptimal in the realm of RS since the required representations for different RS downstream tasks are often varied and complex. In this study, we proposed a unified SSL framework that is better suited for RS image representation learning. The proposed SSL framework, contrastive mask image distillation (CMID), is capable of learning representations with both global semantic separability and local spatial perceptibility by combining contrastive learning (CL) with masked image modeling (MIM) in a self-distillation way. Furthermore, our CMID learning framework is architecture-agnostic, which is compatible with both convolutional neural networks (CNNs) and vision transformers (ViTs), allowing CMID to be easily adapted to a variety of deep learning (DL) applications for RS understanding. Comprehensive experiments have been carried out on four downstream tasks (i.e., scene classification, semantic segmentation, object detection, and change detection) and the results show that models pretrained using CMID achieve a better performance than other state-of-the-art SSL methods on multiple downstream tasks. The code and pretrained models will be made available at https://github.com/NJU-LHRS/official-CMIDhttps://github.com/NJU-LHRS/official-CMID to facilitate SSL research and speed up the development of RS images DL applications.**

***Index Terms*— Contrastive learning (CL), deep learning (DL), masked image modeling (MIM), remote sensing (RS) pretraining, self-supervised learning (SSL).**

I. Introduction

DEEP learning (DL) has gained recognition for its ability to automatically extract representations that accurately reflect the intrinsic properties of images, making it a successful tool in understanding remote sensing (RS) images [1], [2], [3], [4]. While supervised DL paradigms have witnessed great success in RS applications [5], [6], [7], [8], they require large-scale, high-quality annotated data, which are difficult to obtain as accurately annotating RS images is time consuming and requires fine domain expertise. Moreover, the reliance on labeled data can also limit the ability of the model to generalize to new or unseen data, which can be a challenge in RS applications where the environment is constantly changing [9]. In contrast, it is much easier to obtain large amounts of unlabeled RS images due to the growing number of satellites used for Earth observation (EO). This limited labeling capability, along with the massive unlabeled images, makes representation learning on RS images highly relevant for self-supervised learning (SSL).

SSL methods learn from vast amounts of unlabeled data by leveraging the structure present in the data itself to create supervised signals [10], [11], [12]. This possibility to train DL models without human-annotated labels, combined with the impressive results achieved by SSL methods on natural images [11], [13], has prompted numerous investigations into adopting SSL for RS images [14], [15], [16]. Contrastive learning (CL) and masked image modeling (MIM) are currently the most widely adopted SSL methods in the domain of RS [10]. CL involves bringing closer the representations of semantically similar inputs (such as two augmented views of the same image) and separating those of different images [17]. The efficiency of CL in learning separable representations and the availability of naturally semantically similar samples within RS images, such as multitemporal data or data acquired from multiple sensors at a single location, have established CL as the leading SSL paradigm in RS [18], [19]. The flexibility of creating contrastive signals in CL has also driven significant efforts to incorporate unique information about RS images into the CL method, resulting in the acquisition of more specialized representations. For example, multitemporal RS images and geographic information have been combined to generate positive pairs, thus allowing the model to learn spatiotemporal invariant representations [15], [16]. Global land cover products and multimodal RS data have also been integrated with the CL method to learn more generalized representations [20], [21]. Most recently, the MIM SSL method, which learns representations by reconstructing the missing information from the masked image has gained popularity in the field of RS [14],

Manuscript received 11 January 2023; revised 19 February 2023 and 27 March 2023; accepted 14 April 2023. Date of publication 19 April 2023; date of current version 3 May 2023. This work was supported by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235. *(Corresponding author: Xueliang Zhang.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: 502022270062@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3268232

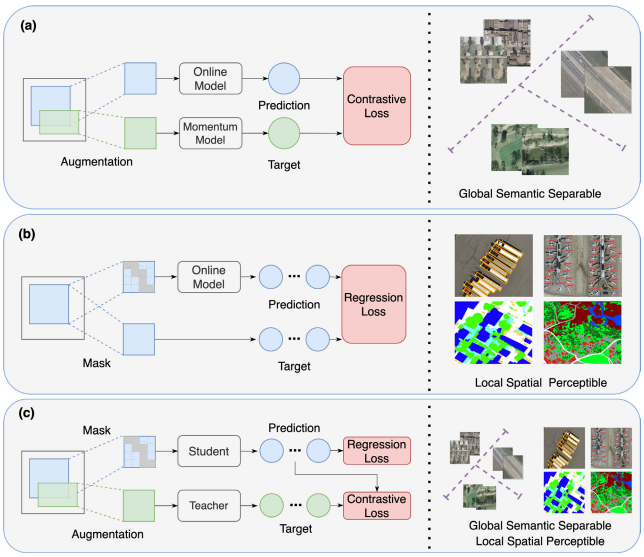

Fig. 1. Comparisons between different SSL paradigms. (a) CL distinguishes global representations of positive pairs from those of negative pairs, making the pretrained model with strong global semantic separability but less local spatial perceptibility. (b) MIM aims at learning local spatial relations for fulfilling the reconstruction task, resulting in the learned representations being more local spatial perceptible but less global semantic separable. (c) Proposed CMID aims to take the best of CL and MIM and learns representations with both global semantic separability and local spatial perceptibility in a unified way.

[22], [23], [24], [25] with the success of vision transformer (ViT) [26] in RS images [25], [27], [28] and the remarkable breakthrough made by MIM in natural images [29], [30]. The recent study [25] applied MIM methods on a customized, large-scale plain ViT model for RS images and conducted extensive experiments on various RS downstream tasks to explore the effectiveness of the MIM learning approach. The results validate the capability of the MIM method in learning rich representations and demonstrate the potential of MIM for representation learning in the field of RS.

Although the existing SSL methods have gained remarkable achievements in representation learning for RS images, they are restricted to utilizing a single-SSL method, either CL or MIM. As a result, the learned representations suffer from a limitation in incorporating both global semantic separability and local spatial perceptibility due to the inherent limitations of CL or MIM methods (Fig. 1). Specifically, CL utilizes whole-image representations to capture interimage relationships by integrating global semantic information, which enhances the separation of representations between different categories or semantics in the feature space, leading to representations with global semantic separability and improved performance in tasks such as linear probing and classification [31], [32]. Hence, as CL merely concentrates on the interimage semantic relationships and ignores the intraimage structure, it fails to perceive the semantic information at different spatial locations within the images, and thus, performs poorly in dense prediction tasks, such as object detection and semantic segmentation [33]. On the other hand, MIM learns the intraimage structure and perceives the contextual information of each pixel within an image, resulting in the learned representations possessing local spatial perceptibility with the expense of global semantic separability. As a result, MIM pretrained models perform well in dense prediction tasks but are not as effective in the classification task [29], [30]. We argue that the most appropriate SSL method for RS images should enhance the performance of the model in various downstream tasks, which means that the learned representations should be both global semantic separable and local spatial perceptible since the required representations for different RS downstream tasks are often varied. For example, RS images are characterized by multiresolution, multiseason, and multispectral, making them vary significantly in time and location [34]. Therefore, a global semantic separable representation is necessary to classify different scenes. On the other hand, the complicated object distribution and spatial arrangement, along with the variations in both object scale and style under different capturing scenes of RS images [35], also require a robust and local spatial perceptible representation in order to accurately detect the objects in these images.

The motivation of learning representations with both global semantic separability and local spatial perceptibility inspires us to design a unified SSL framework that takes the advantage of both CL and MIM. However, it is challenging to fully exploit the advantages of both methods since CL and MIM often use different learning objectives, data augmentations, and model architectures. The different learning objectives of these two methods result in capturing different levels of abstraction and semantics [32], and simply aligning them in the same feature space will lead to semantic confusion and misalignment, ultimately degrading performance on downstream tasks. Moreover, RS images are often known for their multiobject characteristics [36], and the operation of masking out a large portion of images in MIM may vanish the small objects [14] and introduce noise that is not originally present in the images [37], resulting in incompleteness in semantic meaning for the images. Consequently, performing CL on these semantic incomplete images will hinder the learning of separable representations.

In this study, we propose contrastive mask image distillation (CMID) to overcome these challenges and learn both global semantic separable and local spatial perceptible representations in a unified way. CMID leverages teacher–student self-distillation architecture [38], [39] and incorporates the concept of multiple embedding subspaces [15], [40] to take the best of CL and MIM. Specifically, the student encoder maps the masked images into latent embedding, while the teacher encoder encodes the fully augmented images to ensure semantic integrity and provide contrastive supervision to guide the student. The latent embedding is then projected into different feature spaces to perform either reconstruction or discrimination tasks, which helps to prevent confusion between different levels of semantics. In the reconstruction task, the student learns representations that capture local spatial details by recovering the masked parts in both spatial and frequency domains. In the discrimination task, the student learns global semantic separable representations by discriminating the embedding of the student and teacher from those of other images. Additionally, we align the local semantics of the feature vectors at corresponding positions in the teacher's

and student's inputs to overcome the semantic incompleteness caused by the large mask ratio and enhance the local separability of the learned representations. Finally, although the existing MIM methods are largely confined to ViT architectures and have been shown to suffer from information loss when applied to convolutional neural networks (CNNs) [37], [41], CMID has been proven to be effective for both CNN and ViT architectures. This versatility allows for easy adaptation of CMID to a wide range of DL applications for RS images.

The main contributions of this study are as follows.

1) We propose a unified SSL framework, CMID, that learns both global semantic separable and local spatial perceptible representations by combining the CL and MIM, which makes it suitable for a variety of RS downstream tasks. To the best of our knowledge, this is the first SSL framework that combines CL with MIM in the realm of RS.
2) The proposed CMID method adopts a teacher–student self-distillation architecture to fully leverage the benefits of CL and MIM. This approach preserves semantic integrity and learns representations through the distillation of both local and global semantics. Additionally, CMID incorporates the concept of multiple subspaces for latent embeddings to separate the different levels of semantics from CL and MIM, thus preventing semantic confusion.
3) Comprehensive experiments have been carried out on four RS downstream tasks (i.e., scene classification, semantic segmentation, object detection, and change detection) and the results show that models pretrained using CMID achieve better performance than other state-of-the-art SSL methods on multiple downstream tasks. Moreover, these results demonstrate that CMID is effective for both CNN and ViT architectures.

II. Related Work

SSL is a representation learning paradigm that involves training a model to learn from unlabeled data by leveraging pretext tasks to generate labels or supervision from the data itself [10], [11], [12]. The goal of SSL is to learn general representations that can be used for various downstream tasks, such as classification [42] and semantic segmentation [36]. This section briefly discusses the two dominant learning frameworks of SSL [43]: CL and MIM, and their applications to RS images.

A. Contrastive Learning

The core idea of CL is to pull together the representations of different views of the same image (positive samples) and push away that of the other images (negative samples) [17], [44]. SimCLR [45], which utilized a simple siamese network and a plentiful of data augmentation methods, was the first CL method to demonstrate competitive results compared with supervised learning. However, SimCLR directly used negative samples coexisting in the current batch and required a large batch size to ensure the diversity of the negative samples. Considering this limitation, MoCo [11] proposed using a large queue to cache negative samples such that it can take in more negative samples for CL and release the requirement of a large batch size of SimCLR. On the other hand, some CL methods have attempted to eliminate the reliance on negative samples through asymmetric design [13], [46] or reducing redundancy [47]. Specifically, BYOL [13] learned meaningful representations without negative samples by using an asymmetric architecture, where the online encoder predicts the output of a momentum encoder. SimSiam [46] simplified the BYOL architecture by applying stop-gradient operations to replace momentum updates. Barlow-Twins [47] learned representations by reducing the correlations of different images at a batch level. On the other direction, SwAV [48] incorporated clustering into the CL framework to learn more discriminative representations by enforcing consistency between the assignments of the multiaugmented views of the same image. Finally, MoCo-v3 [49] and DINO [39] adopted MoCo [11] and BYOL [13] for pretraining the ViT [26] network using CL methods.

CL pretrained models have achieved great success in classification tasks due to their ability to learn global semantic separability during the pretraining stage. However, their transfer performance for dense prediction tasks, which require more fine-grained and local information, has been limited [50]. In an effort to improve this performance, several recent studies have attempted to learn at a local scale by imposing consistency among pixels [51], [52], feature maps [50], [53], or image regions [54], [55]. Despite these efforts, this approach has been found to be prone to overfitting on the pretraining dataset and has limited performance on downstream tasks [56].

B. Masked Image Modeling

MIM learns rich visual representations via masked parts reconstruction by conditioning on visible context. Following BERT in natural language processing (NLP), iGPT [57] compressed the image to a few pixels, and then directly predicted the masked pixels' color. However, the pixel-based method has a high pretraining computation cost. To reduce computation costs, BEiT [58] proposed using a pretrained discrete variational autoencoder [59] as the visual tokenizer to encode masked patches and learn representations by reconstructing the masked visual tokens. Following this direction, PeCo [60] proposed to inject perceptual similarity during the training of visual tokenizers to benefit MIM pretrained representations. Instead of using the pretrained visual tokenizer, iBOT [61] jointly trained the visual tokenizer and the target network via self-distillation. MAE [29] improved the pretraining efficiency of the MIM method by using an asymmetric encoder–decoder architecture to reconstruct raw images. Compared to MAE, MaskFeat [62] and SimMIM [30] used simple linear layers as the decoder head instead of another transformer as in MAE. MaskFeat further enriched the learned representations by adopting a histogram of gradient (HOG) as the reconstruction target instead of RGB values.

MIM pretrained models excel in transfer learning (TL) with fine-tuning [63] and have demonstrated promising performance in dense prediction tasks [64]. However, they tend

to focus on learning the relationships within each image independently, without considering the interimage relationships, leading the learned representations to be local semantic perceptible but less separable. In addition, most MIM methods have only been applied to ViT, and while there have been some attempts [41] to apply MIM to CNN, they have not performed as well as their ViT counterparts. The underlying issue behind MIM with CNN lies in the fact that CNN performs sliding windows on dense and regular inputs, hindering its ability to concentrate solely on the irregular visible regions to restore lost information. Moreover, as the masked patches shift the distribution of the original image and the convolution operation operates on both the visible and invisible patches, leading to further corruption of the image information [41], [65].

C. Self-Supervised Learning in Remote Sensing

The success of SSL in natural image representation learning has inspired numerous efforts to adapt SSL for RS images. CL is once the most popular SSL method in the RS community. CMC [66] was first adapted to the RS image classification task and demonstrated the applicability of CL in RS images [42]. SimCLR [45] was then incorporated with the idea of neighboring positive pairs to increase the diversity of the training data and learn smoothed representations for RS images [67]. SauMoCo [18] applied MoCo [11] CL method and defined spatial augmentations criteria to uncover semantic relationships among land cover tiles. As a result, SauMoCo was able to characterize different unlabeled RS scenes. Much work has also been done to incorporate additional information into CL to encourage models to learn more specific representations from RS images. Geography-aware MoCo [16] was able to close the gap between self-supervised and supervised learning on various RS images downstream tasks by leveraging spatially aligned images over time to construct temporal positive pairs in CL and using the geolocation regression task as an auxiliary task. Similarly, SeCo [15] used images from the same location at different temporal phases as positive pairs and applied the concept of multiple embedding subspaces to learn representations that encode the different variances and invariances. GeoKR [20] utilized the global land cover products and regarded the geographical location associated with each RS image as geographical knowledge to provide supervision for SSL pretraining. Moreover, there are some works that focus on the application of CL methods to specific downstream RS tasks. For example, CL was integrated with change detection using multimodal bitemporal scenes in an encoder–decoder architecture and achieved promising success in multisensor change detection [68]. GLCNet [69] and IndexNet [36] combined global and local contrast to learn more fine-grained representations for the semantic segmentation of RS images.

Recently, the MIM method has been increasingly gaining popularity in the RS community owing to the success of the ViT in various RS downstream tasks. SITS-Former [22] proposed to learn spatiospectral–temporal representations for time series classification by asking the model to regress central pixels of masked patches given an incomplete time series with some randomly masked patches. SatMAE [23] leveraged temporal and multispectral information in RS images to improve self-supervised pretraining with MIM. RingMo [14] applied the MAE [29] method and designed a new mask strategy for self-supervised representation learning on a 3 million unlabeled RS images dataset. The fine-tuning results on various downstream tasks showed that the new mask strategy was more appropriate for RS images and the learned representations by RingMo were generalized well to various RS downstream tasks. MAE [29] was also applied to pretrain large vision models customized for RS, and the results on various RS downstream tasks demonstrated the effectiveness of MIM pretraining [25].

Although these approaches have achieved certain success, they are limited to learning either global semantic separable or local spatial perceptible representations. Moreover, these methods have prerequisites for the underlying architecture (i.e., CL methods rely on CNN, while MIM methods are confined to ViT). In contrast, the proposed method, CMID, is not only architecture-agnostic but also learns representations with both global semantic separability and local spatial perceptibility, ensuring that the learned representations are generalizable enough to meet the requirements of various RS downstream tasks.

III. Method

A. Overview

The overall structure of the proposed CMID model is illustrated in Fig. 2. CMID employs a teacher–student self-distillation architecture and consists of three branches: the MIM branch, the global branch, and the local branch. The MIM branch leverages the MIM method to learn local spatial perceptible representation. The global branch, using the CL method, focuses on learning global semantic separable representation. Meanwhile, the local branch aims to recover object-level information lost in the MIM branch through self-distillation. The entire network is designed to interact and balance between the learning signals from each branch through its teacher–student architecture. Specifically, given an image x, CMID first generates two views of x, referred to as masked image and augmented image, using either mask augmentation or random data augmentations, and feeds them into the student and teacher, respectively. The teacher and the student share the same architecture (CNN or ViT). The parameters of the student are exponentially moving averaged (EMA) to the parameters of the teacher. The student maps the masked image to latent embedding, while the teacher encodes the augmented image to preserve semantic integrity and provide contrastive supervision to guide the student. The MIM branch applies an MIM head to reconstruct the masked image using the latent embedding output by the student in both spatial and frequency domains. The other two branches project the latent embedding to different feature spaces to avoid semantics confusion and leverage CL to learn separable representations by aligning the embedding of the student and teacher at both global and local scales. Specifically, the global branch utilizes the standard MoCo [11] style CL method to discriminate global representations, while the local branch aligns the local semantics by enforcing

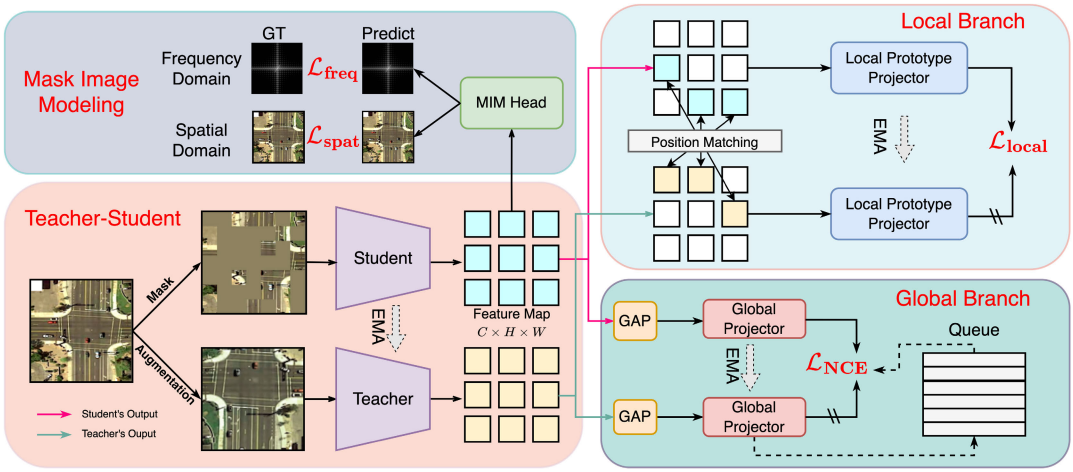

Fig. 2. Overview of the proposed SSL framework, CMID, which adopts a teacher–student self-distillation architecture and consists of three branches.

assignment consistency between position-matched pairs over a set of prototypes to learn object-separable representations and avoid semantic incompleteness and misalignment.

After pretraining using CMID with a large amount of unlabeled data, only the student is used for downstream tasks to evaluate the model performance. In Sections III-B–III-D, we will provide a more detailed explanation of the different components of CMID.

B. Masked Image Modeling Branch

In the MIM branch, we follow the practice of SimMIM [30] and adapt it to the representations learning of RS images with new careful designs.

1) SimMIM Framework: For the input image x, SimMIM first patchifies x into nonoverlapping image patches x^p and randomly generates a mask M to hide a portion of the image patch. The masked patches are then replaced by a learnable mask token [MASK] which is initialized with a value of zero to indicate the presence of missing patches to be predicted. The resulting image is referred to as x_{mask}

$$x_{\text{mask}} = x^p \odot (1 - M) + [\text{MASK}] \odot M \tag{1}$$

where $\odot$ is elementwise multiplication. The SimMIM encoder takes in the masked image x_{mask} and generates a latent embedding. The lightweight one-layer MLP decoder (MIM head in Fig. 2) then uses this embedding to output the predicted reconstruction image x'. Finally, an ℓ_1 loss $\mathcal{L}_{\text{spat}}$ is employed on the masked pixels to update the model parameters

$$\mathcal{L}_{\text{spat}} = \frac{1}{\Omega(x_m)} \parallel x_m - x'_m \parallel_1 \tag{2}$$

where x_m and x'_m denote the sets of original and reconstructed values of the masked pixels, respectively; and $\Omega(\cdot)$ represents the number of elements in a set.

2) Adaptation for RS Images: Although the simplicity and effectiveness of SimMIM, there are some limitations must be taken into account when applying SimMIM into RS images. One issue is that SimMIM replaces masked patches with the [MASK] token, but RS images are known for their multiobject characteristics [36] and the objects are usually densely distributed [14]. The masking operation may cause the dense and small objects in the image to be lost [14], leading to incomplete semantic meaning and making image reconstruction more difficult [14]. Moreover, the zero-initialized mask token [MASK] is not originally presented in the image. Using it directly to replace the masked patches will cause inconsistency between the masked and augmented images, which will degrade the discriminativeness of the learned representations in the other two branches.

To address the above-mentioned issues, we adopt the strategy of using the mean spectral value to fill the masked patches, as suggested in [70], and add the learnable mask token [MASK] to the patch embedding (the output of the patch embedding layer in ViT or the output of the stem layer in CNN), rather than simply replacing the masked patches. The x_{mask} is now expressed as

$$x_{\text{mask}} = x_m^P + \text{MASK} \odot M \tag{3}$$

where x_m^P denotes the set of image patches after the masked patches have been filled with mean spectral value. The consideration behind this design is that the spectral mean value is the direct center (DC) component of the image, resulting in minimum local statistics variation compared with the mask token replacement [70], and it also contains middle-level information about the image. As a result, this masking strategy not only minimizes semantic discrepancy between the masked and augmented images but also reduces the difficulty of recovering the missing information by providing a middle-level information hint. Moreover, we further mitigate the semantic incompleteness caused by the loss of small objects by aligning the local semantics between the student and teacher outputs, which will be described in Section III-D. Finally, we also incorporate the focal frequency loss (FFL) $\mathcal{L}_{\text{freq}}$ [71] in the frequency domain to enforce consistency between the original and the reconstructed image, as this has been shown to be effective to learn representations with high-level semantics [29], [70]

$$\mathcal{L}_{\text{freq}} = \frac{1}{N} \sum_{c=1}^{N} \text{FFL}(x_c, x'_c) \tag{4}$$

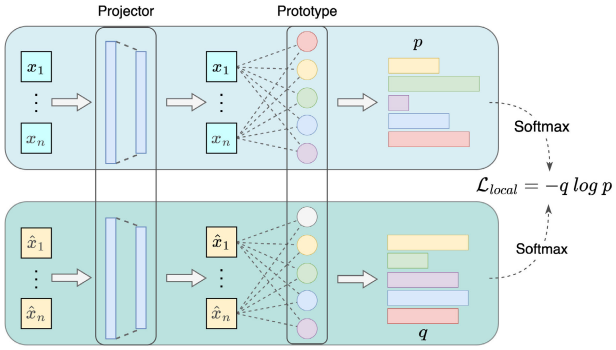

Fig. 3. Detailed process of the local branch. The matched pairs are projected into another feature space and mapped to a set of learnable prototypes to obtain their respective similarities to these prototypes. The cross-entropy loss is then used to minimize the difference in the distributions of these similarities.

where N refers to the number of channels in the input image x; and the subscript c indicates the specific channel of the image. For the detailed formulation of FFL, please refer to [71] and [70].

C. Global Branch

In the MIM branch, the student learns local spatial perceptible representations by capturing fine-grained visual context and semantics but loses global separability. To bridge this gap, the student is trained to recover the global semantic content of the masked image and learn global semantic separable representations by aligning its representation with that of the teacher using a visual dictionary queue in the global branch. The outputs of the student and teacher serve as the query and key in the adopted MoCo [11] CL method. Both the query and key are pooled using global average pooling (GAP) to obtain their respective global representations, which are then projected into another feature space to prevent confusion between different levels of semantics in the MIM and global branches. Finally, we apply infoNCE loss [44] to align their representations by discriminating them from other images

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\langle q, k^+ \rangle / \tau)}{\sum_{i=1}^{K} \exp(\langle q, k_i \rangle / \tau)} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ indicates the cosine similarity; q, k^+ denote the projected global representations of the student and teacher; τ is a temperature scaling the distribution of distance; and k_i is the member of the visual dictionary queue. The queue is built by the teacher's projected global representations and updated in each iteration [11].

D. Local Branch

In Section III-B, we discussed how the mask operation can cause semantic incompleteness by dropping dense and small objects in RS images. This incompleteness not only deteriorates the extraction of local spatial semantics in the MIM branch [14] but also influences the global branch due to the ambiguity it brings to the semantic meaning of the image [32]. To address this issue, we further mitigate semantic incompleteness in the local branch by aligning the local semantics of the student and teacher. This alignment is achieved by enforcing assignment consistency between position-matched pairs over a set of prototypes (Fig. 3). Specifically, we select N position-matched pairs from the feature vectors in the output feature maps of the student and teacher using their absolute positions in original input image x, denoted as $\{(x_i, \hat{x}_i)\}_{i=1}^{N}$. However, since the masked and augmented images are generated by applying two random crops to the original image, there may not be any overlapping regions between the two cropped images [36], [50], [51]. In other words, there are no position-matched feature vectors. To address this issue, we further restrict the size of the minimum cropped image in the random crop process to ensure that there is a certain overlap between the two cropped images. The process for calculating the absolute positions is shown in Fig. 4. Suppose (i_2, j_2, w_2, h_2) representing the (left, top, height, and width) positional parameters of the input cropped image, with the model input and output sizes being $H_{\text{input}} \times W_{\text{input}}$ and $H \times W$. The absolute position (l_1, l_2) of the feature vector at position (u, v) in the feature map can be calculated as

$$l_1 = i_2 + \frac{w_2}{2W} + \frac{W_{\text{input}}}{W}(u-1) \tag{6}$$

$$l_2 = j_2 + \frac{h_2}{2H} + \frac{H_{\text{input}}}{H}(v-1). \tag{7}$$

We calculate the Euclidean distance of each feature vector in the two feature maps using their absolute position, and the top N closest feature vectors are chosen as the position-matched pairs $\{(x_i, \hat{x}_i)\}_{i=1}^{N}$. These matched pairs are then projected into another feature space and mapped to a set of learnable prototypes to obtain their respective similarities to these prototypes. During the training process, these prototypes are gradually learned and model the semantic information of the various objects in the dataset, allowing the model to learn more advanced local object-level semantics than directly aligning the matched pairs using the contrastive loss [36], [50], [69]. Additionally, this approach does not require different matched pairs to correspond to different objects, thereby avoiding the false negatives problem introduced by [36]. Finally, the student is trained to minimize the difference in the distributions over these similarities between the position-matched pairs. Precisely, let $C \in \mathbb{R}^{K \times d}$ denote the K learnable prototypes with dimension d. The similarity distributions p_i and q_i between x_i and $\hat{x}_i$ with respect to C can be expressed as

$$p_i = \text{Softmax}\left(\frac{1}{\tau_s}\langle x_i, C \rangle\right) \tag{8}$$

$$q_i = \text{Softmax}\left(\frac{1}{\tau_t}\langle \hat{x}_i, C \rangle\right) \tag{9}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity and τ_s and τ_t are temperature parameters that control the sharpness of the distribution [39]. Finally, we minimize the cross entropy loss between p_i and q_i

$$\mathcal{L}_{\text{local}} = \frac{1}{N} \sum_{i=1}^{N} -p_i \log q_i. \tag{10}$$

The total loss of CMID is then defined as

$$\mathcal{L} = \lambda_1(\mathcal{L}_{\text{spat}} + \mathcal{L}_{\text{freq}}) + \lambda_2 \mathcal{L}_{\text{NCE}} + \lambda_3 \mathcal{L}_{\text{local}} \tag{11}$$

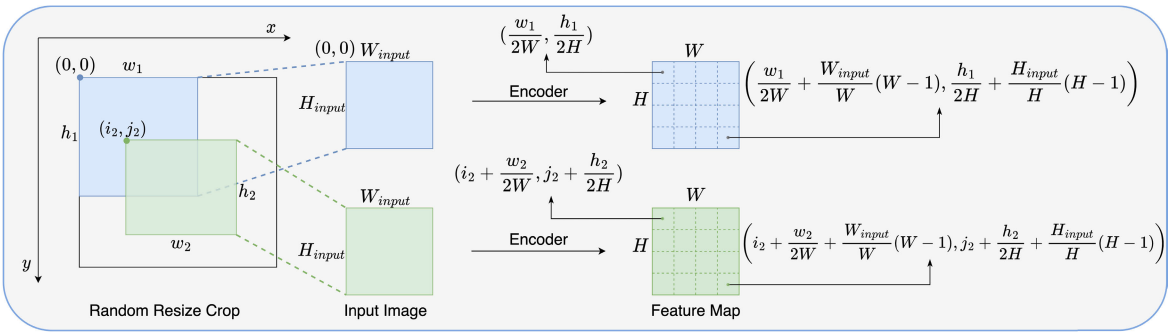

Fig. 4. Detailed process of calculating the absolute position of each feature vector. These absolute positions are calculated using the positional parameters of random resize crop augmentation and the input and output image sizes of the model.

where λ_1, λ_2, and λ_3 act as the weight to balance the three branches.

IV. Experiment

A. Implementation Detail

We use ResNet-50 [72] or Swin transformer-base (Swin-B) [73] as the encoder of the student and teacher. The MIM head in the MIM branch is a 1×1 convolution layer, while the global projector is a one-hidden layer MLP and the local projector is an MLP with two hidden layers. All EMA momentum rates are set to 0.996 and gradually increased to 1 by the cosine annealing scheduler during training. The infoNCE loss temperature τ in (5) and the sizes of the queue are set to 0.2 and 65536 following the best practice of MoCo [11]. The values of τ_s and τ_t in (8) and (9) are set to 0.2 and 0.07, respectively, as suggested in [39]. Based on the results of ablation experiments, we set the mask ratio, number of prototypes K, and the number of selected position-matched pairs N to 0.6, 2048, and 20, respectively. Finally, the weight of each branch (λ_1, λ_2, λ_3) is set to 1 based on the results of our ablation experiments.

B. Ablation Experiment

1) Dataset: The ISPRS Potsdam dataset[1] is used for pretraining and fine-tuning in all ablation experiments. The Potsdam dataset consists of 38 tiles of the same size of 6000×6000 pixel with a spatial resolution of 0.5 m. The dataset has been manually labeled into six categories: low vegetation, tree, building, impervious surface, car, and clutter. We crop these 38 tiles of images to 256×256 pixel with a stride of 128 pixels, yielding a total of 21 888 images for pretraining. For fine-tuning, we crop these images to 512×512 pixel and follow the default train-test split strategy provided by mmsegmentation [74].

2) Experimental Setting: We choose semantic segmentation tasks to evaluate the influence of different components with different parameters on the pretraining performance of CMID, as semantic segmentation requires the pretrained model to have the ability to extract both global and local representations [36], [69].

[1]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

a) Pretrain setting: The ResNet-50 [72] is used as the encoder and is pretrained for 100 epochs on the Potsdam dataset without labels. The Adan [75] optimizer is used, with a batch size of 64, an initial learning rate of 0.003125, and a weight decay of 0.02. The learning rate follows a cosine annealing schedule, starting from 0 to 5 warmup epochs and with a final value of 0.000001.

b) Fine-tune setting: The fine-tuning on the segmentation tasks is implemented using mmsegmentation [74], with UperNet [76] serving as the segmentation decoder. The entire model is trained for 50 epochs with a batch size of 8 using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01, weight decay of 0.0001, and momentum of 0.9. The learning rate is optimized by the polynomial scheduler during the fine-tuning process.

3) Results: All the reported results are the averages of two fine-tuning runs.

a) Three branches: We report the impact of adding or removing different branches on the CMID in Table I. The results show that the best performance is achieved by combining all three branches (row 1 and row 2). Using MoCo [11] in the global branch leads to better results than using the BYOL [13] style CL method. Comparing the use of only the MIM branch (row 4) versus only the CL branches (row 5), we find that the CL performs better, further demonstrating that the MIM method is not well suited for CNN architectures. Combining the MIM and local branches (row 7) gives a result that surpasses those of single or double-branch configurations, indicating that the local branch can effectively compensate for the semantic incompleteness of the MIM branch. CMID collapses when only the global branch is available, as we are not using shuffleBN like the standard MoCo [11] method, which causes the model to learn toward a cheat solution (row 8). Furthermore, the results of ablation experiments on three branches with different weights show that the best performance is achieved when the three branches are balanced (Table II).

b) Mask strategy: We show the effects of using different mask strategies in Table III. The results show that the proposed mask strategy, which fills the mask patches with mean RGB value and adds the mask token to patch embedding, yields the best performance. This demonstrates that the proposed mask strategy results in better interaction between CL and MIM. In addition, the inclusion of reconstructions in the frequency domain further enhances the performance of the pretrained model. Moreover, the visualization results show that

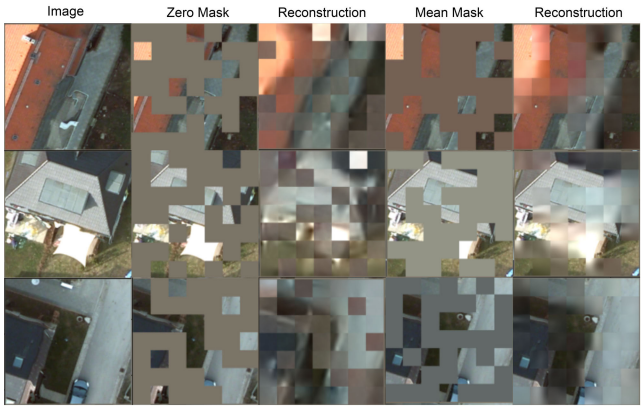

Fig. 5. Reconstruction results for different masking strategies. The results show that using mean RGB value to mask image and adding the mask token to patch embedding leads to better reconstruction results, demonstrating that the proposed mask strategy can provide informative hints for the reconstruction task.

TABLE I
ABLATION RESULTS FOR THREE BRANCHES

Method	MIM	Global Branch	Local Branch	mIoU	OA	mF1
CMID-MoCo	✓	✓	✓	86.97	92.75	92.89
CMID-BYOL	✓	✓	✓	86.33	92.40	92.51
	✓	✓	✗	86.11	92.24	92.39
	✓	✗	✗	84.54	91.21	91.37
	✗	✓	✓	86.06	92.2	92.34
	✗	✗	✓	85.75	92.02	92.18
	✓	✗	✓	86.16	92.26	92.41
	✗	✓	✗	78.04	87.61	87.21

TABLE II
ABLATION RESULTS FOR DIFFERENT BRANCH WEIGHTS

Global Branch (λ_2)	Local Branch (λ_3)	MIM Branch (λ_1)	mIoU	OA	mF1
1	1	1	86.97	92.75	92.89
1	0.5	1	86.13	92.23	92.38
0.5	1	1	86.34	92.38	92.53
1	1	0.5	86.34	92.34	92.51

TABLE III
ABLATION RESULTS FOR DIFFERENT MASK STRATEGIES AND FREQUENCY DOMAIN RECONSTRUCTION

FFT	Mask Strategy	mIoU	OA	mF1
✓	mean + add	86.97	92.75	92.89
	mean + add	86.57	92.53	92.65
✓	zero + replace	85.52	91.82	92.03

the proposed masking strategy achieves better reconstruction results (Fig. 5), further demonstrating that the proposed masking strategy can reduce the difficulty of the reconstruction task and enable the model to learn local spatial perceptible representations more efficiently.

c) Mask ratio: We conduct experiments with various mask ratios and find that a high mask ratio enables the model to learn more robust representations, leading to improved performance on downstream tasks. However, it is important to not set the mask ratio too high, as it can negatively impact model performance (Table IV). Therefore, we set the default mask ratio to 0.6.

d) Number of matched pairs: In this set of ablation experiments, we attempt to identify the optimal minimum cropped size and number of position-matched pairs N that would yield the best performance. The results are shown in Table V. $\text{scale}_{\text{min}}$ represents the ratio between the minimum size of the crop and the original size of the image. As the scale increases, the optimal number of matched samples tends to increase, which supports the notion that aligning more local semantics can improve model performance when there is a large overlap between two cropped images. However, beyond a certain scale, model performance begins to decline. This is likely due to a decrease in representation discriminability in the global branch as the difference between the two views becomes less different when the scale is too large [45]. Finally, we selected the $\text{scale}_{\text{min}}$ of 0.5 and the N of 20 as the default CMID settings.

TABLE IV
ABLATION RESULTS FOR DIFFERENT MASK RATIOS

Mask Ratio	0.30	0.45	0.60	0.75
mIoU	86.33	86.35	**86.97**	86.68

TABLE V
ABLATION RESULTS FOR DIFFERENT MINIMUM CROPPED IMAGE SCALES AND THE NUMBER OF POSITION-MATCHED PAIRS. MIOU AFTER FINE-TUNING IS REPORTED FOR EACH SETTING

$scale_{min}$	N		
	15	20	25
0.2	**86.52**	86.44	86.05
0.3	**86.68**	86.65	86.57
0.4	86.29	**86.69**	86.61
0.5	86.20	**86.97**	86.42
0.6	85.94	**86.76**	86.19

TABLE VI
ABLATIONS RESULTS FOR DIFFERENT NUMBERS OF PROTOTYPES

K	512	1024	2048	4096
mIoU	86.43	86.81	**86.97**	86.69

e) Number of prototypes: We evaluate the influence of the number of prototypes K used in CMID in Table VI. We discovered that the local branch is more robust to the number of prototypes because the results do not vary significantly. As such, we choose 2048 as the default number of prototypes.

C. Pretraining

1) Dataset: We use the MillionAID dataset [77] for CMID pretraining. MillionAID is a massive dataset comprising 1 000 848 nonoverlapping RS images. The images in MillionAID are gathered from Google Earth and captured by various sensors with varying resolutions. On average, the image sizes in the MillionAID dataset range from 110 × 110 pixel to 31 672 × 31 672 pixel. We utilize all the images in the MillionAID dataset for pretraining and resize them to 224 × 224 pixels for network input.

2) Pretraining Setting: We pretrain the CMID network with ResNet-50 or Swin-B as the backbone for 200 epochs in the MillionAID dataset without any label. The Adan optimizer [75] is employed to optimize the model, with a learning rate of 0.0088 and a batch size of 512 for ResNet-50 and 0.002, 256 for Swin-B. We utilize the cosine learning rate

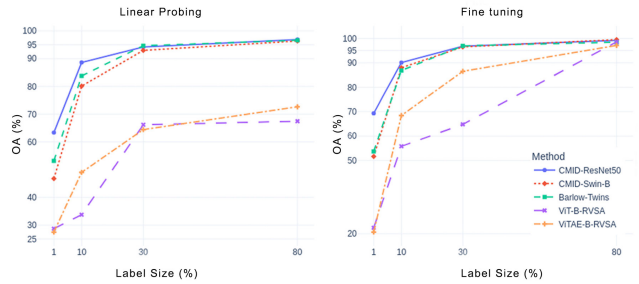

Fig. 6. Evaluation on the UCM classification task with different labels size. For comparison of various ResNet-50-based pretrained models, only the results of the comparison with Barlow-Twins are shown for ease of viewing. Barlow-Twins outperformed the other ResNet-50-based pretrained models, except for CMID, in the classification task.

scheduler with the first 15 epochs for warm-up. The weight decay is set to 0.02. Pretraining is conducted using four NVIDIA V100 GPUs (graphics processing unit). For a fair and holistic comparison between CMID and other SSL methods, we employed the following two comparative strategies.

1) We download the open source pretrained weights of 8 state-of-the-art SSL methods, including BYOL [13], SwAV [48], Barlow-Twins [47], and MoCo-v2 [11] in the field of computer vision and SeCo [15], ResNet-50-SEN12MS [24], ViTAE-B-RVSA [25], and ViT-B-RVSA [25] in the field of RS. We then fine-tuned them in the same settings to evaluate their performance on various RS downstream tasks (Section IV-D-Section IV-G). We consider that comparing with official open-source pretrained models could be more convincing than re-implementing them, as the official pretraining models of various SSL methods are highly tuned and optimized.
2) We pretrained the models using different SSL methods (CMID, MAE [29], SimMIM [30], BYOL [13], Barlow-Twins [47], and MoCo-v2 [11]) under the same pretraining settings and subsequently evaluated their performance on RS downstream tasks. This was done to eliminate any potential influence arising from different pretraining datasets and evaluate the scalability of different SSL methods on small datasets (Section IV-H).

D. Scene Classification

1) UC Merced Land Use Dataset: The UC Merced Land Use (UCM) dataset [78] contains 2100 images with a resolution of 0.3 m. All the images in the UCM dataset have the size of 256 × 256 pixels. The 2100 images equally belong to 21 categories. Thus, each category has 100 images.

2) Experimental Setting: We divide the UCM dataset into a training set and a testing set randomly according to a specific ratio (1:99, 1:9, 3:7, and 8:2). We use the Adan optimizer [75] with a cosine learning rate scheduler and train for 200 epochs. The batch size is set to 64, the base learning rate is set to 0.003125, and the weight decay is set to 0.02. We also employ warmup for ten epochs. We follow the same data augmentation setting as in [25], including RandAug [79], Mixup [80], Cutmix [81], label smoothing [82], and random erasing [83]. During the fine-tuning process, both the backbone and the classification head are updated, but in linear probing, only the classification head is updated. Overall accuracy (OA) is used as the evaluation metric in this downstream task.

TABLE VII

PERFORMANCE OF VARIOUS SSL METHODS ON THE UCM TESTING SET. THE RATIO OF TRAINING SET TO TESTING SET IS 8:2

Method	Backbone	Pre-train Dataset	Pre-train Epoch	Linear Probe	Fine tune
				OA	OA
ImageNet Sup.	ResNet-50	ImageNet	-	94.79	97.92
BYOL	ResNet-50	ImageNet	200	93.23	99.22
Barlow-Twins	ResNet-50	ImageNet	300	96.61	99.16
MoCo-v2	ResNet-50	ImageNet	200	88.80	97.92
SwAV	ResNet-50	ImageNet	200	94.79	98.96
SeCo	ResNet-50	SeCo-1M	200	90.36	97.66
ResNet-50-SEN12MS	ResNet-50	SEN12MS	200	73.18	96.88
MAE	ViT-B-RVSA	MillionAID	1600	67.45	98.56
MAE	ViTAE-B-RVSA	MillionAID	1600	72.66	97.12
CMID	ResNet-50	MillionAID	200	96.88	99.22
CMID	Swin-B	MillionAID	200	96.35	99.48

TABLE VIII

FLOPS AND PARAMETERS SIZE OF DIFFERENT BACKBONE. B, C, H, AND W REPRESENT THE BATCH SIZE, NUMBER OF CHANNELS, AND IMAGE WIDTH AND HEIGHT, RESPECTIVELY

Method	Backbone	Pre-train Dataset	Pre-train Epoch	Potsdam			Vaigingen		
				OA	mIoU	mF1	OA	mIoU	mF1
ImageNet Sup.	ResNet-50	ImageNet	-	92.00	85.69	92.11	90.21	78.53	87.65
BYOL	ResNet-50	ImageNet	200	91.86	85.54	92.07	87.69	73.17	83.97
Barlow-Twins	ResNet-50	ImageNet	300	90.32	83.16	90.67	86.58	71.86	82.02
MoCo-v2	ResNet-50	ImageNet	200	92.74	87.02	92.93	90.26	79.16	88.10
SwAV	ResNet-50	ImageNet	200	92.02	85.74	92.20	87.95	73.76	84.40
SeCo	ResNet-50	SeCo-1M	200	92.04	85.82	92.22	90.14	78.59	87.72
ResNet-50-SEN12MS	ResNet-50	SEN12MS	200	90.50	83.17	90.62	88.62	73.99	84.41
MAE	ViT-B-RVSA	MillionAID	1600	92.40	86.37	92.55	89.54	77.29	86.87
MAE	ViTAE-B-RVSA	MillionAID	1600	92.62	86.61	92.68	89.79	78.17	87.49
CMID	ResNet-50	MillionAID	200	92.94	87.35	93.11	90.23	79.44	88.30
CMID	Swin-B	MillionAID	200	93.52	88.36	93.70	90.50	80.01	88.68

3) Results: The experimental results in Table VII show that all pretrained models achieve promising results when fine-tuning with 80% of the labels. Our CMID pretrained achieves the highest accuracy. Previous studies have shown that the models pretrained using the MIM method excel in TL with full model fine-tuning but perform poorly in linear probing [29], [31], [33]. However, the CMID pretrained models also perform well in linear probing and outperform all other SSL methods, further demonstrating the effectiveness of integrating the MIM and CL methods. It is worth noting that ViTs (Swin-B, ViTAE-B-RVSA, and ViT-B-RVSA) do not perform as well as the CNN network in linear probing, which can be attributed to the different inductive biases of the two architectures. Additionally, we show in Fig. 6 the results of fine-tuning and linear probing with different label sizes. CMID pretrained models show strong performance even when only a small number of labels are available, indicating the robustness of the learned representations during pretraining. Especially, in the one-shot scenario (corresponding to the 1% labels, with one sample per category), CMID substantially outperforms its ResNet-50 or ViTs counterparts. Fig. 6 also shows that, although the CMID pretrained model based on Swin-B performs best with sufficient labels (TOP 1 with 80% labels), it does not perform as well as the ResNet-50 series with limited labels, which we believe is due to the data-hungry nature of ViTs [26].

E. Semantic Segmentation

1) Dataset: We evaluate the semantic segmentation performance of the pretrained models on the Potsdam and Vaihingen[2] datasets. Potsdam has been described in Section IV-B. The Vaihingen dataset contains 33 tiles with a size of

[2]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

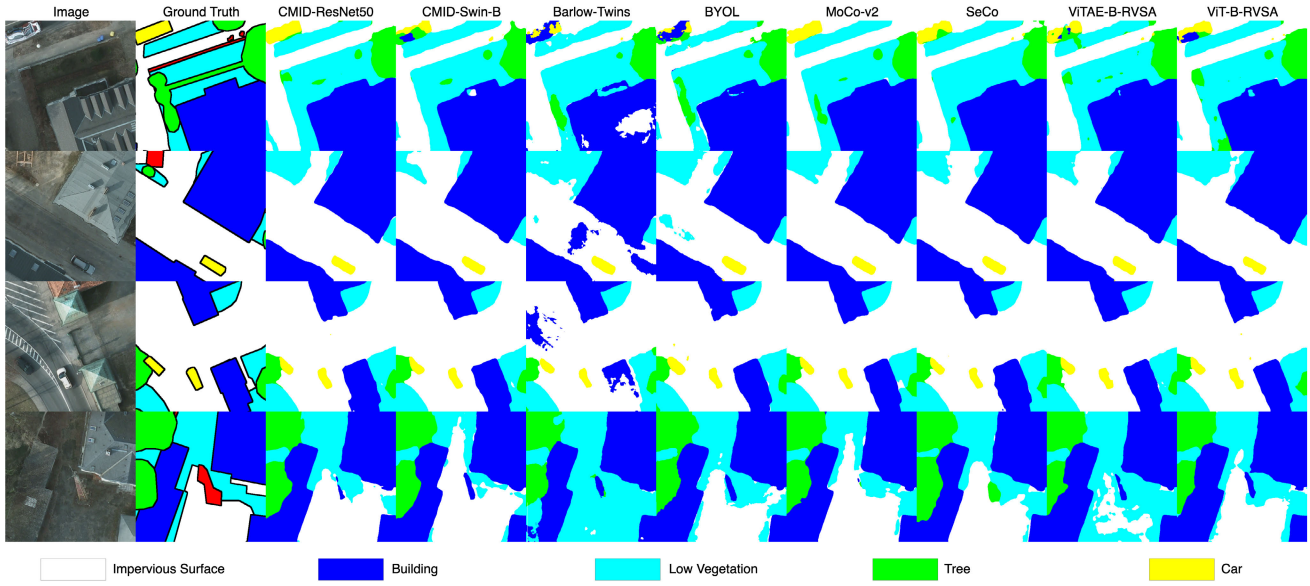

Fig. 7. Visualization results of different methods on the Potsdam dataset. In inference, we disregard the background class, which is designated as red in the ground truth labels.

TABLE IX
SEMANTIC SEGMENTATION RESULTS OF DIFFERENT SSL PRETRAINED MODELS ON THE POTSDAM AND VAIHINGEN DATASETS

Backbone	Input Size (B, C, H, W)	FLOPs (G)	Param. (M)
ResNet-50	(1,3,224,224)	8.26	23.55
ViT-B-RVSA	(1,3,224,224)	33.56	86.00
ViTAE-B-RVSA	(1,3,224,224)	26.33	67.54
Swin-B	(1,3,224,224)	10.61	30.84

[1] The results are evaluated for each backbone in NVIDIA 3080 GPU using the THOP library.

2494 × 2064 pixel and a spatial resolution of 0.9 m. We crop both datasets into patches of 512 × 512 pixel with a stride of 256 pixels and follow the default train–test split strategy provided by mmsegmentation [74].

2) Experimental Setting: We adopt the UperNet [76] framework with the different SSL pretrained models as its backbone. We fine-tune for 50 epochs with a batch size of 8. The ResNet-50-based networks are optimized using the SGD optimizer, where the initial learning rate, weight decay, and momentum are set to 0.01, 0.0005, and 0.9, respectively. The learning rate is adapted by the cosine annealing scheduler. While for the Swin-B, we use the AdamW optimizer with a learning rate of 0.00006 and weight decay of 0.05. For ViT-B-RVSA and ViTAE-B-RVSA, we follow the official fine-tune settings of [25].

3) Results: Table VIII presents the segmentation results of CMID and other SSL methods. When using either ResNet-50 or Swin-B, the CMID pretrained model outperforms the other SSL methods in terms of OA, mean intersection over union (mIoU), and mean F1-score (mF1). In comparison with the ViT-B-RVSA and ViTAE-B-RVSA models, which were pretrained for 1600 epochs, CMID demonstrates a significantly higher level of performance, highlighting the efficiency of CMID's pretraining. Despite the ViT-B-RVSA and ViTAE-B-RVSA [25] are larger and more powerful than ResNet-50 (Table IX), the CMID model, based on ResNet-50, outperforms them in the semantic segmentation task, further indicating that CMID learns more rich and useful representations during the pretraining stage. Additionally, the visualization results in Fig. 7 demonstrate that CMID pretrained models consistently exhibit superior performance. Since semantic segmentation requires both global and local representations [36], [69], the excellent performance of CMID in semantic segmentation demonstrates that CMID can learn representations with both global semantic separability and local spatial perceptibility during pretraining.

F. Object Detection

1) Dataset: DOTA [84] is a large-scale dataset for oriented bounding box (OBB) detection, with a total of 2806 images ranging in size from 800 × 800 pixel to 4000 × 4000 pixel. These images contain a total of 188 282 labeled instances belonging to 15 categories, which are divided into training, validation, and testing sets with 1411, 458, and 937 images, respectively. Following the official setting of mmrotation [85], we sample and crop the DOTA dataset to patches of 1024 × 1024 pixel with a stride of 824 pixels.

2) Experimental Setting: We adopt the OBB detection framework ORCN [86] to perform TL to object detection in mmrotation [85]. For evaluation on ResNet-50, we use the SGD optimizer with a learning rate of 0.005, weight decay of 0.0001, and a maximum gradient norm of 35.0. For Swin-B, we use an AdamW optimizer with a learning rate and weight decay of 0.0001 and 0.05, respectively. For ViT-B-RVSA and ViTAE-B-RVSA, we follow the official fine-tuning hyperparameters [25]. All models are trained for 12 epochs with a batch size of 2, and the learning rate is reduced by a factor of 10 after the eighth and 11th epochs. We report the mean average precision (mAP) and all evaluation results are obtained by submitting the predictions for the testing set to the online server.

TABLE X

RESULTS OF DIFFERENT SSL PRETRAINED MODELS ON THE TESTING SET OF THE DOTA DATASET

Method	Backbone	Pre-train Dataset	Pre-train Epoch	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
ImageNet Sup.	ResNet-50	ImageNet	-	89.12	83.08	52.42	72.99	78.22	81.56	87.90	90.90	85.81	85.01	62.74	65.22	72.78	68.80	57.41	75.62
BYOL	ResNet-50	ImageNet	200	89.35	79.66	49.66	71.16	77.37	81.17	87.82	90.87	85.93	85.27	49.78	63.03	66.48	66.21	61.14	73.62
Barlow-Twins	ResNet-50	ImageNet	300	87.72	70.96	42.29	64.56	74.64	75.45	85.65	90.83	82.25	79.70	38.01	57.81	53.57	64.94	44.76	67.54
MoCo-v2	ResNet-50	ImageNet	200	89.27	78.29	51.21	71.67	78.40	81.32	87.92	90.89	84.34	84.88	56.30	60.91	67.39	65.32	50.72	73.25
SwAV	ResNet-50	ImageNet	200	89.56	78.69	48.74	69.93	77.10	80.56	87.64	90.90	86.40	84.43	52.42	61.87	64.52	63.65	63.20	73.30
SeCo†	ResNet-50	SeCo-1M	200	88.64	73.32	49.94	67.07	76.15	76.48	70.02	90.88	79.46	81.31	41.31	58.55	65.90	65.32	49.41	70.02
MAE	ViT-B-RVSA	MillionAID	1600	89.08	82.34	55.18	72.37	79.23	85.01	88.17	90.88	87.40	86.81	60.66	68.72	76.68	78.35	70.33	78.08
MAE	ViTAE-B-RVSA	MillionAID	1600	89.28	76.97	53.85	69.67	79.79	84.93	88.31	90.84	85.36	85.56	59.15	67.50	77.17	80.08	65.91	76.96
CMID	ResNet-50	MillionAID	200	89.34	84.60	52.55	73.69	78.10	82.71	87.60	90.88	87.43	85.54	67.11	64.97	74.35	71.51	59.04	76.63
CMID	Swin-B	MillionAID	200	89.46	83.54	53.51	74.79	78.72	84.14	87.58	90.90	86.08	85.77	65.97	68.05	74.02	73.27	64.62	77.36

[1] PL: plane. BD: baseball diamond. BR: bridge. GTF: ground track field. SV: small vehicle. LV: large vehicle. SH: ship. TC: tennis court. BC: baseball court. ST: storage tank. SBF: soccer ball field. RA: roundabout. HA: harbor. SP: swimming pool. HC: helicopter.

[2] † represents the result from [27].

Fig. 8. Visualization of detection results on DOTA testing set. The results show that the proposed CMID-Swin-B pretrained model demonstrates comparable detection results to those of the ViT-B-RVSA.

3) Results: The results of the OBB detection are shown in Table X. The detector using the CMID pretrained ResNet-50 as its backbone surpasses other ResNet-50-based self-supervised pretrained models by a significant margin. It is worth noting that while ViTAE-B-RVSA and ViT-RVSA [25] use a longer pretraining strategy and employ a rotational window attention mechanism for RS images object detection task, our CMID-ResNet50 achieves comparable results to ViTAE-B-RVSA and the proposed CMID-Swin-B even outperforms the ViTAE-B-RVSA and yields comparable results to ViT-B-RVSA. The visualization results in Fig. 8 also suggest that the detector using the Swin-B pretrained by CMID as the backbone achieves comparable performance with the ViT-B-RVSA.

TABLE XI

CHANGE DETECTION RESULTS OF DIFFERENT SSL PRETRAINED MODELS ON THE CDD TESTING SET

Method	Backbone	Pre-train Dataset	Pre-train Epoch	mF1
ImageNet Sup.†	ResNet-50	ImageNet	-	95.09
BYOL	ResNet-50	ImageNet	200	96.30
Barlow-Twins	ResNet-50	ImageNet	300	95.63
MoCo-v2	ResNet-50	ImageNet	200	96.05
SwAV	ResNet-50	ImageNet	200	95.89
SeCo	ResNet-50	SeCo-1M	200	96.26
ResNet-50-SEN12MS	ResNet-50	SEN12MS	200	95.88
ImageNet Sup.†	ViTAEv2-S	ImageNet	300	97.02
MillionAID Sup.†	ViTAEv2-S	MillionAID	300	96.81
CMID	ResNet-50	MillionAID	200	96.95
CMID	Swin-B	MillionAID	200	97.11

[1] † represents the result from [27].

G. Change Detection

1) Dataset: The CDD dataset [87], which contains 11 pairs of bitemporal images obtained from Google Earth at different seasons, has a spatial resolution that ranges from 3 to 100 cm per pixel. These images include seven pairs of images with a size of 4725 × 2200 pixel and four images with a size of 1900 × 1000 pixel. We crop the images to a series of patches of 256 × 256 pixel following [88] and generate 10 000 image pairs for training, 3000 image pairs for validation, and 3000 image pairs for testing.

2) Experimental Setting: Following [27], we adopt the BiT [89] framework for evaluation on the change detection task. We use the SGD optimizer with a learning rate of 0.001, a batch size of 16, and a weight decay of 0.0005 for the ResNet-50 series. For the BiT models based on Swin-B or ViTAE, we use the AdamW optimizer with a learning rate of 0.00006 and a weight decay of 0.01. All models are trained for 200 epochs, with the learning rate decreasing linearly during the training process. Once the fine-tuning is complete, we report the mean F1-score (mF1) on the testing set.

3) Results: The results of the change detection are presented in Table XI. The BiT model based on CMID-ResNet50 performs better than other ResNet50-based models and even outperforms the ViTAEv2-S model, which is pretrained on the MillionAID dataset with supervision. In addition, the proposed CMID-Swin-B model demonstrates the best performance on the change detection tasks. The visualization results on the CDD change detection dataset (Fig. 9) further indicate the

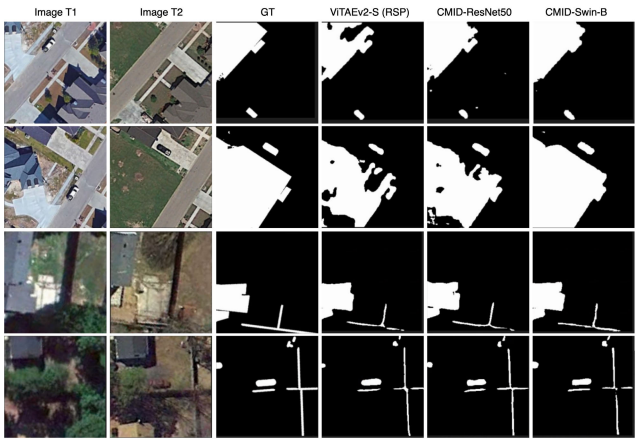

Fig. 9. Visualization of change detection results on CDD testing set. Image T1 and image T2 are the first and second temporals of the same regions. GT refers to ground truth.

capability of the CMID pretrained models in detecting small variations in the images and making more precise predictions. The excellent performance of CMID in the change detection task demonstrates that the representations learned by CMID are invariant enough to avoid pseudochanges caused by imaging conditions, and because the learned representations are separable, the model pretrained by CMID is able to detect changes in different objects within the image.

H. Scalability

We pretrained CMID and several other SSL methods on the relatively small Potsdam dataset under the same pretraining settings to eliminate any possible effects resulting from variations between natural and RS images and evaluate the scalability of different SSL methods. Afterward, we fine-tuned these models to the Potsdam semantic segmentation task to assess the in-domain generalization ability of different SSL methods. Additionally, we fine-tuned these pretrained models to the more challenging DOTA OBB detection task to further evaluate their cross-domain generalization ability.

1) Experimental Setting: The Potsdam dataset is processed in the same manner, as described in Section IV-B during pretraining, without the use of any labels. All models are pretrained from scratch for 400 epochs with a batch size of 64 in the Potsdam dataset. The Adan optimizer [75] with a learning rate of 0.003125 and a weight decay of 0.02 is used for the CMID model. In addition, we implement five state-of-the-art SSL methods (MoCo [25], BYOL [13], Barlow-Twins [47], MAE [29], and SimMIM [30]) using the mmselfsup library [90], with hyperparameters taken from their official settings, except that the learning rate is scaled linearly according to the batch size.

After pretraining, all pretrained models are evaluated on the Potsdam semantic segmentation task and the DOTA OBB detection task using the same settings described in Section IV-E and Section IV-F.

2) Results: The results are presented in Table XII. Despite the fact that all SSL methods pretrained on a small dataset yield degraded results compared with their counterparts trained on a larger dataset, CMID outperforms all other SSL methods

TABLE XII

POTSDAM SEMANTIC SEGMENTATION AND DOTA OBB DETECTION RESULTS OF DIFFERENT SSL METHODS. ALL THE MODELS ARE PRETRAINED ON THE POTSDAM DATASET

Method	Backbone	Potsdam			DOTA
		OA	mIoU	mF1	mAP
ImageNet Sup.	ResNet-50	92.00	85.69	92.11	75.62
BYOL	ResNet-50	90.86	83.80	91.00	59.35
Barlow-Twins	ResNet-50	88.31	79.47	88.31	57.20
MoCo-v2	ResNet-50	91.51	84.88	91.65	66.90
MAE	ViT-B	83.48	71.70	83.22	50.17
SimMIM	Swin-B	89.37	81.09	89.33	64.85
CMID	ResNet-50	92.74	87.04	92.81	72.12
CMID	Swin-B	91.77	85.17	91.83	70.45

and even achieves comparable results to models pretrained with other SSL methods (e.g. MoCo-v2 and BYOL) on large datasets. This is especially evident in the more challenging DOTA OBB detection task, where CMID pretrained models outperform other self-supervised pretrained models by a large margin, highlighting the superior scalability and generalization ability of CMID. It is also worth noting that ViTs do not perform as consistently as CNNs in this setting. This can be attributed to the data-hungry nature of ViTs and the findings of recent research on the relationship between the model size and the pretraining dataset size [91], which suggest that higher learning performance can be achieved with smaller models on a smaller dataset, while larger models may not perform as well as smaller ones. This is further supported by the CMID results, where the CMID-Swin-B model pretrained on the MillionAID dataset performs better, but the CMID-Swin-B model does not perform as well as the CMID-ResNet50 model in this scalability evaluation setting.

I. Failure Case

In this section, we visualize the failure cases encountered when fine-tuning the CMID pretrained models for semantic segmentation and object detection, as shown in Fig. 10. Our aim is to examine the limitations of CMID and offer guidance for future improvement.

1) Semantic Segmentation: We visualize the semantic segmentation results on the Potsdam dataset in Fig. 10(a). The pretraining and fine-tuning settings are the same as Section IV-E. The results show that the pretrained CMID model is inclined to mix up the two categories when they exhibit a high level of similarity. For example, the impervious surfaces and buildings in the first and second rows share a similar color and texture, and the shadows caused by the imaging angle or lighting conditions amplify this similarity, causing the model to misinterpret the two categories. We firmly believe that resolving this issue necessitates an understanding of the interdependence between different categories and the implementation of constraints on this relationship to prevent any misunderstandings or mix-ups between categories.

2) Object Detection: We visualize the failed OBB detection results on the DOTA dataset in Fig. 10(b) following the same setting described in Section IV-F. The results indicate that ResNet-50 pretrained by CMID performs inadequately in detecting small objects. This is likely due to the information loss issue when applying the MIM method to CNNs [37], [41]. Although CMID effectively combines CL and MIM to

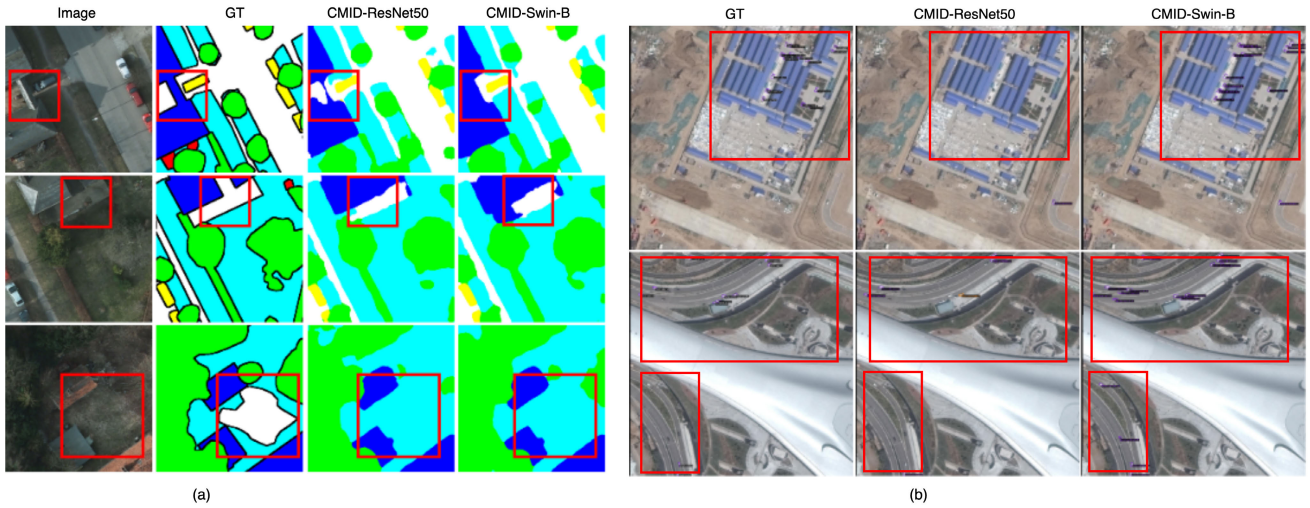

Fig. 10. Visualization of failure cases when fine-tuning the CMID pretrained models for semantic segmentation and object detection tasks. GT refers to ground truth. (a) Semantic segmentation. (b) Object detection.

allow the MIM method to be applied to CNNs, the inherent continuity of the convolution operation restricts it to regular grids and cannot handle irregular, nonoverlapping patches, such as visible and invisible ones, leading to decrease in performance [65]. The key to resolving this issue is to modify the convolution operation to handle irregular, nonoverlapping patches, such as by implementing sparse convolution [92] or partial convolution [93].

V. Discussion

A. Characteristics of the Representations Learned by CMID

The primary goal of CMID is to combine MIM and CL to learn representations that are both global semantic separable and local spatial perceptible. The exceptional performance of CMID in various RS downstream tasks has shown that the representations learned by CMID indeed have these two characteristics. In this section, we will further investigate the representations learned by CMID using various visualization tools.

We start by visualizing the extracted representations on the UCM dataset using the t-SNE [94] method. As shown in Fig. 11, the extracted representations using the proposed CMID method (e and f) are better separated than those obtained from other SSL methods. It is worth noting that the representations extracted using only the MIM method [Fig. 11(c) and 11(d)] are not highly separable, but CMID is able to achieve more globally semantically separable representations by combining CL with MIM. It is also worth noting that the representations extracted by CMID-Swin-B are not as separable as those extracted by CMID-ResNet50. This may explain why CMID-Swin-B does not perform as well as CMID-ResNet50 with limited training data in the classification task (Fig. 6). We expect that further adjustments to the CL learning strategy could enhance the separability of the representations extracted by ViTs.

In Fig. 12, we visualize the attention maps of the self-attention layers for models pretrained with ViTAE-B-RVSA [25] and CMID-Swin-B. The results show that the

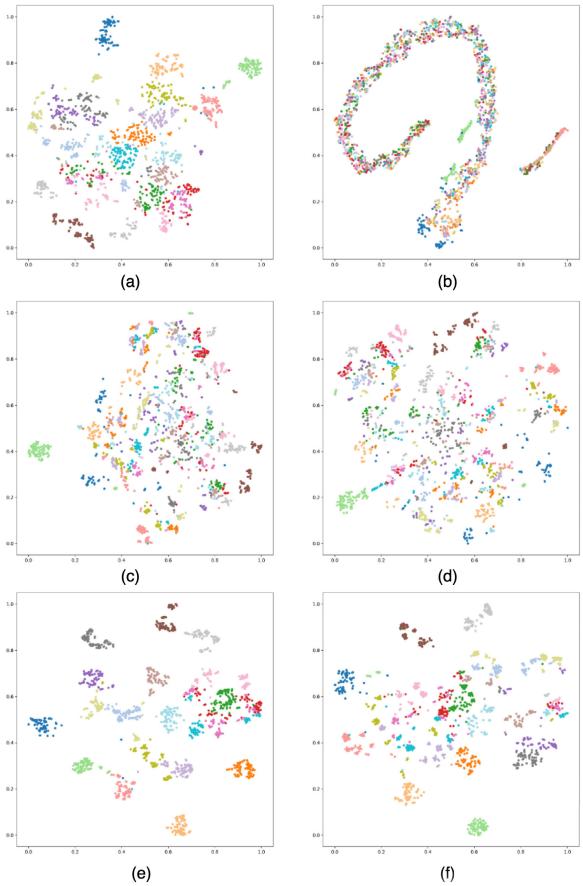

Fig. 11. t-SNE visualization results of representations extracted from different pretrained models are shown in (a)–(f). These results were obtained by extracting the output of the last layer of the pretrained models (which were not fine-tuned) on the UCM dataset. The models shown are as follows. (a) Barlow-Twins. (b) SeCo. (c) ViT-B-RVSA. (d) ViTAE-B-RVSA. (e) CMID-ResNet50. (f) CMID-Swin-B.

ViTAE-B-RVSA pretrained model focuses more on redundant and low-information regions, such as the apron around an aircraft (first row of Fig. 12) or the sea near a ship (fourth row of Fig. 12). This aligns with the MIM method's tendency

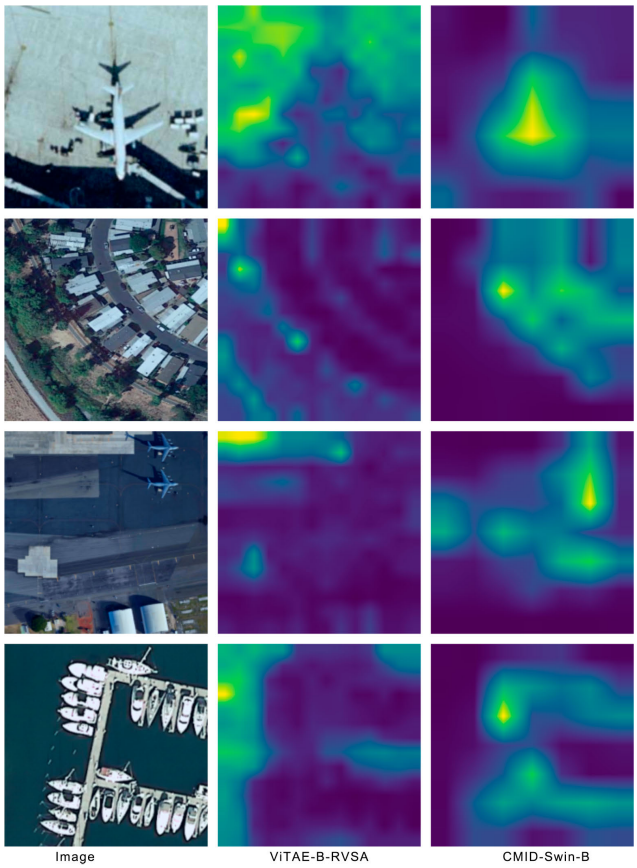

Fig. 12. Visualization the self-attention map of the last layer of the pretrained models by averaging the values of the heads of the last attention layer. These maps demonstrate that the CMID pretrained model is able to effectively focus on the main objects in the image, rather than on other details, indicating its ability to learn and distinguish important representations. These models have not undergone any fine-tuning or been attached to any labels.

to recover noise and low-level image details [62], [95]. In contrast, the proposed CMID pretrained model focuses more on semantically informative regions, which enables the model to more effectively locate the main object in images and improves its performance in downstream tasks, such as segmentation and detection.

Finally, we visualize the most similar feature correspondences between two augmented views of the same image or two images labeled as the same class using the output feature map of the third stage of the pretrained models (Fig. 13). The 12 correspondences with the highest similarity scores are shown. The results demonstrate that CMID is able to accurately match most correspondences even with large changes in color and scale Fig. 13(b) and 13(d). It also has a strong ability to distinguish between different objects without causing confusion, as seen in the precise matching of the same objects in different images of the same class Fig. 13(a) and 13(b). These results indicate the robustness of CMID's learned representations among different augmentations and its ability to discriminate between different objects.

B. Expectation and Limitation

Learning effective representations that are universal across various RS visual understanding tasks, without human annotated labels in training, is a longstanding goal for analyzing

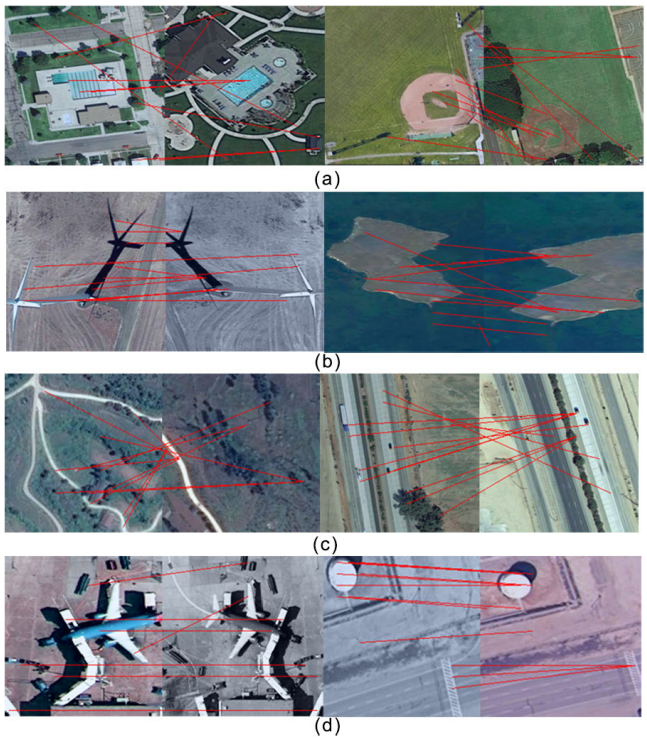

Fig. 13. Visualization of feature correspondence. The correspondence is extracted using CMID pretrained Swin-B (a) and (b) and ResNet-50 (c) and (d). Image pairs in (a) and (c) are sampled from two images of the same class, while (b) and (d) are generated from two augmented views of a single image. These matching results demonstrate the strong ability of CMID to match features at a local scale despite variations in color and spatial locations.

RS images. SSL, with its ability to learn task-agnostic representations by effectively exploiting the huge quantity of unlabeled RS data, has witnessed great success in the field of RS [14], [15], [16], [25]. Many studies have also explored incorporating additional information to enrich the learned representations [22], [23]. However, most existing RS SSL methods are limited to learning either global semantic separable or local spatial perceptible representations, resulting in methods that may work well for a specific downstream task but not for others. We suppose that a successful SSL method should be able to improve the performance of the model for any downstream task and that should be unified in the sense that it can be applied to different network architectures and used with different-sized datasets. Along this direction, we propose CMID to unify the SSL framework for the representation learning of RS images. We anticipate that the proposed CMID will be applied to a broader range of datasets and to more and newer network architectures in order to fully realize its potential. We also expect that the unified learning framework will serve as a baseline for representation learning of RS images and lead to advances in the field of automatic interpretation of RS images.

The proposed CMID SSL has demonstrated impressive results; however, it still has some limitations. First, CMID employs the SimMIM [30] method in the MIM branch to preserve the structure of the image, allowing the application of the MIM method to hierarchical architectures, such as Swin-transformer [73] and ResNet [72]. However, the SimMIM encoder processes both visible and invisible patches, unlike the

MAE [29] method, which only processes the visible patches, resulting in reduced learning efficiency compared to MAE. Thus, a more efficient MIM method that can be used for hierarchical architectures may be a viable alternative to the MIM branch of CMID. Second, while CMID successfully integrates CL and MIM to enable the application of MIM to CNNs, the inherent continuity of the convolution operation restricts it to regular grids, hindering its ability to handle irregular, nonoverlapping patches, such as visible and invisible patches, leading to decreased performance [65]. Thus, there is a need to modify the convolution operation to fully utilize the MIM learning signals and explore the potential of MIM pretraining on CNN. Finally, the representations extracted by CMID-Swin-B are less separable than those extracted by CMID-ResNet50 (Fig. 11). It is assumed that adapting the global branch of CMID to better suit the characteristics of ViT could result in improved global semantic separability of the learned representations.

VI. Conclusion

We present CMID, an SSL framework for representation learning of RS images. CMID combines the advantages of CL and MIM to learn representations with both global semantic separability and local spatial perceptibility in a self-distillation manner. Through extensive experiments on various downstream RS tasks, we demonstrate that CMID significantly enriches the learned representations and outperforms state-of-the-art self-supervised methods on multiple downstream tasks. Additionally, our CMID framework is architecture-agnostic, allowing it to be easily applied to both CNNs and ViTs for a wide range of DL applications in RS image interpretation. We expect CMID will serve as a baseline for representation learning of RS images and lead to advances in the field of automatic interpretation of RS images.

Acknowledgment

The authors are grateful to the High-Performance Computing Center, Nanjing University, Nanjing, China, for their help with GPU resources. They would also like to thank the editor and the anonymous reviewers for their constructive comments.

References

[1] M. H. Saleem, J. Potgieter, and K. M. Arif, "Automation in agriculture by machine and deep learning techniques: A review of recent developments," *Precis. Agricult.*, vol. 22, no. 6, pp. 2053–2091, Dec. 2021.

[2] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, and M. Zuo, "CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, Mar. 2019.

[3] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[4] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[5] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2018.

[6] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[7] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.

[8] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042609.

[9] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," 2022, *arXiv:2206.13188*.

[10] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2020.

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

[12] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6707–6717.

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

[14] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 28, 2022, doi: 10.1109/TGRS.2022.3194732.

[15] O. Manas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[16] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10181–10190.

[17] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[18] 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 Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[19] A. M. Swope, X. H. Rudelis, and K. T. Story, "Representation learning for remote sensing: An unsupervised sensor fusion approach," 2021, *arXiv:2108.05094*.

[20] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405516.

[21] Y. Chen and L. Bruzzone, "Self-supervised SAR-optical data fusion of Sentinel-1/-2 images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406011.

[22] 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," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102651.

[23] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," 2022, *arXiv:2207.08051*.

[24] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1422–1431.

[25] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315.

[26] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[27] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, early access, May 25, 2022, doi: 10.1109/TGRS.2022.3176603.

[28] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," 2022, *arXiv:2209.01206*.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

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

[31] Q. Zhou, C. Yu, H. Luo, Z. Wang, and H. Li, "MimCo: Masked image modeling pre-training with contrastive teacher," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4487–4495.

[32] Z. Huang et al., "Contrastive masked autoencoders are stronger vision learners," 2022, *arXiv:2207.13532*.

[33] C. Tao et al., "Siamese image modeling for self-supervised vision representation learning," 2022, *arXiv:2206.01204*.

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

[35] W. Li, W. Wei, and L. Zhang, "GSDet: Object detection in aerial images based on scale reasoning," *IEEE Trans. Image Process.*, vol. 30, pp. 4599–4609, 2021.

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

[37] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," 2022, *arXiv:2205.13137*.

[38] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3713–3722.

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

[40] T. Xiao, X. Wang, A. A. Efros, and T. Darrell, "What should not be contrastive in contrastive learning," 2020, *arXiv:2008.05659*.

[41] L. Jing, J. Zhu, and Y. LeCun, "Masked Siamese ConvNets," 2022, *arXiv:2206.07700*.

[42] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2021, pp. 1182–1191.

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

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

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

[46] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

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

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

[49] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9640–9649.

[50] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3024–3033.

[51] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16684–16693.

[52] P. O. O. Pinheiro, A. Almahairi, R. Benmalek, F. Golemo, and A. C. Courville, "Unsupervised learning of dense visual representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4489–4500.

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

[54] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "RegionCL: Can simple region swapping contribute to contrastive learning?" 2021, *arXiv:2111.12309*.

[55] T. Xiao, C. J. Reed, X. Wang, K. Keutzer, and T. Darrell, "Region similarity representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10539–10548.

[56] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Revisiting contrastive methods for unsupervised learning of visual representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 16238–16250.

[57] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

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

[59] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[60] X. Dong et al., "PeCo: Perceptual codebook for BERT pre-training of vision transformers," 2021, *arXiv:2111.12710*.

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

[62] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14668–14678.

[63] Y. Wei et al., "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation," 2022, *arXiv:2205.14141*.

[64] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, and R. Girshick, "Benchmarking detection transfer learning with vision transformers," 2021, *arXiv:2111.11429*.

[65] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, and Z. Yuan, "Designing BERT for convolutional networks: Sparse and hierarchical masked modeling," 2023, *arXiv:2301.03580*.

[66] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis.—ECCV*. Glasgow, U.K.: Springer, Aug. 2020, pp. 776–794.

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

[68] S. Saha, P. Ebel, and X. Zhu, "Self-supervised multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4405710.

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

[70] S. Li et al., "Architecture-agnostic masked image modeling—From ViT back to CNN," 2022, *arXiv:2205.13943*.

[71] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13919–13929.

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

[73] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[74] MMS Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[75] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan, "Adan: Adaptive Nesterov momentum algorithm for faster optimizing deep models," 2022, *arXiv:2208.06677*.

[76] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

[77] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[78] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[79] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.

[80] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[81] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[82] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[83] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.

[84] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[85] Y. Zhou et al., "MMRotate: A rotated object detection benchmark using PyTorch," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 7331–7334.
[86] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3520–3529.
[87] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, May 2018.
[88] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[89] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[90] MMS Contributors. (2021). *MMSelfSup: OpenMMLab Self-Supervised Learning Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmselfsup
[91] A. C. Li, A. A. Efros, and D. Pathak, "Understanding collapse in non-contrastive Siamese representation learning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 490–505.
[92] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky, "Sparse convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 806–814.
[93] G. Liu et al., "Partial convolution based padding," 2018, *arXiv:1811.11718*.
[94] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2002, pp. 1–8.
[95] M. Assran et al., "Masked Siamese networks for label-efficient learning," in *Proc. 17th Eur. Conf. Comput. Vis.—ECCV*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 456–473.

Dilxat Muhtar received the B.S. degree in geographic information science from Nanjing University, Nanjing, China, in 2022, where he is currently pursuing the M.S. degree in cartography and geographic information system.

His research interests include self-supervised and transfer learning for remote sensing.

Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information systems and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with the Informatics Institute, University of Missouri, Columbia, MO, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University, where he is currently an Associate Professor with the Department of Geographic Information Science. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.

Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. He has authored four books and over 160 articles. His research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover changes.

Zhenshi Li received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2019, and the M.S. degree in cartography and geographic information system from Nanjing University, Nanjing, in 2022, where he is currently pursuing the Ph.D. degree in cartography and geographic information system.

His research interests include semantic segmentation and weakly supervised deep learning for remote sensing.

Feng Gu received the B.S. degree in remote sensing science and technology from the China University of Mining and Technology, Beijing, China, in 2022. He is currently pursuing the M.S. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China.

His research interests include change detection and deep learning for remote sensing.