

Generic Knowledge Boosted Pre-training For Remote Sensing Images

Ziyue Huang, Mingming Zhang, Yuan Gong, Qingjie Liu, *Member, IEEE*, and Yunhong Wang, *Fellow, IEEE*

Abstract—Deep learning models are essential for scene classification, change detection, land cover segmentation, and other remote sensing image understanding tasks. Most backbones of existing remote sensing deep learning models are typically initialized by pre-trained weights obtained from ImageNet pre-training (IMP). However, domain gaps exist between remote sensing images and natural images (e.g., ImageNet), making deep learning models initialized by pre-trained weights of IMP perform poorly for remote sensing image understanding. Although some pre-training methods are studied in the remote sensing community, current remote sensing pre-training methods face the problem of vague generalization by only using remote sensing images. In this paper, we propose a novel remote sensing pre-training framework, Generic Knowledge Boosted Remote Sensing Pre-training (GeRSP), to learn robust representations from remote sensing and natural images for remote sensing understanding tasks. GeRSP contains two pre-training branches: (1) A self-supervised pre-training branch is adopted to learn domain-related representations from unlabeled remote sensing images. (2) A supervised pre-training branch is integrated into GeRSP for general knowledge learning from labeled natural images. Moreover, GeRSP combines two pre-training branches using a teacher-student architecture to simultaneously learn representations with general and special knowledge, which generates a powerful pre-trained model for deep learning model initialization. Finally, we evaluate GeRSP and other remote sensing pre-training methods on three downstream tasks, *i.e.*, object detection, semantic segmentation, and scene classification. The extensive experimental results consistently demonstrate that GeRSP can effectively learn robust representations in a unified manner, improving the performance of remote sensing downstream tasks. Code and pre-trained models: <https://github.com/floatingstarZ/GeRSP>.

Index Terms—Remote sensing image, pre-training, self-supervised learning

I. INTRODUCTION

DEEP learning models have been widely used in remote sensing (RS) image understanding tasks, such as detection [1], segmentation [2], and scene classification [3]. Most of these interpretation models are initialized with ImageNet [4] pre-trained weights. Although it has been proved that ImageNet pre-trained models generalize well to the RS interpretation tasks, domain gaps still exist between RS and natural images due to different capture views, image resolutions, and

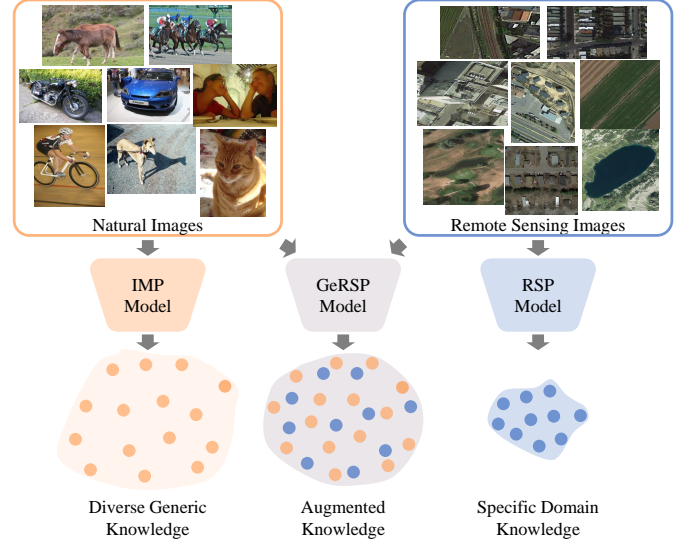


Fig. 1. RS images encompass a wealth of domain-specific knowledge, whereas natural images offer a broader range of diverse generic image knowledge. The motivation of the GeRSP is to enhance the generalization performance of RSP by leveraging the diversity present in natural images.

object appearances, which impedes the RS image understanding performance. This puts forward an urgent requirement for Remote Sensing Pre-training (RSP) techniques [5].

Most RSP methods draw inspiration from general pre-training methodologies, such as MoCo [6], SimCLR [7], and MAE [8], which can be categorized as supervised and self-supervised paradigms. The supervised pre-training paradigm necessitates extensive labelled data for achieving effective pre-trained weights. Nonetheless, acquiring such datasets is costly and demands substantial professional expertise. Recently, the self-supervised pre-training paradigm has been receiving much attention from both the computer vision [6] and the remote sensing community [9]. This paradigm could acquire essential visual representations by constructing label-independent pretext tasks such as instance discrimination [9]–[11], spatial coherence [12], and masked image modeling [13]. Contrastive learning via instance discrimination is the most popular self-supervised method. The key to contrastive learning is to construct positive and negative example pairs. Manas et al. [9] explored utilizing the seasonal variation to construct the seasonal contrastive pairs. Ayush et al. [11] utilized geographical location information and time changes to establish the contrastive learning objective. Liu et al. [10] introduced the consistency of SAR and optical image to realize

This work was supported by the National Natural Science Foundation of China under Grant 62176017. (Corresponding author: Qingjie Liu)

Ziyue Huang, Mingming Zhang, Qingjie Liu, and Yunhong Wang are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China (e-mail: ziyuehuang@buaa.edu.cn; sara_@buaa.edu.cn; qingjie.liu@buaa.edu.cn; yhwang@buaa.edu.cn).

Yuan Gong are with School of Software and Microelectronics, Peking University, Beijing 100871, China. (e-mail: alangy@stu.pku.edu.cn).

contrast learning without negative samples.

Pre-trained models are believed to be able to boost the downstream tasks; however, recent studies [5] have shown that general pre-training methodology may not be suitable for RSP. The suboptimal performance is observed when applying the pre-trained model to the downstream tasks, such as the segmentation task compared to IMP models [5]. The reason might be that IMP models can obtain diverse low-mid level features from natural images [14]. These features play a pivotal role in dense prediction tasks, including semantic segmentation and object detection, possessing stronger transferability [14] and serving as general knowledge. Our experiments and previous research [15] consistently demonstrate that the IMP model can serve as a robust baseline for various remote sensing downstream tasks. Additionally, the differences in semantics between natural and remote sensing images can prevent semantic over-fitting [16], [17], further enhancing transferability.

In contrast, RS images predominantly emphasize objects or scenes on the earth's surface, such as cars, houses, lakes, and airports. Additionally, they are constrained by the bird's-eye view perspective and sensor resolution, thereby restricting the diversity of scenes, perspectives, and detailed object information [18], thus impeding the learning of diverse low-mid level features. Simply scaling up the dataset does not bring more information enrichment. Furthermore, high-level semantic feature alignment in contrastive learning within RSP neglects the learning of low-mid level features [19], thus diminishing performance on dense prediction tasks. IMP can efficiently acquire these features, prompting us to explore simultaneous pre-training using both RS and natural images.

In this study, we tackle this challenge by leveraging the rich knowledge of natural images to boost RSP. Introducing natural images to the RS domain enriches the feature space of RS pre-training models. To achieve this goal, training an IMP model on the RS images to acquire domain-specific knowledge [20] is one straightforward approach. However, this multi-stage training procedure makes the model tend to forget the knowledge gained from the IMP phase, which hinders the pre-trained model from achieving satisfactory results on downstream tasks, as confirmed by our experiments.

To compensate for the shortcomings of the existing pre-training paradigms, we propose **Generic Knowledge Boosted Remote Sensing Pre-training (GeRSP)** to obtain generic and remote sensing domain knowledge, as shown in Fig. 1. In particular, a supervised pre-training branch on natural images is used to obtain general knowledge for downstream tasks. To capture RS domain knowledge, a self-supervised pre-training branch on RS images is co-operated with the supervised pre-training branch on natural images so that the proposed GeRSP simultaneously learns domain-related features from RS images.

In summary, our contributions include the following:

- 1) A novel remote sensing pre-training framework, GeRSP, is proposed to learn robust representations for RS understanding tasks. GeRSP uses a teacher-student architecture to simultaneously learn representations with general and domain knowledge.

- 2) GeRSP contains supervised pre-training and self-supervised pre-training stages: (1) The self-supervised pre-training stage learns domain-related features from unlabeled RS images. (2) The supervised pre-training stage is integrated for general knowledge learning from labeled natural images.
- 3) Three RS downstream tasks are evaluated to compare GeRSP with other pre-training methods, including object detection, semantic segmentation, and scene classification. The experimental results consistently demonstrate that GeRSP effectively improves the performance of remote sensing downstream tasks. Additionally, we perform a visual analysis to further evaluate the effectiveness of GeRSP and its impact on downstream tasks.

II. RELATED WORK

A. Pre-training Methods

Motivated by the observation that humans can leverage existing knowledge to solve new problems [21], transfer learning has been proposed as a solution to this challenge. Transfer learning allows models to benefit from pre-existing knowledge by leveraging pre-trained parameters. This idea of parameter transfer has been widely used in computer vision tasks. There are two general pre-training methods based on whether labeled data is required during the training process: supervised pre-training and self-supervised pre-training.

Supervised pre-training is a practical approach for obtaining pre-trained models. Models with different architectures, such as ResNet [22], ViT [23], and Swin Transformer [24], have been successfully pre-trained on large-scale image datasets like ImageNet [4]. Benefiting from the general visual knowledge obtained from large-scale image datasets, downstream models only require a small amount of task-specific data to perform well. However, collecting and annotating large-scale datasets are still time-consuming and expensive in the real world, which limits the development of supervised pre-training methods. Semi-supervised learning [25], [26] can effectively leverage unlabeled data; however, it is often oriented towards specific task improvements and is less employed for pre-training.

Self-supervised pre-training methods are proposed to address these issues. These methods construct pretext tasks that leverage intrinsic properties of images, thereby facilitating effective feature extraction [27]. The constraint of pretext tasks compels the neural network to extract pertinent image information and generate good visual representations. The self-supervised pretext tasks encompass diverse methodologies, such as image generation [28], image inpainting [8], [29], jigsaw puzzles [30], and image colorization [31]. These tasks are designed based on the inherent structure and characteristics of the images themselves.

Jing et al. [27] summarized the pretext tasks into four categories: generation-based, context-based, free semantic label-based, and cross-modal-based. Generation-based methods encompass tasks such as image generation and image inpainting. Among them, methods based on image generation are primarily utilized for generating more realistic images [32] or expanding datasets [33] rather than focusing on acquiring

a robust feature extractor. Pre-training methods based on image inpainting have gained significant traction within self-supervised pre-training methods, such as MAE [8] and CAE [34]. These methods involve masking specific regions within an image and requiring the network to predict the content of the masked areas. These approaches can effectively benefit downstream tasks that demand semantic understanding by necessitating the network's comprehension of the remaining image blocks and enabling image reconstruction based on contextual cues.

The context-based pretext tasks primarily rely on semantic consistency or spatial context cues within the image as the supervisory signal. The pre-training method based on contrastive learning [6], [7], [35]–[37] has gained significant traction among these tasks. The core idea behind contrastive learning is to minimize the distance between differently augmented views of the same image while maximizing the dissimilarity between unrelated images [6], [7]. In our work, we incorporate contrastive learning into the remote sensing pre-training. Specifically, we employ MoCo technique [6] due to its widespread popularity, code base availability, and results reproducibility.

B. Remote Sensing Pre-training Methods

Similar to other computer vision domains, ImageNet pre-trained models have demonstrated remarkable success in RS image recognition tasks [1], [38]–[43]. Nevertheless, challenges such as the domain gap between natural scenes and RS scenes and limitations in generalization persist in RS pre-training methods. Tong et al. [44] presented a transfer learning approach for scene classification. Initially, the model is pre-trained with a well-annotated large-scale dataset. Subsequently, a semi-supervised transfer learning method is employed on the pre-trained model to achieve pixel-level classification. Building upon this methodology, Long et al. [45] introduced Million-AID, a substantial benchmark dataset comprising one million instances designed explicitly for scene classification. The Million-AID dataset contains globally distributed high spatial resolution RS images in 51 scene categories. After that, a hierarchical multi-task learning framework [46] for pixel-level scene classification demonstrates the strong generalization ability of Million-AID. In a complementary study, Wang et al. [5] conducted extensive experiments to assess the generalization performance of Million-AID pre-training models across multiple downstream tasks. The models employed encompass CNN-based architectures such as ResNet [22], as well as Transformer-based approaches including Swin Transformer [24] and ViTAEv2 [47]. The experimental results highlight that the pre-training models significantly enhance performance in various downstream tasks compared to the ImageNet pre-trained models. However, they also found that only using RS data may lack crucial information for detection and segmentation [5], which inspired our investigation.

Self-supervised pre-training methods have been extensively studied in RS research communities [17]. Numerous studies have employed pre-training methods to enhance the performance of specific RS tasks, such as hyperspectral imagery

classification [48]–[51], synthetic aperture radar (SAR) target recognition [52], and change detection [53], [54]. Some studies leveraged the geographic information associated with RS images to achieve more effective self-supervised pre-training. Jean et al. [12] proposed Tile2Vec, an unsupervised representation learning method inspired by Word2Vec [55]. Tile2Vec assumes that geographically proximate tiles exhibit semantic similarity. Based on this assumption, they employed metric learning for unsupervised tiles learning.

Jung et al. [56] proposed a contrastive learning method based on the SimCLR [7] framework. The method uses the idea of Tile2Vec [12], utilizing three neighbor tiles to obtain the smooth representation for positive samples. Ayush et al. [11] exploited the revisiting characteristics of satellites to construct spatial-aligned image pairs at different times, enabling informative learning. Additionally, a geo-location pretext task was incorporated during training to enhance the representation learning of RS images. SeCo [9] combined temporal variation with other augmentation techniques to enable multi-augmentation contrastive learning. This approach yields representations encompassing time-varying and invariant features, offering advantages for downstream tasks. Scheibenreif et al. [57] addressed land cover classification and segmentation tasks by employing SimCLR [7] with paired satellite data obtained from optical Sentinel-2 and SAR Sentinel-1 sensors. However, it is essential to note that these studies require additional meta-information, including geographical location and time, which imposes more stringent restrictions on their practical application.

While there has been considerable research on supervised and self-supervised pre-training models in the remote sensing domain, the comprehensive exploration of a general RS model with extensive generalization performance still needs to be improved. Risojević et al. [58] proposed a domain-adaptive pre-training method that re-trains an ImageNet pre-trained model using MLRSNet [59] dataset. Their approach outperformed models pre-trained solely on either ImageNet [4] or MLRSNet [59] in scene classification tasks. Likewise, Zhang et al. [20] introduced the ConSecutive pre-training (CSPT) method for RSP. In CPST, the model first performs self-supervised learning on natural scene images through masked image modeling [8] pretext task and then conducts self-supervised training on task-related unlabeled RS data. CSPT aims to bridge the domain gap and transfer knowledge from the natural image domain to the RS domain. However, it does not guarantee the preservation of general features learned from natural images during the second-stage self-supervised learning [60]. Additionally, the substantial size of ViT-based models restricts their practicality in certain situations.

Our approach aims to improve the extraction of low-level general knowledge in RSP by incorporating supervised training with natural images, thereby enhancing spatial information perception capabilities. The motivation of TOV [18] is close to ours. They freeze natural image pre-trained model's shallow and middle layers and subsequently train on the RS dataset. This two-stage approach prevents the forgetting of general knowledge and achieves adaptability to remote sensing images. Compared with TOV, our method employs a joint training

framework and facilitates the adaptation of even the shallow layer to remote sensing images. Furthermore, our approach demonstrates that simply introducing supervised learning with ImageNet [4] can effectively acquire general knowledge without requiring a redundant multi-stage training strategy.

III. GENERIC KNOWLEDGE BOOSTED PRE-TRAINING

A. Overview

Recently, remote sensing pre-training has established its effectiveness in extracting information from remote sensing data, with subsequent applicability to a broad spectrum of downstream tasks [9]–[11], [51]. Simultaneously, ImageNet pre-training persists as a strong baseline due to its adeptness in assimilating transferrable knowledge from extensive large-scale natural data [5], [15]. Nevertheless, optimizing the concurrent utilization of both datasets for acquiring more robust transferable features remains a topic of ongoing research. To fill this gap, we propose GeRSP, a novel pre-training framework incorporating IMP and RSP. GeRSP learns the fundamental knowledge from IMP and captures specific knowledge tailored explicitly to remote sensing imagery. The downstream tasks employ GeRSP pre-trained weights as initialization weights for fine-tuning. This two-stage training methodology is illustrated in Fig. 2.

As illustrated in Fig. 2, GeRSP utilizes a teacher-student architecture to enable collaborative learning. It comprises two learning processes: natural image auxiliary learning (NIAL) on labeled natural images and remote sensing contrastive learning (RSCL) on unlabeled RS images. RSCL simultaneously trains both the teacher network and the student network. During the training process, the teacher network's parameters are updated using the exponential moving average of the student network parameters. Conversely, in the case of NIAL, the training is exclusively focused on the student network. Ultimately, the student network serves as a pre-trained model for downstream tasks. In each iteration of GeRSP, an equal number of natural and remote sensing images are sampled from their respective datasets.

During RSCL, the RS image is subjected to two distinct augmentation strategies, denoted as t and t' , to form positive pairs. Subsequently, the RS image pairs are fed into the teacher and student networks to extract their features. The features are then aggregated using global average pooling (GAP) and projected by independent *projectors* that consists of two fully-connected layers, yielding the features z^{k+} and z^q . Finally, the student network is optimized by minimizing the InfoNCE [61] loss function computed over the positive sample pairs (z^{k+} , z^q) and negative samples z^{k-} for contrastive learning, where z^{k-} are retrieved from a *dynamic queue* that is actively maintained by the teacher network. Further details will be discussed in the subsequent subsection.

Concurrently, labeled natural images from the same batch are utilized for NIAL to ensure adaptability to a wide range of tasks. Data augmentation is also applied to unlabeled natural images, resulting in the augmented images denoted as I_s . Then, the augmented images are inputted into the student network, followed by a GAP operation and an *predictor*.

The *predictor*, distinct from the *projector* used in RSCL, helps alleviate conflicts between NIAL and RSCL. The cross-entropy loss is then used to optimize the student network. After each optimization iteration, both the parameters of the teacher network and the dynamic queue are updated. Subsequent sections will detail each component, including data augmentation (Sect. III-B), the backbone network, projectors, and predictors for feature extraction in NIAL and RSCL (Sect. III-C), the loss functions (Sect. III-D), and the update strategies for parameters and the dynamic queue (Sect. III-E).

B. Data Augmentation

Data augmentation introduces variability in the input images, aiming at increasing the difficulty of the contrastive learning pretext task. Thus, the pre-trained model can acquire more meaningful features from augmented images rather than merely memorizing the input images [6], [7], [37]. Therefore, the proposed GeRSP framework adopts the strong augmentation strategy described in [62] for both RSCL and NIAL, enhancing the transferability of representations pre-trained on remote sensing (RS) images. Besides, the pre-trained model becomes more relevant to RSCL by employing strong data augmentation during NIAL, enabling more efficient feature learning.

The pipeline for strong augmentation is depicted in Fig. 3. Initially, images are cropped from the corresponding original image with a ratio ranging from 0.2 to 1 and then resized to a scale of 224. Subsequently, color jitter is applied with a probability of 0.8, employing brightness, contrast, saturation, and hue factors of 0.4, 0.4, 0.4, and 0.16, respectively. Moreover, images are converted to gray-scale with a probability of 0.2. Finally, images are flipped with a probability of 0.5 and subjected to Gaussian blur with a probability of 0.5.

C. Feature Extractor

Considering the practicality of the pre-trained model [9], we select ResNet50 [22] as the backbone network for pre-training. During RSCL, we employ shuffling batch normalization (BN) [6] in the backbone to eliminate the correlation between BN parameters and the mini-batch, which can effectively avoid information leakage. Specifically, the order of samples within the mini-batch is shuffled before input to the backbone network. The original order is recorded to be restored during contrastive learning.

After the backbone network has extracted image features, Global Average Pooling (GAP) is applied to reduce the dimensionality and obtain features in \mathbb{R}^{2048} . As illustrated in Fig. 2, the non-linear *projector* and *predictor* are employed to process these features further.

The *projector* consists of two fully connected layers with ReLU activation. It has a hidden dimension of 2048 and an output dimension of 128, yielding the features z^{k+} and z^q . On the other hand, the *predictor* is a single fully connected layer that maps features to logits z^s for classification. The introduction of non-linearity in the *projector* can prevent dimension collapse [63] and enhance the performance of multitask learning. Since RSCL aims to achieve feature invariance by strong

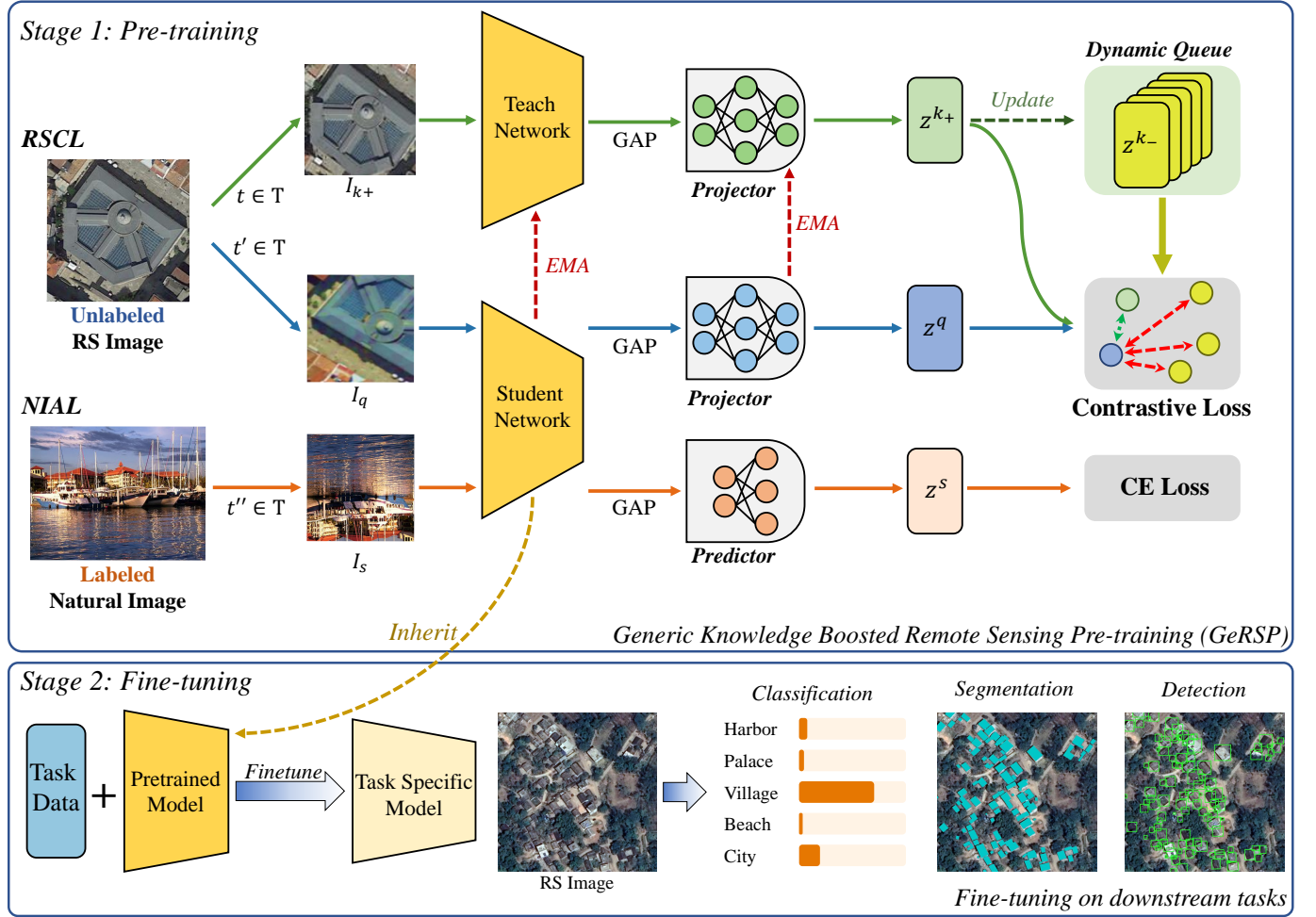


Fig. 2. The overall framework of our proposed Generic Knowledge Boosted Remote Sensing Pre-training (GeRSP). GeRSP integrates two learning processes: natural image auxiliary learning (NIAL) on labeled natural images and remote sensing contrastive learning (RSCL) on unlabeled RS images. NIAL utilizes labeled natural images for training. NIAL involves training the model using labeled natural images, while RSCL adopts a contrastive learning approach. The trained model is subsequently fine-tuned on various downstream tasks using task-specific data.

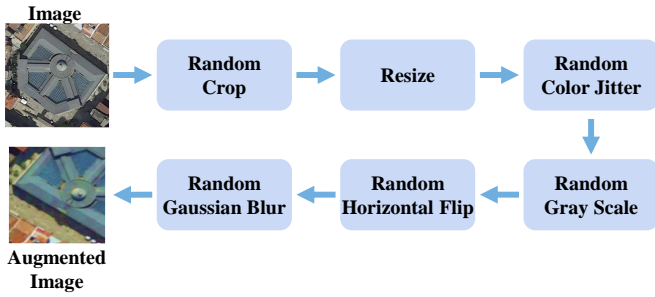


Fig. 3. Data Augmentation Pipeline for pre-training.

augmentation in input images, it can adversely affect task-specific characteristics learning such as image color, contrast, and position. Therefore, the *projector* is crucial for RSCL and can learn semantic invariance features while preserving more information for the backbone network. Additionally, the *projector* serves as a bridge to close the information gap between NIAL and RSCL.

D. Loss Function

The loss function for GeRSP consists of two terms: the cross-entropy loss for NIAL, denoted as L_{CE} , and the contrastive loss for RSCL, denoted as L_{CT} . For L_{CE} , we firstly normalize z^s by using the softmax function, which yields a probability vector $\mathbf{p} = (p_1, p_2, \dots, p_K)$, where K represents the number of categories. Then, the cross-entropy loss L_{CE} is computed by using the following equation:

$$L_{CE}(\mathbf{p}, \mathbf{y}) = - \sum_{i=1}^K y_i \log(p_i) \quad (1)$$

where y_i represents the ground truth label for the corresponding category. For L_{CT} , RSCL utilizes the InfoNCE loss as the contrastive loss function. Considering features z^{k+} and z^q obtained from the *projectors*, as well as features derived from the negative queue comprising historical features $\mathbf{Z}^{k-} = z^{k-}$, the calculation of the contrastive loss [61] is as follows:

$$L_{CT}(z^q, \mathbf{Z}^{k-}) = - \log \frac{\exp(s_{q,k+}/\tau)}{\exp(s_{q,k+}/\tau) + \sum_{k-} \exp(s_{q,k-}/\tau)} \quad (2)$$

where τ is the temperature parameter that controls the intensity of contrast, $s_{q,k} = z^q \cdot z^k / ||z^q|| ||z^k||$ is the cosine similarity between features. The cross-entropy loss used in NIAL is actually similar to the InfoNCE loss [6] employed in RSCL, allowing for the joint training of NIAL and RSCL. The total loss is defined as below:

$$L_{total} = L_{ct} + \alpha L_{ce} \quad (3)$$

where coefficient α is employed to strike a balance between NIAL and RSCL, with a default value of 1.

E. Parameter & Queue Update

The teacher network employs momentum update [6] on the network parameters to facilitate a stable training process. Denoting parameters of the teacher network and the student network as W_t and W_s respectively, the update rule for W_t is as follows:

$$W_t = mW_t + (1 - m)W_s \quad (4)$$

where the momentum coefficient m is set to 0.996. The dynamic queue is implemented as a First-In-First-Out (FIFO) queue with a maximum capacity of 65,536. It is updated after each iteration and serves as a storage for features generated by the teacher network. Given a batch size of 128, it takes approximately 500 iterations to complete a full update of the queue.

IV. PRE-TRAINING

A. Implementation Details

To compare the effectiveness of our proposed GeRSP with other pre-training methods, we use ResNet50 [22] as the backbone and select several pre-training methods, such as supervised pre-training and self-supervised pre-training. To validate the superiority of GeRSP over RSP, we choose SeCo [9], GeoAware [11], CACO [64], TOV [18], MoCo [6], and MoCov2 [35] as comparative methods. To demonstrate the advantages of GeRSP over IMP, we compare it with supervised pre-training and MoCo [6]. Furthermore, we conduct training using MoCo on mixed data as a stronger baseline to show that GeRSP can better utilize mixed data.

The unlabeled RS image dataset Million-AID (MAID) [45] and the labeled natural image dataset ImageNet [4] are utilized for pre-training. MAID [45] is a large-scale remote sensing scene classification dataset consisting of 1,000,848 images with 51 scene categories. MAID is collected from Google Earth with broad resolutions, ranging from 0.5 to 153 m per pixel.

All pre-training methods are trained with the stochastic gradient descent (SGD) optimizer for 100 epochs, with initial values of 0.05 for the learning rate, 0.90 for weight decay, and 0.00005 for momentum. The learning rate for GeRSP is optimized using the cosine annealing scheduler with restarts [65]:

$$lr_{cur} = lr_{min} + \frac{1}{2}(lr_{max} - lr_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \quad (5)$$

The minimum learning rate, denoted as lr_{min} , is set to 0.10, while the maximum learning rate, denoted as lr_{max} , is set

to 0.01. During a single round of training, T_{max} is defined as the maximum number of epochs, which is set to 20. T_{cur} represents the current epoch number within the current round. Once T_{cur} reaches T_{max} , it is reset to 0 to initiate the subsequent round, and lr_{cur} denotes the current learning rate.

For MoCo and MoCov2, learning rates are optimized using the step linear scheduler with a step size of 30 and a learning rate decay of 0.1. Pre-training methods are trained in a distributed manner, utilizing data parallelism across 8 RTX-2080 GPUs, with a total batch size of 128. After the pre-training stage, irrelevant components, including the predictor, projector, and the teacher network, are removed, and the student network acts as the pre-trained model for downstream tasks. For previous RSP methods (SeCo, GeoAware, CACO, and TOV), we directly download their pre-trained parameters and evaluate them on downstream tasks to objectively compare the performance of each method.

V. FINETUNING ON DOWNSTREAM TASKS

To evaluate the effectiveness of pre-training methods, we finetune different pre-trained models on three downstream tasks: scene classification, object detection, and semantic segmentation. All pre-trained models in these tasks adhere to the same configurations and hyper-parameter settings.

A. Scene Classification

To align pre-trained models with the requirements of classification tasks, they undergo augmentation by adding an average pooling layer and a single fully connected layer at the final stage. The models are optimized by mini-batch stochastic gradient descent with momentum (SGDM) algorithm for 100 epochs, with a batch size of 64. For SGDM, the initial learning rate, weight decay, and momentum are set to 0.01, 0.0001, and 0.9, respectively. A linear-step scheduler is utilized to ensure training stability, employing step values of [30, 60, 90] and a decay ratio of 0.1. For image pre-processing, the images are scaled to 224×224 pixels. Additionally, random flipping is applied with a probability of 0.5, meaning each image has a 50% chance of being flipped. Two scene classification datasets are employed for validation:

- EuroSAT [66]: This dataset consists of Sentinel-2 satellite images captured over Europe. It consists of 27,000 images belonging to 10 distinct categories. Each class comprises approximately 2,000 to 3,000 images, with a resolution of 64×64 pixels.
- NWPU-RESISC45 [40]: This dataset, developed by Northwestern Polytechnical University (NWPU) for Remote Sensing Image Scene Classification (RESISC), includes 31,500 images. The dataset covers 45 scene categories, with 700 images per category. The images have spatial resolutions ranging from 30 to 0.2 meters, and their size is 256×256 pixels.

The evaluation metric employed for classification accuracy is the Top-1 accuracy. Following [5], we employ 20% of the data for training and reserved 80% for testing. The models are repeatedly trained and evaluated five times at each setting. The average value μ and standard deviation σ of the results across various trials were documented as $\mu \pm \sigma$.

TABLE I
THE COMPARISON OF DIFFERENT PRE-TRAINING METHODS ON SCENE CLASSIFICATION TASK.

Pre-training	Dataset	EuroSAT Top-1	NWPU-RESISC45 Top-1
IMP	ImageNet	97.84 ± 0.04	92.48 ± 0.13
MoCo-IN [6]	ImageNet	97.55 ± 0.29	90.77 ± 0.19
MoCo-MAID [6]	MAID	97.52 ± 0.10	91.19 ± 0.15
MoCo-IN-MAID [6]	ImageNet + MAID	97.47 ± 0.10	91.15 ± 0.20
MoCov2 [35]	MAID	97.52 ± 0.07	91.28 ± 0.04
SeCo [9]	1M Sentinel-2	97.74 ± 0.25	90.40 ± 0.15
Geo-Aware [11]	GeoImageNet	97.87 ± 0.11	92.37 ± 0.19
CACO [64]	CACO 1M	97.67 ± 0.09	90.81 ± 0.18
TOV [18]	TOV-NI + TOV-RS	97.80 ± 0.06	92.59 ± 0.21
GeRSP	ImageNet + MAID	97.87 ± 0.15	92.67 ± 0.16
GeRSP-200	ImageNet + MAID	97.87 ± 0.10	92.74 ± 0.09

B. Object Detection

To assess the performance of pre-trained models, we employ three detection methods, e.g., Faster R-CNN [67], RetinaNet [68], and Dy-Head [69]. These detection methods are implemented using MMDetection [70] toolbox. The pre-trained backbone parameters are replaced with the parameters obtained from our experiments. The detectors are trained using the SGDM optimizer with a learning rate of 0.001, momentum of 0.90, and weight decay of 0.0001. The learning rate is reduced by a factor of 10 at 16 epochs and 22 epochs, respectively. The experiments are conducted on two GPUs with a batch size of 4 over 24 epochs. During training and testing, images are resized to 800×800 pixels. Only random flipping is applied during training. For testing, non-maximum suppression (NMS) with an intersection over union (IoU) threshold of 0.3 is employed to remove duplicated detections and retains a maximum of 1,000 detections. The widely used DOTA [71] and DIOR [72] datasets are selected in our experiments:

- DOTA [71]: This dataset comprises 2,806 images with a total of 188,282 instances belonging to 15 different categories. The images have varying sizes ranging from 800 × 800 pixels to 4,000×4,000 pixels. In our experiment, we utilize horizontal bounding box (HBB) annotations, which require minimal modifications to enable Faster R-CNN for RS object detection. The images are cropped into patches of size 800×800 pixels with a stride of 640. The performance on the cropped validation set is reported.
- DIOR [72]: The DIOR dataset consists of 23,463 images with a total of 192,472 instances. It covers 20 object categories and offers a significantly more diverse distribution of instances and finer classification than other datasets.

During the evaluation, we compare the Average Precision (AP) of each category, and the mean Average Precision (mAP) is also considered. Specifically, we adopt the evaluation protocol of COCO [73] and use the AP calculated under the IoU threshold of 0.5 as the evaluation criterion.

C. Segmentation

Two classic segmentation methods (i.e., PSANet [74] and DeepLabV3+ [75]) are chosen in our experiments. The two

segmentation models are all trained in 80,000 iterations with a batch size of 4, using the SGD algorithm with a learning rate of 0.01, weight decay of 0.0005, and momentum term of 0.9. During training, images are resized to 2,048×512 pixels, randomly cropped to 512×512 patches, and randomly flipped along the horizontal axis. For more accurate segmentation results, images are resized to 1,024×1,024 during testing. High spatial resolution land-cover semantic segmentation is selected as the segmentation task on land-cover dataset LoveDA.

- LoveDA [76]: The dataset contains 5,987 satellite images with 166,768 annotated objects from three cities: Nanjing, Changzhou, and Wuhan. LoveDA covers 536.15 km^2 and each image includes multi-scale objects, complex background, and inconsistent class distributions. There are 2,522 and 1,669 images in the training and validation sets, respectively. The typical resolution of images in the dataset is 1,024×1,024.

The mean IoU (mIoU) is chosen as the metric for evaluation.

VI. EXPERIMENTS AND ANALYSIS

A. Results

We selected ImageNet supervised pre-training (IMP), MoCov1 [6], MoCov2 [35], SeCo [9], GeoAware [11], CACO [64], and TOV [18] as comparison pre-training methods. ImageNet supervised pre-training and MoCov1 serve as the baseline for GeRSP. To illustrate the impact of the dataset on the pre-training method in downstream tasks, we pre-train the backbone network with MoCo on ImageNet, MAID, and ImageNet+MAID datasets, respectively. The obtained pre-trained models are denoted as MoCo-IN, MoCo-MAID, and MoCo-IN-MAID, respectively. To demonstrate the sustained effectiveness of GeRSP in larger-scale training, we conduct training for 200 epochs, referred to as GeRSP-200. The results of semantic segmentation are shown in Table I, Table II, Table III and Table IV, respectively.

Overall Performance: Across all the obtained results, GeRSP consistently exhibited improvements. Notably, GeRSP delivers substantial enhancements in detection, attaining superior performance across all methods and datasets. Moreover, GeRSP demonstrates the capability to enhance both segmentation and classification endeavors. Compared to training from scratch, all pre-training methods prove efficacious in augmenting the performance of fine-tuning on downstream tasks, thereby emphasizing the indispensability of pre-training in remote sensing cognitive tasks.

Scene Classification: Table I reports the top-1 accuracy of our proposed GeRSP and other pre-training methods on EuroSAT and NWPU-RESISC45. As shown in Table I, GeRSP outperforms the other pre-training methods on the two datasets, showing its effectiveness. For MoCo-IN, MoCo-MAID, and MoCo-IN-MAID, GeRSP improves the top-1 accuracy by +2.0%, +1.6%, and +1.6% on NWPU-RESISC45, respectively. The experiments show domain gaps exist between RS images and natural images when comparing MoCo-IN with GeRSP. Moreover, pre-training only on RS images can lead to the lack of general features of objects, which reduces the generalization performance on downstream tasks when

TABLE II
FINE-TUNING RESULTS ON THE DIOR [72] OBJECT DETECTION TASK. WE USE THREE DETECTION METHODS TO VERIFY THE PRE-TRAINED MODELS.

Methods	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
<i>Faster R-CNN</i>																					
From Scratch	45.4	28.2	63.1	59.9	18.7	57.4	32.1	38.1	43.4	40.2	48.6	40.4	37.0	69.8	45.7	43.5	75.1	27.9	31.8	70.0	45.8
IMP	58.6	74.0	70.9	86.0	38.9	77.5	55.2	59.9	69.4	74.2	80.4	51.3	55.5	75.1	62.9	57.0	83.6	50.2	39.6	81.9	65.1
MoCo-IN [6]	53.8	73.5	67.6	86.1	41.8	75.4	55.3	61.9	73.2	74.9	77.2	57.6	57.3	77.2	67.7	55.4	84.6	56.2	39.2	82.3	65.9
MoCo-MAID [6]	53.8	73.3	69.4	85.9	40.6	75.1	56.3	65.7	73.1	76.1	77.7	57.4	57.3	77.1	64.9	57.4	84.5	53.4	40.3	83.1	66.1
MoCo-IN-MAID [6]	53.7	73.4	69.2	84.9	41.7	73.5	55.3	62.7	72.7	74.6	78.4	57.4	56.7	77.0	64.3	54.3	85.5	59.2	39.7	82.2	65.8
MoCov2 [35]	50.9	72.7	69.2	85.8	41.6	74.0	57.1	62.4	70.6	75.0	78.2	56.8	56.5	77.1	65.2	54.3	84.6	54.8	39.7	83.0	65.5
SeCo [9]	56.3	56.6	71.3	81.4	33.4	71.1	48.2	55.1	63.5	65.6	71.7	47.9	49.9	74.5	49.0	51.8	82.4	46.5	37.6	80.3	59.7
Geo-Aware [11]	52.7	68.4	73.3	83.7	36.6	75.5	50.4	63.0	68.1	69.5	75.9	47.7	53.5	74.8	60.2	57.6	85.0	46.8	39.6	80.7	63.1
CACO [64]	53.8	64.6	73.4	82.6	35.2	72.7	49.0	56.7	65.6	69.9	75.1	48.7	50.7	75.7	54.8	54.6	84.5	44.6	38.4	80.4	61.6
TOV [18]	63.9	72.4	77.1	85.0	40.3	77.3	57.3	66.1	71.9	75.1	80.0	51.1	55.4	76.9	63.9	59.8	85.9	49.0	41.1	82.8	66.7
GeRSP	61.5	76.2	69.1	86.7	42.3	77.6	59.9	62.4	72.8	77.6	78.4	54.5	58.4	75.3	68.4	57.2	86.3	55.1	41.0	82.3	67.1
GeRSP-200	63.6	78.1	69.4	87.1	44.0	78.5	62.9	67.0	76.0	76.8	79.6	54.7	58.7	75.3	60.3	57.3	86.4	56.3	41.3	80.4	67.8
<i>RetinaNet</i>																					
From Scratch	51.1	37.9	64.4	59.0	19.8	59.4	38.0	43.5	54.3	51.0	58.2	45.3	39.7	68.9	61.9	37.5	77.6	27.8	27.8	67.0	49.5
IMP	66.6	77.4	74.5	87.2	35.8	79.9	59.4	55.6	75.4	80.4	79.7	49.5	54.8	75.8	68.8	52.7	85.5	50.1	39.6	82.4	66.5
MoCo-IN [6]	57.1	70.8	71.0	85.3	35.8	74.8	54.2	51.4	70.4	74.8	74.9	49.9	54.4	77.1	59.6	54.5	84.8	46.5	39.7	82.5	63.5
MoCo-MAID [6]	57.7	72.7	71.4	86.0	35.0	75.8	52.6	55.3	73.6	75.1	75.6	51.4	53.8	76.5	64.6	53.4	85.6	43.4	40.8	82.8	64.2
MoCo-IN-MAID [6]	51.2	70.9	72.1	86.7	35.0	75.5	57.6	52.6	70.4	75.8	74.5	49.9	52.8	75.7	63.2	51.2	85.0	43.8	38.5	80.2	63.1
MoCov2 [35]	54.1	71.8	70.9	86.7	35.3	74.7	59.3	52.1	70.2	74.8	74.7	51.0	53.8	75.7	61.5	51.9	85.0	42.1	39.9	82.3	63.4
SeCo [9]	54.7	60.9	71.1	81.1	30.3	72.7	56.3	53.2	69.6	73.7	73.2	50.7	48.5	75.0	65.0	47.9	84.0	45.4	34.9	78.6	61.3
Geo-Aware [11]	60.4	67.9	71.8	83.6	31.3	77.4	51.9	55.1	71.5	72.2	75.1	46.6	50.9	73.9	62.3	51.6	85.7	42.4	37.6	78.1	62.4
CACO [64]	56.0	67.9	72.3	81.6	30.8	74.3	56.5	53.9	71.6	74.9	75.2	49.8	50.9	75.1	69.6	49.1	84.8	41.4	36.4	80.1	62.6
TOV [18]	63.4	74.3	75.1	83.1	37.5	77.2	60.9	59.8	73.9	81.6	77.7	51.8	54.3	76.4	68.6	56.1	86.1	47.5	40.1	82.2	66.3
GeRSP	70.3	76.9	75.3	87.2	39.4	80.3	61.4	59.3	77.3	80.5	76.8	50.5	56.0	76.8	64.3	54.3	86.8	50.3	42.2	83.8	67.5
GeRSP-200	70.0	77.7	74.5	87.8	41.2	80.3	63.6	61.9	76.9	80.3	78.7	51.9	56.8	78.1	72.6	57.0	87.7	50.0	43.0	84.6	68.7
<i>DyHead</i>																					
From Scratch	58.6	61.1	70.5	68.6	28.3	65.3	46.4	50.5	65.2	63.2	62.9	51.8	46.5	77.6	66.5	47.5	78.9	50.1	35.4	75.7	58.5
IMP	63.8	82.6	76.4	86.0	43.2	79.5	65.0	67.5	78.7	80.2	79.0	57.0	58.6	82.7	67.8	64.9	86.3	60.6	46.6	86.6	70.6
MoCo-IN [6]	64.1	78.7	74.9	85.8	43.0	74.7	59.4	65.8	75.4	77.1	77.4	59.4	57.9	85.9	70.9	67.5	85.6	59.1	46.9	85.5	69.8
MoCo-MAID [6]	60.6	80.7	72.0	86.5	42.2	74.3	60.7	61.8	75.5	74.8	76.4	57.7	57.4	84.4	69.8	61.6	86.3	55.5	46.7	85.7	68.5
MoCo-IN-MAID [6]	61.3	78.9	75.3	86.2	42.7	75.8	62.0	61.0	75.7	77.1	75.0	57.8	57.8	84.2	67.3	62.2	85.1	58.7	45.6	85.2	68.7
MoCov2 [35]	64.0	80.1	74.8	86.4	44.0	74.9	61.1	61.7	75.8	75.1	76.3	59.3	59.4	85.0	70.2	64.4	85.0	55.5	47.3	86.4	69.3
SeCo [9]	57.5	75.2	73.9	81.3	37.4	72.9	61.6	60.0	72.5	74.2	72.5	56.6	53.6	80.4	66.3	57.1	82.9	58.6	41.4	83.0	66.0
Geo-Aware [11]	63.0	77.4	76.1	85.1	42.2	76.5	61.2	61.6	75.9	72.2	74.7	56.7	57.6	83.7	67.2	63.5	85.5	60.0	45.1	84.0	68.4
CACO [64]	59.4	76.5	75.3	83.2	40.9	73.3	58.3	62.6	73.8	75.7	73.5	56.5	55.9	83.0	69.7	58.7	83.6	54.6	43.0	82.9	67.0
TOV [18]	64.7	81.7	77.5	85.6	44.8	76.0	65.7	64.9	79.6	77.8	77.6	57.1	57.2	81.7	71.1	65.3	84.8	63.3	44.8	85.0	70.2
GeRSP	68.4	83.6	75.9	87.4	47.0	78.4	63.3	67.8	79.4	79.1	77.6	57.7	59.1	84.2	68.4	66.8	86.6	59.0	48.1	86.4	71.2
GeRSP-200	72.0	84.1	76.8	86.9	47.3	79.0	62.3	70.9	81.4	78.9	78.9	58.2	60.7	83.9	70.4	67.7	86.9	62.8	48.9	86.6	72.2

comparing MoCo-MAID with GeRSP. The proposed GeRSP, on the other hand, learns representations with general and special knowledge simultaneously through a unified framework. Compared with SeCo, CACO, TOV, and Geo-Aware, GeRSP consistently improves top-1 accuracy, indicating that GeRSP can learn robust representations with both general knowledge and domain specializations.

Object Detection: Table II compares the detection performance of GeRSP on DIOR with other pre-training methods. The comparison is made by fine-tuning with Faster R-CNN [67], RetinaNet [68], and DyHead [69]. We observe that GeRSP consistently outperforms other pre-training methods. Compared to IMP, GeRSP improves mAP from 65.1% to 67.1% (+2.0%) on Faster R-CNN, improves mAP from 66.5% to 67.5% (+1.0%) on RetinaNet, improves 70.6% to 71.2% (+0.6%) on DyHead. MoCo-MAID only uses MAID for training, so it can only obtain domain knowledge about remote sensing images and lacks more robust generalization performance. Compared to MoCo-MAID, GeRSP improves mAP from 66.1% to 67.1% (+1.0%) on Faster R-CNN, GeRSP

improves mAP from 64.2% to 67.5% (+3.3%) on RetinaNet, GeRSP improves 68.5% to 71.2% (+2.7%) on DyHead.

As evidenced by Table II, MoCo-IN-MAID, despite being pre-trained using both the ImageNet and MAID datasets, fails to enhance performance on downstream tasks and exhibits a certain degree of degradation. Conversely, our GeRSP framework consistently achieves effective performance improvement. This observation underscores GeRSP's ability to effectively leverage information from multiple datasets, resulting in the acquisition of more generalized features. Additionally, it is worth noting that the performance of pre-trained weights obtained through existing pre-training methods in the detection task displays instability, with their fine-tuning performance on the three detection methods slightly trailing behind that of IMP. In contrast, GeRSP exhibits exceptional stability in its performance in detection while continuously demonstrating improvement over time.

Table III compares the fine-tuning effects of different pre-training methods on DOTA. We employed Faster R-CNN, RetinaNet, and DyHead as the detection models to validate the

TABLE III
FINE-TUNING RESULTS ON THE DOTA [71] OBJECT DETECTION TASK. WE USE THREE DETECTION METHODS TO VERIFY THE PRE-TRAINED MODELS.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
<i>Faster R-CNN</i>																
From Scratch	76.5	44.0	23.2	21.5	63.7	73.3	83.4	85.5	13.7	48.0	21.6	25.1	66.6	47.3	19.8	47.6
IMP	84.1	63.3	41.4	60.9	66.9	77.1	86.2	92.6	53.0	61.4	61.0	56.4	76.1	52.3	44.1	65.1
MoCo-IN [6]	84.1	63.3	39.9	59.1	66.7	74.5	86.6	91.0	57.4	60.1	59.3	50.7	78.5	48.0	40.2	63.9
MoCo-MAID [6]	83.3	66.1	40.3	55.3	66.9	74.2	85.3	90.8	60.5	58.3	57.2	52.7	77.2	46.7	38.5	63.6
MoCo-IN-MAID [6]	83.4	62.1	39.9	53.8	66.3	74.5	85.3	91.8	59.0	58.6	54.8	56.6	76.9	48.9	33.8	63.0
MoCov2 [35]	83.4	58.3	40.4	58.4	67.7	75.3	86.2	91.8	54.2	58.4	59.5	56.8	78.1	51.2	42.3	64.2
SeCo [9]	82.0	57.3	35.8	42.6	65.4	76.3	86.5	90.8	39.2	57.7	51.8	46.7	75.2	50.5	34.7	59.5
Geo-Aware [11]	84.1	66.1	36.9	56.0	68.0	76.2	86.0	91.6	53.5	61.5	61.4	52.1	74.3	51.0	43.8	64.2
CACO [64]	82.0	63.6	38.3	49.1	67.5	76.9	85.5	91.2	44.4	57.8	55.8	52.8	75.0	46.6	37.0	61.6
TOV [18]	83.7	66.5	41.4	60.6	67.4	76.3	86.8	91.6	60.1	61.5	62.0	55.8	75.2	51.4	45.6	65.6
GeRSP	85.3	63.6	42.7	60.7	67.2	75.2	85.9	91.7	60.4	63.0	61.2	56.8	78.2	50.4	46.5	65.9
GeRSP-200	85.2	64.3	44.9	60.5	65.4	75.0	86.1	92.8	65.0	63.2	60.4	56.3	76.7	50.5	44.7	66.1
<i>RetinaNet</i>																
From Scratch	74.7	41.1	17.7	22.8	57.0	64.6	71.1	83.8	9.2	40.7	21.8	26.3	67.5	33.4	8.8	42.7
IMP	83.3	64.6	34.1	54.6	58.2	70.6	74.5	92.8	55.1	57.4	54.7	54.9	74.0	48.1	25.9	60.2
MoCo-IN [6]	83.6	59.2	32.6	49.9	61.2	71.8	76.5	92.3	54.8	60.0	49.5	46.6	76.8	45.9	23.2	58.9
MoCo-MAID [6]	84.3	62.3	35.6	52.0	60.8	72.1	76.5	92.0	57.3	58.6	47.5	50.1	77.3	47.1	24.4	59.9
MoCo-IN-MAID [6]	82.8	62.8	34.5	45.5	58.4	71.9	75.4	93.0	53.8	57.2	48.6	51.6	75.7	46.5	29.9	59.2
MoCov2 [35]	83.7	63.8	35.0	53.3	60.2	71.6	76.4	91.8	51.9	57.4	48.6	48.6	77.0	48.1	32.3	60.0
SeCo [9]	80.8	56.9	30.1	38.7	60.8	69.1	75.1	89.5	41.9	53.9	43.2	43.2	72.8	43.7	16.6	54.4
Geo-Aware [11]	81.0	62.3	27.6	52.9	59.2	67.3	73.8	91.3	53.5	54.6	46.8	49.5	71.2	47.1	26.6	57.7
CACO [64]	79.9	60.0	32.7	42.3	59.0	68.3	74.9	91.2	47.4	53.1	43.4	50.9	74.0	43.5	22.5	56.2
TOV [18]	82.3	66.1	36.9	51.9	61.4	70.7	76.9	91.0	52.6	57.8	50.8	57.3	74.5	49.9	29.0	60.6
GeRSP	84.9	65.6	36.3	54.3	60.0	71.3	74.7	93.7	67.4	59.2	54.8	56.4	75.8	50.1	31.8	62.4
GeRSP-200	85.5	65.3	38.1	55.5	60.7	72.5	75.9	94.2	65.2	58.7	54.9	57.3	77.4	50.6	32.6	63.0
<i>DyHead</i>																
From Scratch	79.7	49.2	27.1	37.1	63.0	72.1	81.9	86.9	26.6	52.1	23.3	37.9	75.4	46.1	18.5	51.8
IMP	85.5	64.4	40.3	59.7	66.6	76.1	84.9	92.6	54.0	65.5	47.9	54.7	79.9	49.1	51.0	64.8
MoCo-IN [6]	85.4	59.6	37.9	52.7	68.7	75.5	84.1	92.9	59.7	66.6	48.7	51.9	81.2	48.1	35.4	63.2
MoCo-MAID [6]	84.2	65.5	40.5	55.1	67.7	74.7	84.2	93.2	60.3	66.7	49.4	52.3	80.7	46.2	32.2	63.5
MoCo-IN-MAID [6]	84.7	65.1	39.3	53.5	67.1	74.6	84.2	92.8	58.1	66.3	45.5	54.5	79.8	47.1	31.2	62.9
MoCov2 [35]	85.2	62.3	40.2	49.5	67.0	74.9	84.1	92.6	53.7	66.4	46.8	51.8	80.5	50.0	38.6	62.9
SeCo [9]	82.1	61.0	35.0	46.5	63.8	72.3	83.2	89.5	44.9	59.4	31.1	46.1	78.7	47.6	28.4	58.0
Geo-Aware [11]	83.8	62.7	37.4	57.3	65.0	74.4	83.6	90.7	48.0	64.5	45.3	58.8	78.5	48.9	37.6	62.4
CACO [64]	82.6	59.7	38.7	56.2	65.0	75.9	83.7	90.2	48.2	61.1	40.9	51.8	79.3	44.8	28.8	60.5
TOV [18]	85.5	66.0	44.1	59.3	64.8	76.2	84.4	91.4	61.8	63.6	46.9	56.5	80.2	49.5	42.8	64.9
GeRSP	87.0	66.3	42.4	60.3	68.3	75.7	85.0	93.7	65.5	68.5	51.5	56.5	80.0	51.2	40.2	66.1
GeRSP-200	87.3	66.9	44.3	62.4	67.7	76.1	84.5	93.3	65.3	69.5	52.5	60.5	80.5	51.3	49.0	67.4

results. GeRSP consistently achieves notable improvements in detector performance and outperforms other methods across most regions. Specifically, GeRSP proves more advantageous than IMP in detection tasks, showcasing an increase of +0.8% on Faster R-CNN, +2.2% on RetinaNet, and +1.3% on DyHead. Furthermore, GeRSP demonstrates more significant improvement than MoCo-MAID, with gains of +2.3% on Faster R-CNN, +2.5% on RetinaNet, and +2.6% on DyHead.

Notably, MoCo-MAID, which solely utilizes remote sensing images, fails to exhibit any advantages over IMP in the detection task, showing decreases of -1.5% on Faster R-CNN, -0.3% on RetinaNet, and -1.3% on DyHead. Moreover, the more advanced self-supervised pre-training method MoCov2 fails to yield performance improvements. In contrast, our GeRSP method effectively enhances detection performance. Moreover, from the GeRSP-200 row, it is evident that extended training further enhances detection performance. GeRSP proves highly effective in elevating the capabilities for detection.

Segmentation: Table IV compares mIoU of PSANet [74] and DeepLabV3+ [75] initialized by GeRSP and other pre-

training methods on the LoveDA dataset.

For PSANet, GeRSP consistently outperforms other pre-training methods on LoveDA. Compared with MoCo-IN, MoCo-MAID, and MoCo-IN-MAID, GeRSP improves mIoU by +0.99%, +2.38%, and +1.19%, which demonstrates that both pre-training dataset and method play essential roles in representation learning. Compared to GeRSP, MoCo-IN exhibits poor performance due to its insufficient understanding of knowledge in the RS domain. Although MoCo-MAID utilizes remote sensing images for pre-training and acquires domain knowledge specific to such images, it lacks generalization ability. In the case of MoCo-IN-MAID, the pre-training method also influences the pre-training process.

For DeepLabV3+, GeRSP outperforms all other supervised pre-training and self-supervised pre-training methods on LoveDA. As shown in Table IV, GeRSP improves mIoU by +2.42%, +1.86%, and +0.8% when compared with MoCo-IN, MoCo-MAID, and MoCo-IN-MAID. The results are the same as in PSANet, demonstrating that GeRSP efficiently learns representations for remote sensing downstream tasks.

Compared with SeCo, GeRSP consistently improves mIoU on the two semantic methods, especially +4.89% for PSANet and +5.43% for DeepLabV3+. This observation shows that GeRSP can effectively enhance semantic segmentation performance.

TABLE IV
RESULTS OF PSANET [74] AND DEEPLABV3+ [75] WITH DIFFERENT PRE-TRAINED BACKBONES ON LOVEDA [76] DATASET.

Pre-training	Dataset	PSANet mIoU	DeepLabv3+ mIoU
IMP	ImageNet	49.24	48.39
MoCo-IN [6]	ImageNet	48.54	46.64
MoCo-MAID [6]	MAID	47.15	47.20
MoCo-IN-MAID [6]	ImageNet + MAID	48.34	48.26
MoCov2 [35]	MAID	44.79	48.57
SeCo [9]	1M Sentinel-2	44.64	43.63
Geo-Aware [11]	GeoImageNet	49.37	48.76
CACO [64]	CACO 1M	48.81	48.89
TOV [18]	TOV-NI + TOV-RS	49.33	49.7
GeRSP	ImageNet + MAID	49.53	49.06
GeRSP-200	ImageNet + MAID	49.56	50.56

TABLE V
SENSITIVITY OF BALANCE COEFFICIENT α IN GERSP.

α	EuroSAT Top-1	NWPU-RESISC45 Top-1	DOTA mAP	DIOR mAP	PSANet mIoU	DeepLabv3+ mIoU
0	97.52 \pm 0.10	91.19 \pm 0.15	63.6	66.1	47.15	47.20
0.5	97.79 \pm 0.10	92.54 \pm 0.09	65.6	67.6	49.37	49.36
1	97.87 \pm 0.15	92.67 \pm 0.16	65.9	67.1	49.53	49.06

B. Visualization

Fig. 4 visually explains model predictions on scene classification tasks through class activation maps (CAM). The CAM is implemented by Grad-CAM++ [77]. The principle of Grad-CAM++ [77] is to use a weighted combination of the positive partial derivatives as weights to sum activation maps and construct CAM capable of explaining numerous instances. We compare the models obtained by INT Supervised and GeRSP, as shown in the second and third rows of Fig. 4. Intuitively, the GeRSP model can focus well on the instances in RS images, such as school, palace, and baseball diamond. It can also sense multiple objects in the image, such as ships in the harbor shown in the third column of Fig. 4. This demonstrates that GeRSP can improve the model's ability to extract semantic information from RS images. On the contrary, the IMP model is difficult to capture RS semantic information and lacks the generalization ability in RS cognitive tasks. Fig. 5 shows that the IMP model struggles to recognize the scene categories in images, whereas the GeRSP model can recognize these scene concepts through unsupervised learning.

We also undertake a stage-wise linear evaluation, as outlined by Wang et al. [19], to elucidate the underlying sources of generalization of GeRSP. Following [19], we freeze the parameters of the pre-trained model and subsequently conduct linear evaluation training using features obtained from each stage. The results of linear evaluation reflect the semantic information contained in the features. Additionally, since the

quality of low-mid level features significantly impacts the performance of semantic discrimination tasks, the performance of linear evaluation indirectly reflects the caliber of low-mid level features. As shown in Fig. 6, overall, as we progress through the stages, there is a gradual enhancement of semantic information, resulting in a continuous improvement in linear evaluation performance. Compared with other competitive methods, it is evident that both IMP and our GeRSP demonstrate comparable linear evaluation performance across different stages, consistently outperforming other methods at every stage, especially in the first and second stages. This phenomenon suggests that ImageNet supervised training can acquire high-quality image features across different levels, particularly emphasizing low to mid-level features, which is advantageous for semantic discrimination tasks in remote sensing images. Therefore, our GeRSP, through the introduction of natural image supervised learning, acquires rich generic image knowledge that can be effectively applied to subsequent remote sensing tasks.

We conduct experiments on the balance coefficient α in Eq. 3. As shown in Table V, when there is no supervised loss, e.g., α set to 0, the model's performance decreases across all tasks. When α is set to 0.5 or 1, the natural image supervised loss is introduced into the pre-training, significantly improving performance. Also, from the results, the scale of α has a relatively minor impact on the pre-training effectiveness, demonstrating the insensitivity of GeRSP to α .

VII. CONCLUSION

This paper proposes Generic Knowledge Boosted Remote Sensing Pre-training (GeRSP) for remote sensing pre-training. GeRSP aims to obtain a pre-trained model suitable for initializing deep learning models for RS understanding tasks. The proposed method leverages a teacher-student architecture to harness the benefits of both supervised pre-training and self-supervised pre-training, mitigating the impact of domain gaps between RS images and natural images. During the self-supervised pre-training process, GeRSP acquires domain-specific features from unlabeled RS images. In contrast, in the supervised pre-training process, it learns general knowledge from labeled natural images. By integrating self-supervised and supervised pre-training, GeRSP simultaneously learns representations with general and special knowledge. Subsequently, GeRSP's effectiveness is evaluated by conducting three remote sensing downstream tasks: object detection, semantic segmentation, and scene classification. Consistently, the experimental results demonstrate that GeRSP is an effective pre-training method in remote sensing, enhancing the performance of various downstream tasks.

GeRSP effectively mitigates the limitations of contrastive learning in perceiving fine-grained features in RS images, ensuring transferability across various tasks, especially segmentation and detection tasks. In addition to introducing ImageNet supervised learning, using masked image modeling [8], [78]–[82] and explicitly specified learning of image descriptors [80]–[82] can achieve similar goals. Replace supervised training with these methods, potentially resulting in enhanced

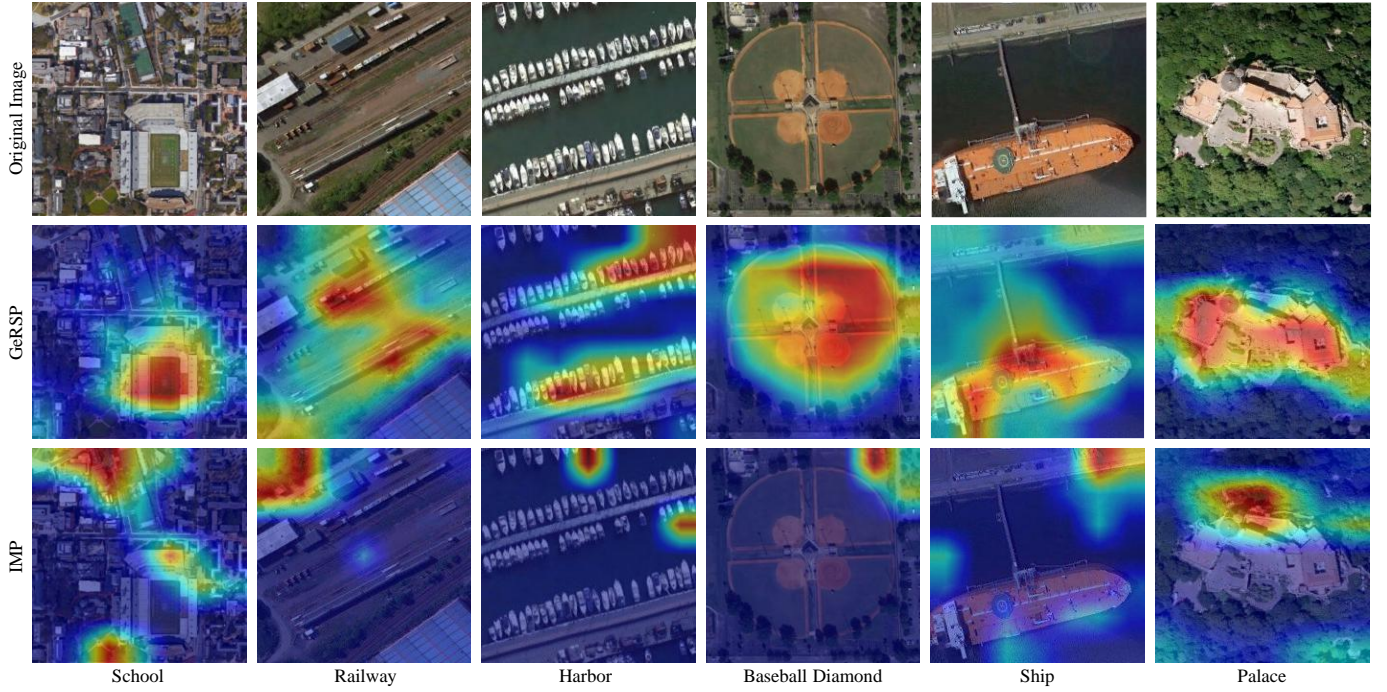


Fig. 4. Class activation maps (CAMs) visualization of GeRSP model and IMP model on six categories.

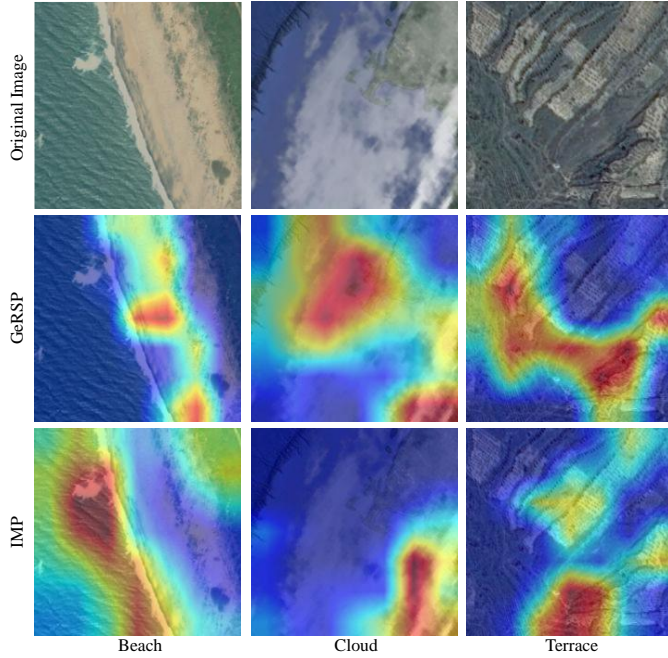


Fig. 5. Class activation maps (CAMs) visualization of GeRSP model and IMP model on beach, cloud, and terrace.

generalization performance, which we will further consider in future.

REFERENCES

- [1] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2849–2858.
- [2] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of cnss," *ISPRS J. Photogramm. Remote Sens.*, vol. 3, pp. 473–480, 2016.
- [3] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

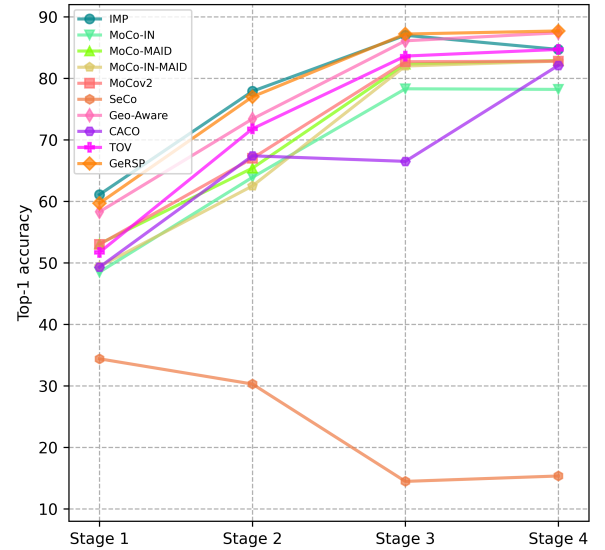


Fig. 6. Top-1 accuracy of stage-wise evaluation [19] on NWPU-RESISC45 [40]. All parameters of the pre-trained models are frozen during the training process. During training, features from various stages are extracted, flattened, and fed into a linear layer for classification, achieving a stage-wise linear evaluation.

- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [5] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, 2022.
- [6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9729–9738.
- [7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 1597–1607.
- [8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 000–16 009.
- [9] 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 Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.
- [10] C. Liu, H. Sun, Y. Xu, and G. Kuang, "Multi-source remote sensing pretraining based on contrastive self-supervised learning," *Remote Sens.*, vol. 14, no. 18, p. 4632, 2022.
- [11] K. Ayush, B. Uzket, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10 181–10 190.
- [12] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3967–3974.
- [13] S. He, Q. Li, Y. Liu, and W. Wang, "Semantic segmentation of remote sensing images with self-supervised semantic-aware inpainting," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [14] N. Zhao, Z. Wu, R. W. Lau, and S. Lin, "What makes instance discrimination good for transfer learning?" *Proc. Int. Conf. Learn. Represent.*, 2020.
- [15] I. Corley, C. Robinson, R. Dodhia, J. M. L. Ferres, and P. Najafirad, "Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters," *arXiv preprint arXiv:2305.13456*, 2023.
- [16] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 5414–5423.
- [17] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, 2022.
- [18] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "Tov: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2023.
- [19] Y. Wang, S. Tang, F. Zhu, L. Bai, R. Zhao, D. Qi, and W. Ouyang, "Revisiting the transferability of supervised pretraining: an mlp perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 9183–9193.
- [20] T. Zhang, P. Gao, H. Dong, Y. Zhuang, G. Wang, W. Zhang, and H. Chen, "Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain," *Remote Sens.*, vol. 14, no. 22, p. 5675, 2022.
- [21] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Represent.*, 2021.
- [24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
- [25] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 22 243–22 255, 2020.
- [26] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [27] 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, 2020.
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2536–2544.
- [30] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 69–84.
- [31] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 649–666.
- [32] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *Proc. Int. Conf. Learn. Represent.*, 2019.
- [33] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.
- [34] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.
- [35] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
- [36] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9640–9649.
- [37] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent—a new approach to self-supervised learning," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21 271–21 284, 2020.
- [38] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14 680–14 707, 2015.
- [39] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, 2015.
- [40] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [41] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.
- [42] G. Mátyus, S. Wang, S. Fidler, and R. Urtasun, "Hd maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 3611–3619.
- [43] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.
- [44] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. of Environment*, vol. 237, p. 111322, 2020.
- [45] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.
- [46] Y. Long, G.-S. Xia, L. Zhang, G. Cheng, and D. Li, "Aerial scene parsing: From tile-level scene classification to pixel-wise semantic labeling," *arXiv preprint arXiv:2201.01953*, 2022.
- [47] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *arXiv preprint arXiv:2202.10108*, 2022.
- [48] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [49] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.
- [50] M. Zhu, J. Fan, Q. Yang, and T. Chen, "Sc-eadnet: A self-supervised contrastive efficient asymmetric dilated network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.
- [51] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 1182–1191.
- [52] Z. Wen, Z. Liu, S. Zhang, and Q. Pan, "Rotation awareness based self-supervised learning for sar target recognition with limited training samples," *IEEE Trans. Image Process.*, vol. 30, pp. 7266–7279, 2021.

- [53] Y. Chen and L. Bruzzone, "Self-supervised remote sensing images change detection at pixel-level," *arXiv preprint arXiv:2105.08501*, 2021.
- [54] X. Ou, L. Liu, S. Tan, G. Zhang, W. Li, and B. Tu, "A hyperspectral image change detection framework with self-supervised contrastive learning pretrained model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7724–7740, 2022.
- [55] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [56] 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, 2021.
- [57] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 1422–1431.
- [58] V. Risojević and V. Stojnić, "Do we still need imagenet pre-training in remote sensing scene classification?" *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 1399–1406, 2022.
- [59] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, "Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, 2020.
- [60] S. Purushwalkam, P. Morgado, and A. Gupta, "The challenges of continuous self-supervised learning," *Proc. Eur. Conf. Comput. Vis.*, 2022.
- [61] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [62] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 3833–3845, 2020.
- [63] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021.
- [64] U. Mall, B. Hariharan, and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 5261–5270.
- [65] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.
- [66] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [67] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.
- [68] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017, pp. 2980–2988.
- [69] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 7373–7382.
- [70] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.
- [71] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3974–3983.
- [72] 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, 2020.
- [73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
- [74] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.
- [75] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [76] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- [77] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.
- [78] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "Cmid: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, 2023.
- [79] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 15619–15629.
- [80] Y. Wang, H. H. Hernández, C. M. Albrecht, and X. X. Zhu, "Feature guided masked autoencoder for self-supervised learning in remote sensing," *arXiv preprint arXiv:2310.18653*, 2023.
- [81] W. Li, Y. Wei, T. Liu, Y. Hou, Y. Liu, and L. Liu, "Self-supervised learning for sar atr with a knowledge-guided predictive architecture," *arXiv preprint arXiv:2311.15153*, 2023.
- [82] 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 Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 14668–14678.



Ziyue Huang received the BS degree in information and computing science from the School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in computer science with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and object detection.



Mingming Zhang received the BS degree in information and computer science from Liaoning University, Shenyang, China and the MS degree in software engineering from Beihang University, Beijing, China. She is currently pursuing the Ph.D. degree in computer science with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University. Her research interests include remote sensing image analysis and computer vision.



Yuan Gong received the BS degree in computer science from the School of Computer Science and Engineering at Beihang University, Beijing, China, in 2022. He is currently pursuing a Master's degree in software engineering at the School of Software and Microelectronics, Peking University, Beijing, China.



Qingjie Liu received the BS degree in computer science from Hunan University, Changsha, China, in 2007, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2014. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. He is also a Distinguished Research Fellow with the Hangzhou Institute of Innovation, Beihang University, Hangzhou. His current research interests include image fusion, object detection, image segmentation, and change detection. He is a member of

the IEEE.



Yunhong Wang received the BS degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the MS and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively.

She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering,

Beihang University, Beijing, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing. She is a Fellow of IEEE, IAPR, and CCF.