Geographical Supervision Correction for Remote Sensing Representation Learning

Wenyuan Li, Keyan Chen and Zhenwei Shi* *Member, IEEE*

***Abstract*—Global land cover (GLC) products can be utilized to provide geographical supervision for remote sensing representation learning, which has significantly improved downstream tasks' performance and decreased the demand of manual annotations. However, the time differences between remote sensing images and GLC products may introduce deviations in geographical supervision. In this paper, we propose a Geographical supervision Correction method (GeCo) for remote sensing representation learning. Deviated geographical supervision generated by GLC products can be corrected adaptively using the correction matrix during network pre-training and joint optimization process is designed to simultaneously update the correction matrix and network parameters. Additionally, we identify prior knowledge on geographical supervision to guide representation learning and restrict the correction process. The prior knowledge named "minor changes" implies that the geographical supervision may not change significantly, whereas the prior knowledge named "spatial aggregation" implies that land covers are aggregated in their spatial distribution. According to the prior knowledge, corresponding regularization terms are proposed to prevent abrupt changes in geographical supervision correction process and excessive smoothing of network outputs, thereby ensuring the adaptive correction process's correctness. Experimental results demonstrate that our proposed method outperforms random initialization, ImageNet pre-training, and other representation learning methods on a variety of downstream tasks. In particular, when compared to the method that learns representations directly from deviated geographical supervision, it is proved that our method can eliminate the influence of deviations and further improve the effect of representation learning.**

***Index Terms*—representation learning, remote sensing images, scene classification, semantic segmentation, object detection, cloud / snow detection**

I. Introduction

The deep learning method has possessed a high capacity for feature learning and demonstrated remarkable performance in remote sensing scene classification [1, 2], object detection [3–6] and semantic segmentation [7–11]. However, its effect depends on a large number of manual annotations, and the process of producing annotations is time- and labor-intensive. Expert domain knowledge is also required for remote sensing images, making large-scale annotation more difficult.

To alleviate deep learning's reliance on large number of manual annotations, considerable recent researches have been conducted on self-supervised representation learning methods. These methods usually follow a two-stage paradigm, including network pre-training and fine-tuning. During the first stage, they automatically extract supervision information from the data itself or the external environment by designing various pretext-tasks, thereby completing the network pre-training. No manual annotated data is necessary during this stage. For the second stage, the corresponding annotated data will be used to fine-tune the pre-training model for downstream tasks. It can improve downstream tasks' performance and decrease the reliance on annotated data.

The critical factor affecting self-supervised representation learning's performance is the ability of pretext tasks to extract effective supervision information. The straightforward way to design a pretext task is to obtain supervision information via image transformation [12–19], such as image colorization [17, 18] and inpainting [15, 16]. However, the performance of the pre-training model obtained in this manner is limited and cannot be expected to exceed that of the widely used ImageNet [20] pre-training model. This could be because, in the absence of manual annotations, the supervision information obtained through image transformation methods is always biased toward a particular level of representations, making it difficult to ensure the model's generalizability when transferred to downstream tasks. To address this issue, contrastive learning methods [21–31] construct positive and negative samples from multiple images and improve representation learning's performance by maximizing differences between positive and negative samples. Recent contrastive learning methods have outperformed ImageNet [20] pre-training on a variety of downstream tasks. Another methods of constructing pretext tasks are to obtain supervision information from the external environment or multi-modal data, such as by combining voice and text with corresponding images [32–34].

In contrast to general image representation learning, remote sensing images frequently contain a large number of similar land covers. It may affect the contrastive learning performance as narrowing the discrepancies between positive samples and negative samples. Additionally, remote sensing images lack auxiliary data such as sound and text. But they usually contain associated geographical knowledge that can be used to perform remote sensing image processing and representation learning [35–37]. The geographical knowledge includes geographical location, global land cover (GLC) products, open street map [38] and etc. They can be utilized to provide geographical supervision for remote sensing image processing tasks, such as using open street maps [38] to provide annotated information for semantic segmentation. The GeoKR method [39] lever-

The work was supported by the National Natural Science Foundation of China under the Grant 62125102. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Wenyuan Li,Keyan Chen and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

ages geographical location information and global land cover (GLC) products [40] to generate geographical supervision for representation learning, which can improve downstream tasks' performance and reduce reliance on annotated data.

However, there are certain discrepancies between GLC products and remote sensing images in terms of producing times, resolutions and etc. As a result of these discrepancies, a significant amount of deviations may be introduced into the generated geographical supervision, impairing the effectiveness of remote sensing representation learning. To mitigate the effect of deviations, we propose a deviated **Ge**ographical supervision **Co**rrection method (**GeCo**) to improve the performance of remote sensing representation learning. During the networks pre-training, a correction matrix W is developed to correct the geographical supervision adaptively. Joint optimization process is designed to update correction matrix and networks parameters simultaneously.

To ensure that the corrected geographical supervision is more closely aligned with its actual distribution, we conduct a systematic analysis of the generated geographical supervision and discover two priors. They are named as "minor changes" and "spatial aggregation", which can restrict the correction process. The prior knowledge "minor changes" reflects the fact that geographical supervision may not change significantly over time. We propose a regularization term to ensure that geographical supervision correction adheres to this prior knowledge by bringing the correction matrix close to the identity matrix. The prior knowledge "spatial aggregation" reflects the fact that land covers present an aggregated state in the spatial distribution. We propose a regularization term that lets networks produce uneven predictions, ensuring that the prior "spatial aggregation" is satisfied. In addition, it can also keep the network from producing excessively smooth results and ensure the effectiveness of representation learning.

We adopt Levir-KR dataset [39] and three GLC products, GLobeLand30 [40], GLCFCS30 [41] and FromGLC30 [42], for representation learning. We demonstrate our method's effectiveness in scene classification, semantic segmentation, and object detection of remote sensing images. GeoKR [39], ImageNet pre-training method [20], and contrastive learning methods including Moco [21], SimCLR [22], and BYOL [23] are selected as comparison methods. Our experimental results demonstrate that our method effectively reduces the interference of deviated geographical supervision on representation learning and enhances the performance of downstream tasks.

The contributions can be summarized as follows:

- An adaptive correction method of deviated geographical supervision is proposed to improve the performance of representations learning in remote sensing images. The correction matrix is designed to correct the deviated geographical supervision adaptively while network pre-training.
- Regularization terms are proposed to ensure that the corrected geographical supervision conforms to prior knowledge of land cover change and spatial distribution, therefore preventing the geographical supervision correction and representation learning process from collapsing.
- Experimental results prove that method can eliminate the influence of deviations and further improve the performance of representation learning.

The rest of this paper is organized as follows. In Section II, we introduce the related work. In Section III, we give a detailed introduction of our proposed method. In Section IV, the experimental results are presented. Conclusions are drawn in section V.

II. Related Work

A. Self-supervised Representation Learning

Self-supervised representation learning divides the deep learning-based image processing problem into two stages. In the first stage, several pretext tasks are designed to extract supervision information from the data itself or the external environment in order to construct an appropriate loss function and finish network pre-training. Due to the fact that this procedure does not require manual annotations, any scale dataset can be utilized for pre-training. In the second stage, a small quantity of annotated data is used to fine-tune the pre-training model for specific downstream tasks (image classification, object detection, semantic segmentation and etc.). It can improve the performance of downstream tasks and reduce their demand for manual annotations.

Prior to self-supervised representation learning, the ImageNet pre-training model + fine-tuning paradigm is typically adopted. Its issue is that because the ImageNet [20] dataset is manually annotated, it is extremely challenging to continue expanding its scale. Self-supervised learning does not require manually labeled data during the pre-training stage, so it has the ability to exploit an infinite amount of data. However, it is highly dependent on the efficiency of pretext tasks. Pretext tasks that are well-designed can aid the network in extracting effective representation. They can be classified into three categories based on the way they get supervision information: from single image [12–19, 43], from multiple images [21–31], and from external environment [32–34].

- **from single image**

The majority of early researches obtains supervision and constructs loss functions from a single image transformation. For instance, use the input color image as the label, convert it to grayscale image, and create a pretext task based on the image coloring task [17, 18]. Occlude a random portion of the image and use the image inpainting task as a pretext task for completing the network's pre-training [15, 16, 19]. These pretext tasks are also frequently used image processing tasks, which can be completed effectively only when the network understands the image, allowing for the extraction of effective representations. However, the pre-training model generated by these methods is prone to overfitting, and the effect is generally inferior to the ImageNet pre-training model.

- **from multiple images**

Contrastive learning [21–31] is the most frequently used and effective method. During network pre-training, the contrastive learning method constructs the loss function using the relationship between representations of multiple images. Firstly, positive samples of the input image are obtained through data augmentation, and negative samples are re-selected from the training data. Then, the input image's representation and the positive sample's representation are constrained to be close to one another, while the input image's representation and negative sample's representation are constrained to be far apart.

Contrastive learning significantly improves self-supervised learning's performance and outperforms the ImageNet [20] pre-training model in a variety of downstream tasks. The division of positive and negative samples, particularly the selection of negative samples, is critical for contrastive learning to be effective. [25] proposes that negative samples be saved and updated in a memory bank, which can significantly increase the number of negative samples during pre-training stage and improve the quality of negative samples. However, this method requires a large amount of memory to save the memory bank, and updating the negative samples might take a long time. On this basis, he et al. propose MoCo [21, 26, 44] series methods that save and update negative samples using a momentum encoder rather than a memory bank, which not only saves memory but also improves the quality of negative samples and significantly improves the effect of comparative learning. [22] proposes an end-to-end contrastive learning method in which each image is treated as a positive sample and the remaining data as negative samples, but experiments demonstrate that this method requires a larger batch size.

- **from external environment**

Along with the methods outlined above, supervision information can be gleaned from external data such as text and sound that closely match the image [32–34]. For instance, [33] collects some images and text data concurrently on the Internet, use text to supervise image representation learning, and achieve favorable results.

B. Representation Learning for Remote Sensing Images

Due to the fact that self-supervised learning does not require annotated data, the efficiency of representation learning is relatively low. To replicate the ImageNet pre-training effect, some additional information must be added, such as the relationship between multiple images or external text and sound data. However, it is nearly impossible to obtain the corresponding text and other data for remote sensing images. Additionally, due to the similarity of land covers, a large number of remote sensing images exhibit similar visual characteristics, making it difficult to ensure sufficient discrimination between positive and negative samples. The majority of researches on remote sensing representation learning makes reference to some form of geographical knowledge that cannot be acquired through general image representation learning. As a result, learning representations for remote sensing images is distinct from learning representations for other types of images.

[35, 37] select positive and negative samples using geographical location information as a reference, which makes the selection more rational and enhances the effectiveness of representation learning. [36] creates a new head with geographical location based on the original self-supervised representation learning framework. Additionally, [39] incorporates global land cover products and geographic location as geographical knowledge to provide supervision for representation learning. It has the potential to enhance the effectiveness of self-supervised learning. However, due to the differences in producing times, the supervision provided by the geographical knowledge will invariably introduce deviations and distort the effect of the pre-training model.

III. Method

In this section, we detail our method for deviated **Ge**ographical supervision **Co**rrection (**GeCo**) to improve the performance of remote sensing representation learning. The overall is shown in fig. 1 [1].

A. Overall Framework

We use GlobeLand30 2020, FromGLC30 2017, and GLCFCS30 2020 to generate geographical supervision for remote sensing images. The process for generating geographical supervision is identical to that described in section III-B.

Because remote sensing images and GLC products have different producing times and resolutions, deviations may still be introduced into the generated geographical supervision. The learning process can be expressed as follows under the guidance of deviated geographical supervision:

$$\min_{\theta} L(X, A \mid \theta), \tag{1}$$

where L represents the loss function for representation learning. X and A represent input image and its corresponding deviated geographical supervision, respectively. θ denotes network parameters.

To reduce the impact on deviations, we propose an adaptive correction method of deviated geographical supervision during network pre-training in order to maximize the stability and effectiveness of representations. We define the correction matrix W, which is capable of correcting the deviated geographical supervision A into :

$$\tilde{A} = A^T \times \mathbf{W}. \tag{2}$$

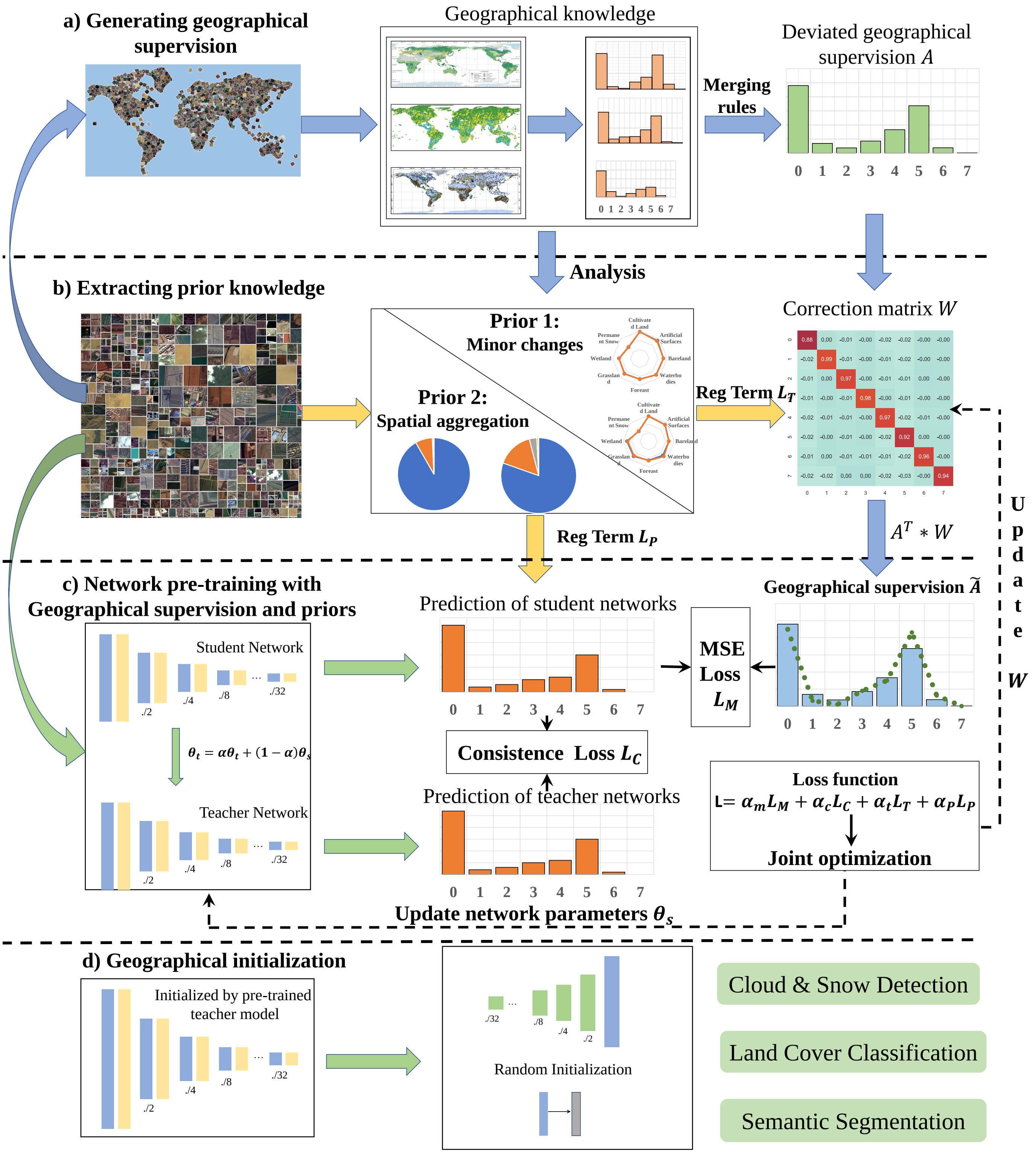
In general, at the spatial level, deviations manifest themselves as variations in the spatial distribution of various land covers. The specific form of deviation at the level of geographical supervision is the difference between the proportion of land covers in the entire dataset obtained using GLC products and the actual proportion of land covers, which can be described as $D = A - \tilde{A}$.

Simultaneously, we develop a joint optimization method for optimizing both the network parameters θ and the correction matrix W. The optimization process can be described as follows:

$$\min_{\theta, W} L(X, \tilde{A} \mid \theta, \mathbf{W}). \tag{3}$$

To make the correction process more stable and to prevent it collapsing, regularization terms are devised to constrain

[1] Some figures of geographical knowledge cites from [40–42]


a) Generating geographical supervision
Geographical knowledge
0 1 2 3 4 5 6 7
Merging rules
Deviated geographical supervision A
0 1 2 3 4 5 6 7
Analysis
b) Extracting prior knowledge
Prior 1:
Minor changes
Prior 2:
Spatial aggregation
Cultivated Land
Artificial Surfaces
Bareland
Waterbodies
Foreast
Grassland
Wetland
Permanent Snow
Reg Term L_T
Correction matrix W
Reg Term L_P
$A^T * W$
c) Network pre-training with Geographical supervision and priors
Student Network
./2
./4
./8
./32
$\theta_t = \alpha\theta_t + (1-\alpha)\theta_s$
Teacher Network
Prediction of student networks
0 1 2 3 4 5 6 7
MSE Loss L_M
Geographical supervision $\tilde{A}$
0 1 2 3 4 5 6 7
Consistence Loss L_C
Prediction of teacher networks
0 1 2 3 4 5 6 7
Loss function
L= $\alpha_m L_M + \alpha_c L_C + \alpha_t L_T + \alpha_P L_P$
Joint optimization
Update W
Update network parameters θ_s
d) Geographical initialization
Initialized by pre-trained teacher model
Random Initialization
Cloud & Snow Detection
Land Cover Classification
Semantic Segmentation

	0	1	2	3	4	5	6	7
0	0.88	0.00	-0.01	-0.00	-0.02	-0.02	-0.00	-0.00
1	-0.02	0.99	-0.01	-0.00	-0.01	-0.02	-0.00	-0.00
2	-0.01	0.00	0.97	-0.00	-0.01	-0.01	0.00	-0.00
3	-0.01	-0.00	-0.01	0.98	-0.00	-0.01	-0.00	-0.00
4	-0.02	-0.01	-0.01	-0.00	0.97	-0.02	-0.01	-0.00
5	-0.02	-0.00	-0.01	-0.00	-0.02	0.92	0.00	-0.00
6	-0.01	0.00	-0.01	-0.01	-0.01	-0.02	0.96	-0.00
7	-0.02	-0.02	0.00	0.00	-0.02	-0.03	-0.00	0.94

Fig. 1. Detailed structure of the proposed method. As the generated geographical supervision contains deviations, the proposed method is designed to learn effective representations from deviated geographical supervision. It can adaptively correct deviated geographical supervision. Regularization terms can help to keep the adaptive correction process from collapsing.

adaptive correction of geographical supervision based on two priors, named as "minor changes" and "spatial aggregation". The prior knowledge "minor changes" reflects the fact that geographical supervision may not change significantly over time, while prior knowledge "spatial aggregation" reflects the fact that land covers present an aggregated state in the spatial distribution. However, large deviations caused by major changes do occur on occasion. To cope with major changes, we adopt the reasonable geographical supervision generation strategy (see section III-B), the mean-teacher network structure and the stable correction matrix update method (see section III-D) to ensure that the network can ignore the impact of major changes as much as possible during the pre-training stage.

The mean-teacher structure [45], which has been extensively utilized in noise labels [46] and representation learning [21], consists of student and teacher networks. Student and teacher networks both have a similar structure. We begin with ResNet50 [47] as the backbone $f(\dots)$ and add a single fully connected layer to serve as the projection head $h(\dots)$. The outputs of student and teacher networks are represented as O_s and O_t.

The following formula expresses the total pre-training loss function:

$$L = \alpha_m * L_M + \alpha_c * L_C + \alpha_t * L_T + \alpha_p * L_P. \quad (4)$$

L_M is the student networks' prediction loss function in the form of mean square error (MSE):

$$L_M = \|O_s - \tilde{A}\|_2^2, \quad (5)$$

where O_s represents prediction of student networks and $\tilde{A}$ represents the corrected geographical supervision. L_C denotes the consistence loss function that may cause teacher networks to restrict the update magnitude of student networks in order to ensure a smooth training process:

$$L_C = \|O_s - O_t\|_2^2, \quad (6)$$

where O_s and O_t represent outputs of student networks and teacher networks, respectively.

L_T and L_P denote the regularization terms that constrain network pre-training and automatic supervision correction, where will be introduced in section III-C.

B. Geographical Supervision Extraction

Global land cover (GLC) products we use in this paper include GlobeLand30 [40], GLCFCS30 [41] and FromGLC30 [42]. Table I depicts detailed information about these products.

GlobeLand30: The GlobeLand30 [40] version of 2000 and 2010 become open access in 2014. The Ministry of Natural Resources launched GlobeLand30 update in 2017. The latest version is GlobeLand30 2020. It includes 10 land cover types in total, namely cultivated land, forest, grassland, shrubland, wetland, water bodies, tundra, artificial surface, bare land, permanent snow and ice. The whole product is divided into hundreds of regular tiles. Each tile contains a complete record of an area's land covers, which are saved as a separate TIF image.

GLCFCS30: GLCFCS30 [41] contains 30 types of land covers, which was first released in 2015 and updated in 2020. The whole product is also composed of hundreds of tiles.

FromGLC30: FromGLC30 [42] is available in three versions of 2010, 2015 and 2017. It contains the same ten common land cover types as GlobeLand30.

To conduct a unified analysis of various land cover products, we design land cover merging rules that primarily merge the land covers in GLCFCS30 to make them identical to GlobeLand30 and FromGLC30. Table II contains information about the types of land covers and the rules for merging them.

As they are not intended to be used with deep learning methods, we need to extract the appropriate information from these products and convert it to geographical supervision. Geographical supervision can be used to build loss functions for deep learning, ensuring that the pre-trained model is useful for subsequent remote sensing image processing tasks. The procedure for extracting geographical supervision is depicted as follow.

We extract the geographical location of each remote sensing image X_i firstly, which is represented by a vector as follows:

$$GT_i = [lon_s, delta\_lon, 0, lat_s, 0, delta\_lat], \quad (7)$$

where lon_s and lat_s represent the longitude and latitude of the image's upper-left corner. $delta\_lon$ and $delta\_lat$ denotes the pixel resolution, or the change in longitude and latitude caused by a single pixel movement.

As tiles are used to store GLC products, we should locate the tile that contains the remote sensing image X_i, denoted by T_j. Because the coverage area of each tile is significantly larger than the coverage area of the remote sensing image, we hypothesis that tile T_j contains the remote sensing image X_i as long as the latitude and longitude of the image's upper-left corner are contained within the tile. GT_j is used to represent the geo-location of tile T_j. Additionally, because GlobeLand30 is projected using the UTM coordinate system, we first convert the UTM coordinates of each tile to latitude and longitude coordinates.

The region of tile T_j that corresponds to the remote sensing image X_i can be cropped. Assuming the width and height of X_i are represented as w and h, the latitude and longitude of the image's bottom-right corner can be calculated using the following formulas:

$$lon_e = GT_i(0) + GT_i(1) * w, \quad (8)$$

$$lat_e = GT_i(3) + GT_i(5) * h. \quad (9)$$

The corresponding coordinates of the region in the tile can be calculated as follows using the position information for the image's upper-left and lower-right corners.

$$x_{min} = (GT_i(0) - GT_j(0))/GT_j(1), \quad (10)$$

$$y_{min} = (GT_i(3) - GT_j(3))/GT_j(5), \quad (11)$$

$$x_{max} = (lon_e - GT_j(0))/GT_j(1), \quad (12)$$

$$y_{max} = (lat_e - GT_j(3))/GT_j(3). \quad (13)$$

TABLE I
DETAIL INFORMATION ABOUT SELECTED GLOBAL LAND COVER (GLC) PRODUCTS, INCLUDING GLOBELAND30, GLCFCS30 AND FROMGLC30.

GLC Product	Region	Versions	Number of types	Satellite	Resolution
GlobeLand30 [40]	Global	2000, 2010, 2020	10	Landsat, HJ-1	30 meter
GLCFCS30 [41]	Global	2015, 2020	30	Landsat	30 meter
FromGLC30 [42]	Global	2010, 2015, 2017	10	Landsat	30 meter

TABLE II
DETAIL INFORMATION ABOUT THE TYPES OF LAND COVERS AND THE RULES FOR MERGING THEM.

GlobeLand30	FromGLC30	GLCFCS30
Cultivated land	Cropland	Rainfed cropland Herbaceous cover Tree or shrub cover (Orchard) Irrigated cropland
Artificial surface	Impervious surface	Impervious
Bareland	Bareland	Bare areas Consolidated bare areas Unconsolidated bare areas
Waterbodies	Water	Water body
Forest	Forest	Evergreen broadleaved forest Deciduous broadleaved forest Open deciduous broadleaved forest Closed deciduous broadleaved forest Evergreen needle-leaved forest Open evergreen needle-leaved forest Closed evergreen needle-leaved forest Deciduous needle-leaved forest Open deciduous needle-leaved forest Closed deciduous needle-leaved forest Mixed leaf forest
Grassland	Grassland	Grassland Lichens and mosses Sparse vegetation Sparse herbaceous
Wetland	Wetland	Wetlands
Permanent snow & ice	Snow/Ice	Permanent ice and snow
Shrubland	Shrubland	Shrubland Evergreen shrubland Deciduous shrubland Sparse shrubland
Tundra	Tundra	————————

Cropping tile T_j with the obtained position information yields the land cover coverage map M_i of X_i. For GLCFCS30, it is necessary to merge the land covers using merging rules. We use $M_i(k)$ to represent the number of land cover k and $A_i(k)$ to represent the proportion of land cover k in the image X_i, where $\sum_k A_i(k) = 1$.

Using the preceding procedure, geographical supervision for the image X_i can be generated using the three selected GLC products, GlobeLand30 2020, GLCFCS30 2020 and FromGLC 2017, denoted as A_i^{Globe}, A_i^{GLC} and A_i^{From}. The final supervision information is obtained by evaluating the average of them:

$$A_i = \{A_i^{Globe} + A_i^{GLC} + A_i^{From}\}/3. \tag{14}$$

During pre-training stage, A_i is utilized to calculate loss functions with corresponding image X_i.

C. Prior Knowledge and Corresponding Regularization Terms

We also need to teach the network the prior knowledge unrelated to the image content, such as prior knowledge of land cover change and spatial distribution. The prior knowledge can be deduced from the geographical supervision of the whole pre-training dataset, as denoted as $A = \{A_1, A_2, \dots, A_i, \dots, A_M\}$. M is the number of images in the pre-training dataset. The deduced prior knowledge are named "minor changes" and "spatial aggregation", which are described as follow.

The prior knowledge "minor changes": As the producing times of GLC productions are significantly different from ones of remote sensing images, this may result in the generation of deviated geographical supervision. In this paper, a correction matrix W is proposed to correct the deviated geographical supervision: $\widetilde{A} = A^T \times \mathbf{W}$. The prior knowledge "minor changes" represent the fact that deviations between the corrected geographical supervision and deviated geographical supervision should be as small as possible. This can be interpreted intuitively as meaning that land covers may not change significantly over times. For example, while cities, cultivated land and water bodies may change slightly along their edges, they rarely appear or vanish entirely, especially on a short time scale.

The prior knowledge "spatial aggregation": This prior knowledge is deduced from the fact that the spatial distribution of land covers is uneven, with noticeable spatial aggregation. For example, the fact that artificial surfaces comprise an absolute majority around cities, whereas forests and bareland occupy a minority. Forests and grassland dominate in mountainous locations, with artificial surfaces accounting for just a minor portion.

We design two corresponding regularization terms in order to apply the prior knowledge to representation learning.

Regularization term L_T: The regularization term L_T is deduced from the prior knowledge "minor changes", which states that the majority of land covers do not change significantly over time. By restricting the correction matrix W, it assures that adaptive correction of geographical supervision adheres to this prior. This regularization term is solely applicable to the correction matrix W, which is constructed as follows:

$$L_T = \|\mathbf{W} - I_W\|_F^2, \tag{15}$$

where I_W denotes the unit diagonal matrix. During the network pre-training, the prediction loss function tends to update W in response to the network's outputs, whereas the regularization term tends to maintain the geographical supervision unchanged. Finally, a trade-off will be reached, which implies that correction of deviated geographical supervision are made on the premise that the network can learn effective representations of geographical supervision.

Regularization term L_P: The regularization term L_P is derived from the prior knowledge "spatial aggregation", which states that land covers have an aggregated state in their spatial distribution and the majority of remote sensing images contain just one or two dominant land covers. It is accomplished by instructing student networks to output uneven predicted results using the following formula:

$$L_P = -P * logP, \tag{16}$$

where P denotes the softmax value generated by the student network, $P = softmax(O_s)$. W and θ represent the correction matrix and network parameters θ. It can also be regarded as a regularization term, denoted as L_P, to constrain the network pre-training and correction matrix updating. When all elements in P have the same value, L_P takes the maximum value. Conversely, when just one element in P is 1 and the others are 0, L_P is assigned the minimum value 0. Since network predictions participate in both the process of updating of W and network parameters θ, this regularization term can ensure that corrected geographical supervision meets the requirement "spatial aggregation" by updating W. Additionally, by directing the predictions of networks to satisfy the prior knowledge, it can prevent the network from producing excessively smooth results, hence avoiding the collapse of network pre-training.

D. Joint Optimization

We design a joint optimization process that includes network parameters θ and correction matrix W. It can ensure geographical supervision is corrected concurrently while network pre-training. It divided the parameter updating process into two stage: 1) updating θ with fixed W and 2) updating W with fixed θ.

Updating θ with fixed W: In the first stage, we fix the correction matrix W and update the network parameters θ. The network parameters θ can be divided into student network parameters θ_s and teacher network parameters θ_t. During the network's pre-training, θ_t does not participate in backward process, but instead receives updates from the student network using the moving average method at a predetermined interval:

$$\theta_t = \alpha * \theta_t + (1 - \alpha) * \theta_s, \tag{17}$$

where α controls the update speed of the teacher networks. The student network parameters θ_s are updated using the random gradient descent method.

Updating W with fixed θ: To ensure the stability of the correction matrix W update process and to avoid the impact of violent fluctuations in deviated geographical supervision correction during network pre-training, we employ the gradient accumulation method to gradually accumulate the gradients of W. After accumulating to a certain interval, we update W using the gradient descent method.

E. Implementation Details

We can obtain a pre-trained model of the networks following the pre-training stage. Model parameters that have been pre-trained contain information about remote sensing images and their associated geographical knowledge. We then design distinct deep learning networks for each remote sensing image processing task. Each of these networks is built on the same structure's backbone network f. The backbone network is initialized using the parameters of the teacher network in the pre-trained model. Each task necessitates the creation of a head network, with the parameters in the head initialized randomly.

We develop our codes using Pytorch-1.7. The input image is 256×256 pixels in size. For loss function's coefficients, we set $\alpha_m = \alpha_c = 100$ and $\alpha_t = \alpha_p = 10$. The learning rate is set to 0.001 for updating network parameters and 0.0005 for updating correction matrix. The training batch size is set to 114 and gradient's cumulative step for updating correction matrix is set to 8.

Due to the fact that shrubland and tundra are underrepresented and poorly distinguished from other land covers, we do not include them in the generation of geographical supervision. Consequently, there are eight land covers for representation learning, including cultivated land, artificial surface, bare land, water bodies, forest, grassland, wetland and permanent snow & ice. Therefore, the correction matrix W has dimensions of 8×8. The correction matrix is independent of the input remote sensing images and is related to land covers. Consequently, the final optimized correction matrix documents the correction of various land cover across the entire dataset.

IV. Experimental Results

A. Experimental Setup

We will conduct thorough evaluation of our method based on the following aspects. The generated geographical supervision is first analyzed to determine the plausibility of our proposed prior knowledge. Then, to validate the effect of our method, we compare it to contrastive learning methods in computer vision and representation learning methods in remote sensing images on downstream tasks such as scene classification, semantic segmentation, and object detection. To visually analyze the effect of our method, we also visualize the initial features of the pre-trained models obtained by different methods on the downstream dataset. Finally, we conduct ablation experiments to assess the impact of various strategies. The pre-training dataset, the used comparison method, the dataset, the network structure, and implementation details on downstream tasks are all described below.

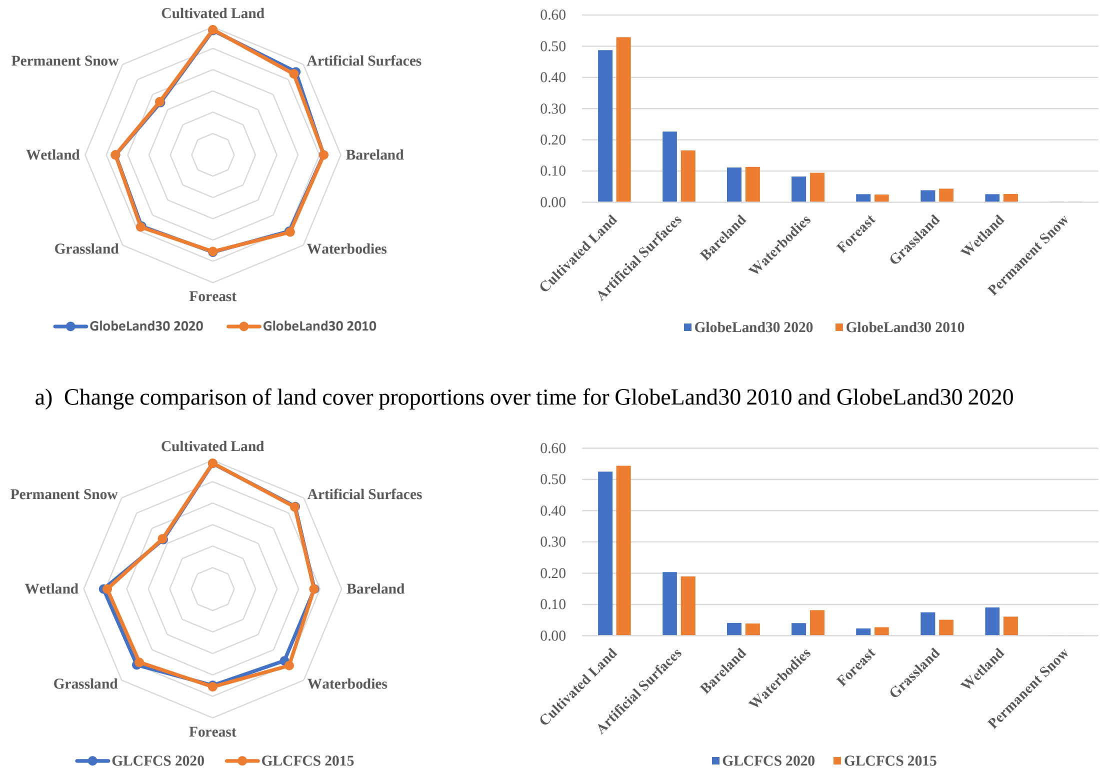
The remote sensing images in the pre-training dataset originate from the Levir-KR dataset, which is described in detail in [39]. Images are captured using a variety of sensors and resolutions, including gaofen-1 multi-spectral sensor (GF1-PMS) with resolution of 2 meters, gaofen-2 multi-spectral sensor (GF2-PMS) with resolution of 0.8 meters, gaofen-1 wide file of view (GF1-WFV) with resolution of 16 meters, and gaofen-6 wide file of view (GF6-WFV) with resolution of 16 meters. All images are converted to RGB images with 256×256 pixels for processing convenience.

TABLE III
RMSE VALUES BETWEEN VARIOUS LAND COVERS IN GLC PRODUCTS WITH DIFFERENT RELEASE TIME.

	Cultivated Land	Artifical Surfaces	Bareland	Waterbodies	Foreast	Grassland	Wetland	Permanent Snow
GlobeLand30 2010 and 2020	0.2812	0.2356	0.0421	0.2057	0.0675	0.1240	0.1252	0.0135
GLCFCS 2015 and 2020	0.1894	0.1179	0.0.0467	0.1451	0.0.0774	0.1220	0.1528	0.0071

TABLE IV
MAE VALUES BETWEEN VARIOUS LAND COVERS IN GLC PRODUCTS WITH DIFFERENT RELEASE TIME.

	Cultivated Land	Artifical Surfaces	Bareland	Waterbodies	Foreast	Grassland	Wetland	Permanent Snow
GlobeLand30 2010 and 2020	0.1117	0.0810	0.0035	0.0552	0.0072	0.0212	0.0195	0.0003
GLCFCS 2015 and 2020	0.1095	0.0539	0.0083	0.0437	0.0131	0.0561	0.0593	0.0004



a) Change comparison of land cover proportions over time for GlobeLand30 2010 and GlobeLand30 2020

b) Change comparison of land cover proportions over time for GLCFCS30 2015 and GLCFCS30 2020

Fig. 2. (better view in color) Change comparison of land cover proportions over times. a) For GlobeLand30 2010 and GlobeLand30 2020. b) For GLCFCS30 2015 and GLCFCS30 2020.

We validate our method's effectiveness in remote sensing image scene classification, semantic segmentation, cloud / snow detection and object detection tasks using the corresponding downstream datasets. Different proportions of training data are used in order to evaluate the effect of different methods on different scale data. Comparison methods are grouped into three types. The first group includes random initialization and ImageNet pre-training. They are the widely used initialization methods. The second group includes three recent contrastive learning methods: MoCo [21], SimCLR [22], and BYOL [23]. Finally, the third group of comparison method is GeoKR [39] method for remote sensing image representation learning. It can be regarded as the method that learn representations from deviated geographical supervision. All pre-training methods obtained from methods in the last two groups are performed on the identical Levir-KR dataset [39], and the same comparative experiment employs the identical training strategy.

For scene classification task, UCMerced [48] and RSSCN7 [49] datasets are used to fine-tune various pre-training models. The UCMerced dataset is divided into 21 categories, each of which contains 100 images measuring 256×256 pixels in size. Its resolution is 0.3 meters. The RSSCN7 dataset is divided into seven categories, each of which contains 400 images measuring 400×400 pixels in size. For processing convenience, we downsample them to 256×256. Each dataset is divided into a training set and a test set in a 3:1 ratio. In terms of network settings, we add a fully connected layer for classification after the global pool of pre-training networks. The learning rate is set to 0.005, and the model is trained for a total of 200 epochs. The final evaluation index is the average accuracy of all categories.

For semantic segmentation tasks, we use the Vaihingen [50] dataset with resolution of 0.09 meters to examine our method's effect. As cloud and snow detection are fundamentally related to the remote sensing semantic segmentation problem, we select Levir_CS dataset [51] to evaluate our method's performance on cloud and snow detection task. The Vaihingen dataset contains six categories of land covers. We divide it into a training set, a validation set, and a test set in the ratio 3:1:1, and then crop each image into patches measuring 256×256 pixels in size. Levir_CS is a dedicated dataset for cloud and snow detection in remote sensing images, which are captured from Gaofen-1 WFV with resolution of 16 meters. We proportionately divide it into training, validation, and test sets, and then down-sample each image to 256×256 pixels. In terms of network configuration, the portion of the pre-training network following global pooling is removed and replaced by up-sampling and convolution layers. For the Vaihingen dataset, the learning rate is set to 0.005 and for the Levir_CS dataset, it is set to 0.001. Each model is trained for 200 epochs, with the learning rate decreasing to 90% after every 20 epochs. During training, the mIoU (mean average of Intersection-over-Union) of the validation set is calculated after every 20 epochs, and the model with the highest mIoU is retained as the final result.

In addition, we use the Levir dataset [52] with resolution 0.2 to 1 meters to validate our method's performance on the object detection task. Three objects are included in the Levir dataset: airplanes, ships, and oil tanks. The dataset is divided into a training set, a validation set, and a test set in the ratio 3:1:1. In terms of network configuration, the CenterNet [53] network structure is used. The learning rate is set to 0.005. After 200 epochs, the learning rate drops to 90%. Each model is trained for 2000 epochs. We calculate the mAP (mean average precision) of the validation set every 50 epochs trained and save the best model as the final result.

B. Prior Knowledge Assessment

This section will examine the rationality and effectiveness of prior knowledge. Calculating the degree of deviations quantitatively explains the rationality of the proposed prior knowledge ”minor changes.” The effectiveness of the prior knowledge ”spatial aggregation” is demonstrated through an analysis of the image's dominant land cover.

Root mean squad error (RMSE) and mean absolute error (MAE) are used as indicators to measure the land cover distribution of different products, with each single image serving as the smallest analysis unit. The following is the calculation formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i^x - A_i^y)^2}, \tag{18}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i^x - A_i^y|, \tag{19}$$

where A_i^x and A_i^y represent generated supervision information of ith image with different GLC products x and y. Additionally, for the convenience of the subsequent analysis, we specify that the total number of land cover j generated by GLC product x in the pre-training dataset is $S_x(j)$, as calculated using the following formula:

$$S_x(j) = \sum_{i=1}^{n} A_i^x(j). \tag{20}$$

We use $P_x(j)$ to represent the proportion of land cover j, which is calculated as follows:

$$P_x(j) = S_x(j) / \sum_{k} S_x(k). \tag{21}$$

In the following experiment, we depict P_x as a histogram and $log(S_x)$ as a radar chart to intuitively show comparison of land covers with varying proportions.

1) Analysis of Land Cover Changes over Time: We analyze land cover changes over time using the GlobeLand30 versions released in 2010 and 2020, as well as the GLCFCS30 versions released in 2015 and 2020. Tables III and Table IV list RMSE and MAE values between various land covers. Fig. 2 illustrates a comparison of land cover proportions over time.

As can be seen, GlobeLand30 and GLCFCS30 exhibit consistency with time change, as evidenced by the fact that the change proportion is roughly equal for each land cover. Additionally, the evolution of various land covers over time demonstrates some differences. But even the most changed land covers such as cultivated land, artificial surfaces, and waterbodies, the changes are still minor in comparison to the total amount. The analysis above demonstrates that the majority of land covers undergo only minor changes on a spatial or temporal scale, and rarely undergo significant changes.

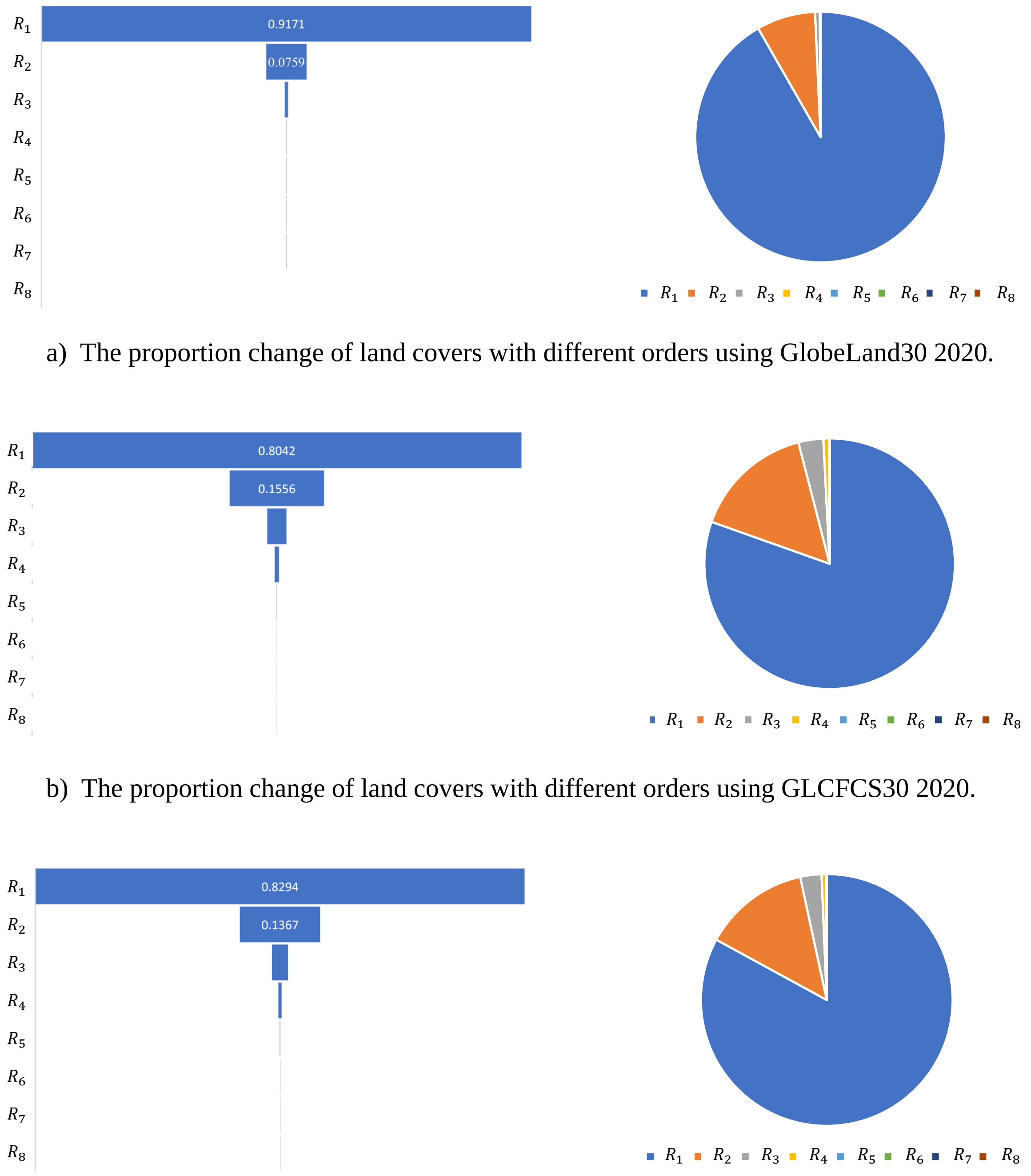
2) Analyses of Land Covers' Spatial Aggregation: For each remote sensing image X_i, we compute the geographical supervision A_i and sort it by the proportion of each land cover in descending order. The land cover with the highest proportion is denoted by the symbol $R_{i,1}$, while the one with the lowest proportion is denoted by $R_{i,8}$. We use the following formula to calculate the average proportion value of the kth largest land covers for all images in the pre-training dataset.

$$R_k = \frac{1}{n}\sum_{i}^{n} R_{i,k}, \tag{22}$$

TABLE V

THE DISTRIBUTION OF LAND COVERS WITH VARYING PROPORTIONS IN THE PRE-TRAINING DATASET. DIFFERENT ROWS DENOTE THE RESULT OF VARIOUS GLC PRODUCTS, WHEREAS DIFFERENT COLUMNS DENOTE THE R_k.

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
GlobeLand30 2020	0.9171	0.0759	0.0063	0.0007	0.0001	0.0000	0.0000	0.0000
GLCFCS30 2020	0.8042	0.1556	0.0321	0.0073	0.0006	0.0001	0.0000	0.0000
FromGLC30 2017	0.8294	0.1367	0.0273	0.0056	0.0008	0.0001	0.0000	0.0000



a) The proportion change of land covers with different orders using GlobeLand30 2020.

b) The proportion change of land covers with different orders using GLCFCS30 2020.

c) The proportion change of land covers with different orders using FromGLC30 2017.

Fig. 3. The proportion changes of land covers with different R_k and different GLC products. a) The proportion change of land covers with different R_k using GlobeLand 2020. b) The proportion change of land covers with different R_k using GLCFCS30 2020. c) The proportion change of land covers with different R_k using FromGLC 2017.

where $R_{i,k}$ refers to the kth largest land covers for ith image in the pre-training dataset. R_k represents the kth largest land cover proportion for the pre-training dataset.

Table V lists values of R_1 to R_8. Different rows denote the result of various GLC products, whereas different columns denote the R_k. Fig. 3 illustrates the proportional change qualitatively using funnel chart and pie chart. As can be observed, the proportion of land covers with the largest proportion has typically over 80%, while the proportion of land covers with the third largest proportion has been less than 5%, indicating that the majority of images contains only one or two types of land covers.

The objective of this experiment is to demonstrate that the spatial distribution of land covers is uneven, with noticeable "spatial aggregation". For example, the fact that artificial surfaces comprise an absolute majority around cities, whereas forests and bareland occupy a minority. Forests and grassland dominate in mountainous locations, with artificial surfaces accounting for just a minor portion.

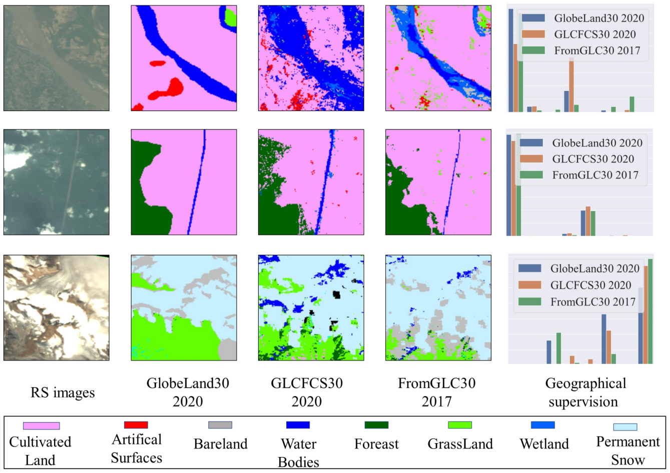


Fig. 4. (better view in color) Generated land cover map. Different colors are assigned to different land cover types. Histograms are used to display the geographical supervision.

Fig. 4 depicts land cover maps and geographical supervision generated using various GLC products to better illustrate the prior knowledge of land cover changes and spatial distribution. Each row is an illustration. The first column is the remote sensing image, the second through fourth columns are the visualization of land cover maps, and the fifth column is the geographical supervision calculated based on the various land cover maps. According to the proposed prior and assessment of the land covers, the change of land covers is generally small, and for the majority of images, the geographical supervision produced by different GLCs is comparable to the results shown in the Fig. 4.

Despite the fact that there are some deviations between the results generated by different products, the majority of them exist at the edges of land covers. This is likely due to the different resolutions of the image and GLC products, as well as the minor changes in the land covers. These deviations have no effect on the land cover types depicted on the land cover map and only result in minor variations in the values of the geographical supervision counterparts. Our proposed method can correct these deviations adaptively to enhance the performance of representation learning.

C. *Performance on Downstream Tasks*

1) Scene Classification: Table VI illustrates the experimental results. Different columns in Table VI represent various training proportions. Methods marked with * indicate that the representation learning starts from the ImageNet pre-training model.

As demonstrated by the experimental results, when compared to random initialization, our method significantly improves scene classification performance, especially when training data is scarce. At the same time, despite the fact that the ImageNet [20] pre-training model is not obtained with remote sensing images, it outperforms random initialization, indicating that it can provide some general image representations. Additionally, while contrastive learning methods are also pre-trained with remote sensing images, their performance is harmed by the problem of unstable training due to the large number of images with similar land covers. The performance of remote sensing representation learning can be significantly improved by incorporating geographical knowledge. However, when compared to the effect of GeoKR, it is clear that there are some deviations in the supervision information generated by geographical knowledge, which impairs representation learning performance. Our method can effectively mitigate these deviations and improve performance of downstream tasks.

2) Semantic Segmentation: Table VII presents the segmentation results for the Vaihingen dataset, while Table VIII and Table IX present the cloud and snow detection results. The first row of tables indicates the various training proportions.

The experimental results demonstrate that our method outperforms random initialization and ImageNet pre-training, particularly when training data is scarce. The comparison to GeoKR demonstrates that our method has the potential to significantly improve segmentation performance. This improvement is more noticeable when the training dataset is smaller. Indeed, when sufficient training data is available, we can observe a gradual narrowing of the gap between different methods. This demonstrates that having sufficient annotation data can also help mitigate the issue of deviated geographical supervision.

Fig. 5 displays the semantic segmentation results of various methods applied to the Vaihingen dataset [50]. As shown in the figure, our method improves the overall performance of semantic segmentation similar to the effect demonstrated by the quantitative results. Compared to contrastive learning methods, our method can improve discrimination accuracy for categories with less discrimination, such as "Tree" and "Low vegetation". These enhancements are most evident in the sharper segmentation edges obtained by our method. In addition, our method is more effective at extracting edge information from small targets, such as "car".

Fig. 6 illustrates the cloud and snow detection performance of various methods. As can be seen, various methods for cloud detection are effective. However, when it comes to snow

TABLE VI
Scene classification results on UCMerced and RSSCN7 dataset. Best results are marked in bold. Each column shows varying proportions of training data, whereas each row represents classification accuracy of different methods. Methods marked with * indicate that the representation learning starts from the ImageNet pre-training model.

	UCMerced					**RSSCN7**				
	5%	10%	20%	50%	100%	5%	10%	20%	50%	100%
Random	0.4247	0.5238	0.6362	0.7689	0.9105	0.4790	0.6143	0.6610	0.7179	0.8357
ImageNet	0.6622	0.8273	0.8806	0.9022	0.9562	0.6752	0.7829	0.8276	0.8593	0.9229
MoCo [21]	0.6533	0.7911	0.8381	0.9073	0.9378	0.6819	0.7790	0.8329	0.8710	0.9024
SimCLR [22]	0.5746	0.7403	0.8127	0.9168	0.9365	0.6914	0.7581	0.7919	0.8138	0.8710
BYOL [23]	0.5702	0.7632	0.8787	0.8737	0.9422	0.6948	0.7748	0.7733	0.8252	0.8714
GeoKR [39]	0.6229	0.7867	0.8368	0.8978	0.9416	0.8029	0.8548	0.8843	0.8924	0.9005
GeoKR* [39]	0.7048	**0.8470**	**0.9092**	0.9549	0.9695	0.8448	**0.8933**	0.8933	0.9152	**0.9419**
GeCo	0.6959	0.8063	0.8781	0.9384	0.9613	0.8286	0.8581	0.8900	0.9176	0.9391
GeCo*	**0.7448**	0.8425	0.8933	**0.9632**	**0.9746**	**0.8514**	0.8625	**0.9054**	**0.9304**	0.9404

TABLE VII
Semantic segmentation results on Vaihingen dataset. mIoU is used as the evaluation index. Best results are marked in bold. Each column shows varying proportions of training data, whereas each row represents mIoU of different methods. Methods marked with * indicate that the representation learning starts from the ImageNet pre-training model.

	0.25%	0.5%	1%	2%	5%	10%	20%	50%	100%
Random	0.3054	0.3369	0.3846	0.3757	0.4194	0.4727	0.5106	0.6309	0.6448
ImageNet	0.2974	0.3575	0.3470	0.4050	0.4640	0.4455	0.5177	0.6611	0.7015
MoCo [21]	0.3295	0.3463	0.4800	0.4354	0.5450	0.5929	0.6128	0.6406	0.6819
SimCLR [22]	0.3270	0.3349	0.4102	0.4450	0.5098	0.5939	0.5979	0.6417	0.6651
BYOL [23]	0.2325	0.3179	0.3976	0.4913	0.5925	0.6447	0.6829	0.6870	0.7271
GeoKR [39]	0.3634	0.4285	0.5165	0.5783	0.6209	0.6423	0.6796	0.6861	0.7110
GeoKR* [39]	0.3607	0.4155	0.5150	0.5665	0.6390	0.6397	0.6978	0.7159	0.7268
GeCo	**0.3802**	0.4411	**0.5323**	**0.5800**	0.6355	0.6582	0.6807	0.7108	0.7203
GeCo*	0.3754	**0.4466**	0.5297	0.5758	**0.6462**	**0.6641**	**0.7175**	**0.7228**	**0.7277**

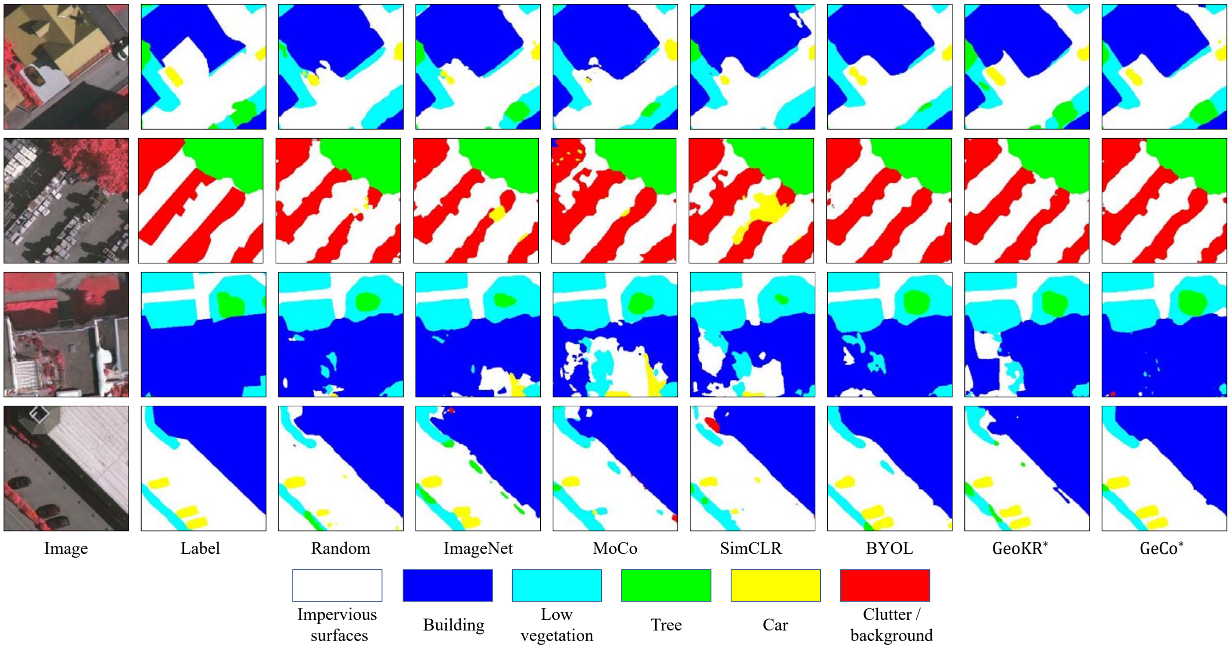


Fig. 5. (better viewed in color) Semantic segmentation results on Vaihingen dataset. The first column represents the input image, while the second column represents the label. The final seven columns depict results of different methods.

TABLE VIII

CLOUD DETECTION ON LEVIR_CS DATASET. IOU IS USED AS THE EVALUATION INDEX. BEST RESULTS ARE MARKED IN BOLD. EACH COLUMN SHOWS VARYING PROPORTIONS OF TRAINING DATA, WHEREAS EACH ROW REPRESENTS CLOUD DETECTION ACCURACY OF DIFFERENT METHODS. METHODS MARKED WITH $\star$ INDICATE THAT THE REPRESENTATION LEARNING STARTS FROM THE IMAGENET PRE-TRAINING MODEL.

	0.5%	1%	2%	5%	10%	20%	50%	100%
Random	0.6582	0.6364	0.6198	0.6905	0.6764	0.6889	0.7093	0.7320
ImageNet	0.6892	0.6586	0.7077	0.6817	0.6618	0.7219	0.7026	0.7344
MoCo [21]	0.6703	0.6611	0.6920	0.6755	0.6816	0.6853	0.7210	0.7338
SimCLR [22]	0.5994	0.6581	0.6363	0.6690	0.6532	0.7128	0.6651	0.7256
BYOL [23]	0.6712	0.6830	0.7020	0.7078	0.6881	0.7230	0.7349	0.7359
GeoKR [39]	0.6905	0.6954	0.7130	0.7181	0.7371	0.7563	0.7484	0.7622
GeoKR* [39]	0.6930	0.6989	0.7099	0.7337	0.7233	0.7332	0.7510	0.7507
GeCo	**0.6941**	0.7010	**0.7219**	**0.7355**	**0.7474**	0.7530	**0.7646**	**0.7765**
GeCo*	0.6898	**0.7018**	0.7121	0.7333	0.7409	**0.7587**	0.7502	0.7617

TABLE IX

SNOW DETECTION ON LEVIR_CS DATASET. IOU IS USED AS THE EVALUATION INDEX. BEST RESULTS ARE MARKED IN BOLD. EACH COLUMN SHOWS VARYING PROPORTIONS OF TRAINING DATA, WHEREAS EACH ROW REPRESENTS SNOW DETECTION ACCURACY OF DIFFERENT METHODS. METHODS MARKED WITH $\star$ INDICATE THAT THE REPRESENTATION LEARNING STARTS FROM THE IMAGENET PRE-TRAINING MODEL.

	0.5%	1%	2%	5%	10%	20%	50%	100%
Random	0.0057	0.1578	0.2513	0.2928	0.3190	0.4154	0.4012	0.4447
ImageNet	0.2326	0.2743	0.3425	0.2911	0.3115	0.3670	0.4060	0.4700
MoCo [21]	0.2631	0.3088	0.3176	0.3176	0.2750	0.3288	0.4349	0.4447
SimCLR [22]	0.1630	0.2599	0.2747	0.2942	0.3237	0.2635	0.3304	0.4031
BYOL [23]	**0.2669**	0.2704	0.3491	0.3722	0.3558	0.4221	0.4361	0.4486
GeoKR [39]	0.2025	0.3059	0.3778	0.4070	0.4749	0.5421	0.5102	0.5730
GeoKR* [39]	0.2532	**0.3566**	0.3938	0.4526	0.4006	0.4628	0.5055	0.4992
GeCo	0.1389	0.3101	**0.3943**	**0.4664**	0.4903	0.5266	**0.5547**	**0.5778**
GeCo*	0.2261	0.3342	0.3857	0.4411	**0.5005**	**0.5454**	0.5455	0.5375

TABLE X

OBJECT DETECTION ON LEVIR DATASET. MAP IS USED AS THE EVALUATION INDEX. BEST RESULTS ARE MARKED IN BOLD. EACH COLUMN SHOWS VARYING PROPORTIONS OF TRAINING DATA, WHEREAS EACH ROW REPRESENTS OBJECT DETECTION ACCURACY OF DIFFERENT METHODS. METHODS MARKED WITH $\star$ INDICATE THAT THE REPRESENTATION LEARNING STARTS FROM THE IMAGENET PRE-TRAINING MODEL.

	0.5%	1%	5%	10%	50%	100%
Random	0.0192	0.0522	0.2139	0.4678	0.7009	0.7178
ImageNet	0.0175	0.0551	0.3189	0.5250	0.7191	0.7370
MoCo [21]	0.0092	0.0589	0.3425	0.5787	0.6512	0.6826
SimCLR [22]	0.0092	0.0398	0.1632	0.5048	0.7243	0.7229
BYOL [23]	0.0071	0.0452	0.2396	0.5391	0.7215	0.7370
GeoKR [39]	**0.0715**	0.0729	0.3886	0.5979	0.7164	0.7288
GeoKR* [39]	0.0231	0.0740	0.3716	0.5969	**0.7395**	0.7632
GeCo	0.0230	0.0981	0.4048	0.5488	0.7370	0.7835
GeCo*	0.032	**0.1417**	**0.4671**	**0.6297**	0.7329	**0.7972**

detection, the various methods exhibit significant differences. There are two primary causes of errors: one is that clouds and snow cannot be correctly distinguished, as illustrated in the second and third rows of the figure, and the other is that land is misinterpreted as cloud or snow, as illustrated in the final row. Both errors can be significantly reduced using our proposed method. As can be seen from the results, GeCo improves the ability to discriminate and decreases the area of misjudgment, although there is also the issue of cloud and snow being indistinguishable. The GeCo method, in particular, can significantly reduce the probability of misjudging ground objects as snow.

3) Object Detection: The results of object detection are shown in Table X. Different columns in the table represent various training proportions. The experimental results demonstrate that our method significantly improves the object detection performance. When the proportion of labeled data is between 0.5% and 1%, all methods have a relatively small effect. It is proved that although representation learning can enhance object detection's effectiveness, it also requires a

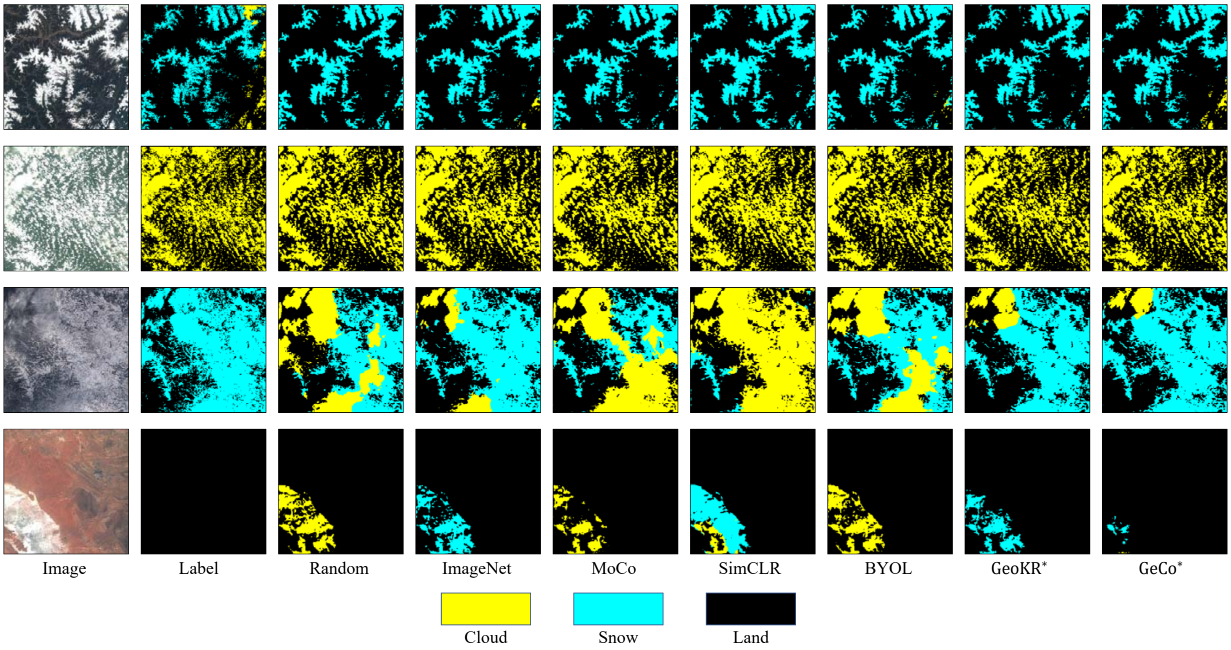


Fig. 6. (better viewed in color) Cloud / snow detection results on Levir_CS dataset. The first column represents the input image, while the second column represents the label. The final seven columns depict results of different methods.

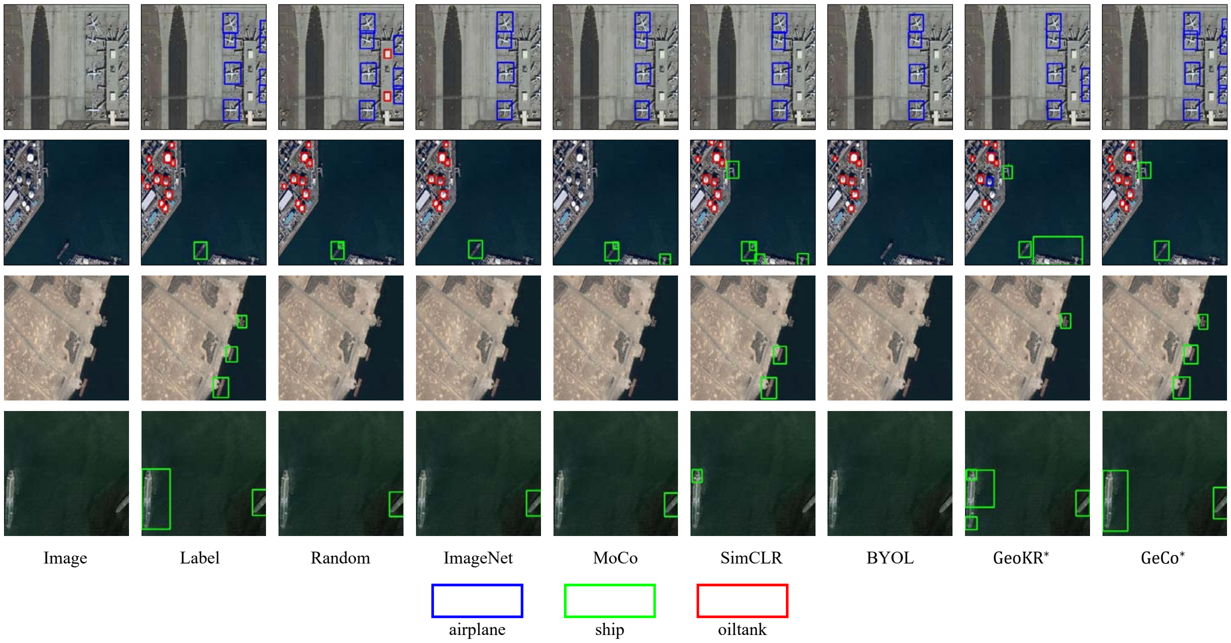


Fig. 7. (better viewed in color) Object detection results on Levir dataset. The first column represents the input image, while the second column represents the label. The final seven columns depict results of different methods.

certain amount of labeled downstream dataset. After utilizing more than 50% of the annotated data, almost all methods produced acceptable results, and their differences are becoming small. Our method, on the other hand, still increases object detection accuracy by approximately 3%, demonstrating that it can continue to increase performance without adding labels.

Fig.7 depicts the object detection results of various pre-training methods. It is typically challenging for deep learning-based object detection methods to identify objects with a dense distribution and a location near the image's edge. As shown in the figure, various methods for detecting these challenging targets are unstable, but our proposed method can improve detection performance and stability. As shown in the first row of the figure, almost all methods can accurately detect the airplane in the middle of the image, but many methods cannot detect the targets at the image's edge. However, our method may be accurate to detect them. In the second row of the figure, the detection results of certain dense oil tanks are displayed.

The majority of methods have issues with missed objects and false positives, but our method performs significantly better. In addition, our method can achieve favorable results for some onshore ship detection tasks, as shown in the last two rows.

4) Discussion: It can be seen from the above results, our method achieves the best results in the vast majority of semantic segmentation task and cloud / snow detection tasks, and in general outperforms GeoKR and other methods. On the scene classification task, our method performs marginally better than the GeoKR method, but in some cases, the GeoKR method achieves better results than our method. In the two cases with less training data, GeoKR achieves better results than our method for the object detection task.

On the classification task of scene classification, the reason why our method is not as effective as GeoKR in some cases may be that the scene classification task is relatively straightforward. Therefore, the task of scene classification may have a relatively high tolerance for deviations in land cover. As long as information about land covers is included in the pre-training model, the performance of the task can be effectively improved.

For the more difficult task of semantic segmentation and cloud / snow detection, our method yields the best results in almost every cases. From the visualization results, it is evident that our method is capable of achieving more refined semantic segmentation results. In the cloud and snow detection task, our method significantly improves the ability to differentiate cloud and snow and reduces the likelihood of misidentifying ground objects as cloud and snow.

On the object detection task, although our method is inferior to GeoKR in some cases, it is evident that the advantages of GeoKR are not particularly apparent in these cases. Similarly, in other cases, our method outperforms the GeoKR method by a significant margin. Therefore, in terms of overall performance, our method is superior and more stable. In conjunction with the visualization results, we are able to conclude that our method can achieve better results with more complex data.

In conclusion, when considering the overall performance of initialization, our method outperforms other methods in terms of both the number of cases with the best results and the refinement degree of results. Our method is significantly superior to other methods for improving downstream tasks, particularly for tasks that are more closely related to land covers or more difficult.

D. Feature Visualization

It is known that an effective initialization is crucial for deep learning methods. In addition to directly comparing the accuracy of pre-trained models on downstream tasks, the initialization effects of different pre-trained models can be visualized by visualizing their feature maps. In order to visualize the initialization effect of various pre-training models, we use them to extract the features of the downstream data and obtain the heat map of the extracted features via dimensionality reduction and normalization. The visualization of feature maps is shown in Fig. 8. In the visualized feature map, the larger the value, the larger the response to the corresponding region of the image. The first row represents the result of feature visualization on the scene classification, which are accomplished by up-sampling the output of the final backbone layer. The final three rows represent the feature visualization results for semantic segmentation, cloud / snow detection and object detection, respectively. Since these tasks require the use of multiple layer fusion strategies, we up-sample the outputs of the different pooling layers during visualization and then superimpose them to obtain the final feature map.

For the feature map of random initialization method, the response is nearly uniformly distributed around the image's perimeter and does not differ significantly between images. This demonstrates that the random initialization method can only provide an initialization scheme that allows the network to converge, but has no effect on the enhancement of downstream tasks. ImageNet initialization can provide discriminative features, but the ImageNet dataset lacks remote sensing images. It has nothing to do with land cover and doesn't seem very plausible. From the feature visualization results of contrastive learning, it is clear that these methods can learn some discriminative information from remote sensing images, but their level of refinement is not as good as ours. Due to the incorporation of geographical knowledge, it is evident that the initial features and edges generated by our method are more realistic and distinct. This may provide a more reasonable initialization, thereby enhancing the performance of downstream tasks.

E. Effectiveness Analysis on Geographical Supervision Correction

We firstly conduct ablation experiments to demonstrate the effect of each component of our method, as shown in Table XI, Table XII, Table XIII, Table XIV and Table XV. The baseline is constructed according to the GeoKR method [39]. The followings are the definitions of each ablation item:

- **GLC.** Utilize GLC products to generate geographical supervision for representation learning. The results are deviated from GeoKR.
- **Ensemble (Ens).** Utilize multiple GLC products to generate geographical supervision. Take the average value as the final supervision A.
- **Correction (Corr).** The Correction matrix is added to adaptively correct the deviated geographical supervision, and network parameters are updated using the joint optimization method.
- **Regularization (Reg).** The regularization terms are incorporated into the process of correcting geographical supervision.

The first row of each table represents the method that employs no strategy (baseline). ImageNet's initialization is served as the baseline for ablation experiment. The second row represents the addition to the baseline of geographical supervision by the GLC product. The third row represents the geographical supervision obtained by integrating multiple GLC products, while the last two rows represent the addition

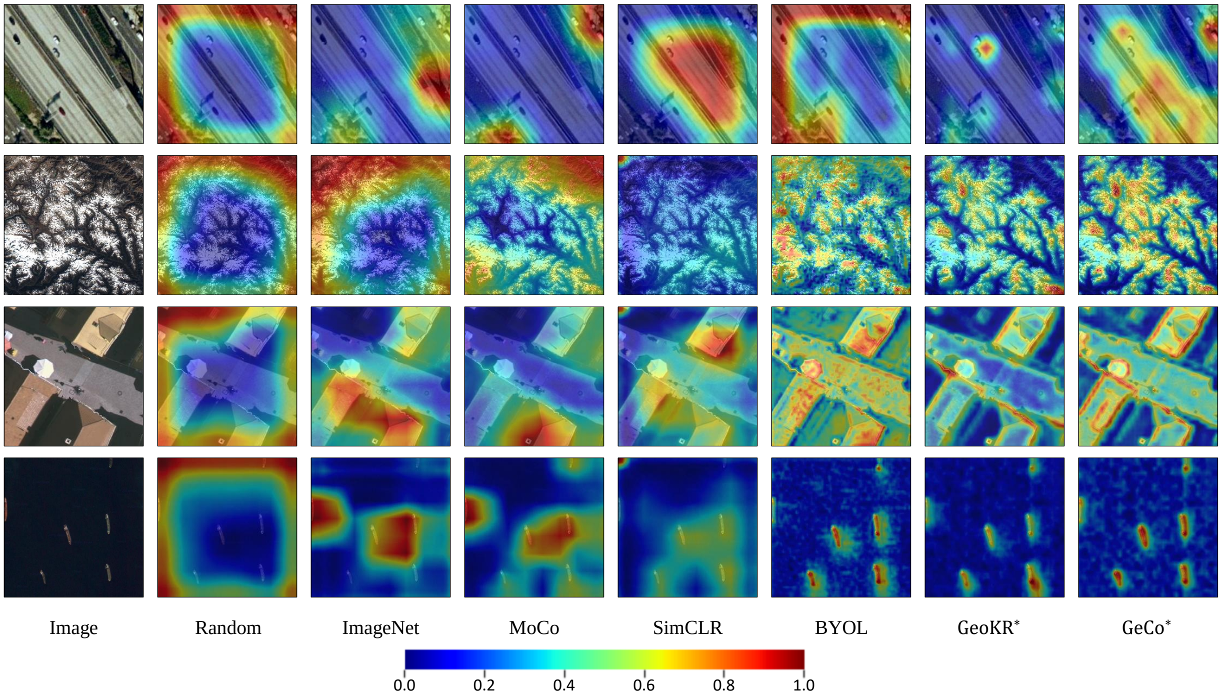


Fig. 8. (better view in color) Visualization of feature maps on downstream datasets. Different columns denote various pre-training methods.

TABLE XI
Ablation studies on UCMerced and RSSCN7 dataset. Ablations are performed on 1) GLC, 2) Ensemble (Ens), 3) Correction (Corr), 4) Regularization (Reg).

	Ablations			**UCMerced**					**RSSCN7**				
GLC	Ens	Corr	Reg	5%	10%	20%	50%	100%	5%	10%	20%	50%	100%
×	×	×	×	0.6622	0.8273	0.8806	0.9022	0.9562	0.6752	0.7829	0.8276	0.8593	0.9229
✓	×	×	×	0.7048	**0.8470**	0.9092	0.9549	0.9695	0.8448	**0.8933**	0.8933	0.9152	**0.9419**
✓	✓	×	×	0.6768	0.8222	**0.9054**	0.9619	0.9648	0.8510	0.8890	0.8881	0.9286	0.9329
✓	✓	✓	×	0.7149	0.8330	0.8889	0.9562	0.9721	0.8338	0.8619	0.8714	0.9286	0.9157
✓	✓	✓	✓	**0.7448**	0.8425	0.8933	**0.9632**	**0.9746**	**0.8514**	0.8625	**0.9054**	**0.9304**	0.9404

TABLE XII
Ablation studies on Vaihingen dataset [50]. Ablations are performed on 1) GLC, 2) Ensemble (Ens), 3) Correction (Corr), 4) Regularization (Reg).

GLC	Ens	Corr	Reg	0.25%	0.5%	1%	2%	5%	10%	20%	50%	100%
×	×	×	×	0.2974	0.3575	0.3470	0.4050	0.4640	0.4455	0.5177	0.6611	0.7015
✓	×	×	×	0.3607	0.4155	0.5150	0.5665	0.6390	0.6397	0.6978	0.7159	0.7268
✓	✓	×	×	0.3398	0.4426	0.4945	0.5425	0.6425	0.6572	0.7114	0.7253	0.7253
✓	✓	✓	×	0.3373	**0.4478**	0.4943	**0.6037**	0.6364	**0.6841**	0.7063	**0.7266**	**0.7350**
✓	✓	✓	✓	**0.3754**	0.4466	**0.5297**	0.5758	**0.6462**	0.6641	**0.7175**	0.7228	0.7277

TABLE XIII
Ablation studies on cloud detection. Ablations are performed on 1) GLC, 2) Ensemble (Ens), 3) Correction (Corr), 4) Regularization (Reg).

GLC	Ens	Corr	Reg	0.5%	1%	2%	5%	10%	20%	50%	100%
×	×	×	×	0.6892	0.6586	0.7077	0.6817	0.6618	0.7219	0.7026	0.7344
✓	×	×	×	0.6930	0.6989	0.7099	0.7337	0.7233	0.7332	**0.7510**	0.7507
✓	✓	×	×	0.6862	0.6926	0.7062	0.7179	0.7353	0.7485	0.7504	0.7650
✓	✓	✓	×	**0.6949**	0.6872	**0.7236**	**0.7397**	0.7284	0.7407	0.7483	**0.7684**
✓	✓	✓	✓	0.6898	**0.7018**	0.7121	0.7333	**0.7409**	**0.7587**	0.7502	0.7617

TABLE XIV
ABLATION STUDIES ON SNOW DETECTION. ABLATIONS ARE PERFORMED ON 1) GLC, 2) ENSEMBLE (ENS), 3) CORRECTION (CORR), 4) REGULARIZATION (REG).

GLC	Ens	Corr	Reg	0.5%	1%	2%	5%	10%	20%	50%	100%
×	×	×	×	0.2326	0.2743	0.3425	0.2911	0.3115	0.3670	0.4060	0.4700
✓	×	×	×	**0.2532**	**0.3566**	0.3938	**0.4526**	0.4006	0.4628	0.5055	0.4992
✓	✓	×	×	0.1860	0.3145	0.3581	0.4232	0.4794	0.4967	**0.5576**	**0.5475**
✓	✓	✓	×	0.2440	0.3269	**0.4001**	0.4513	0.4931	0.5079	0.5062	0.5336
✓	✓	✓	✓	0.2261	0.3342	0.3857	0.4411	**0.5005**	**0.5454**	0.5455	0.5375

TABLE XV
ABLATION STUDIES ON LEVIR DATASET. ABLATIONS ARE PERFORMED ON 1) GLC, 2) ENSEMBLE, 3) CORRECTION, 4) REGULARIZATION.

GLC	Ens	Corr	Reg	0.5%	1%	5%	10%	50%	100%
×	×	×	×	0.0175	0.0551	0.3189	0.5250	0.7191	0.7370
✓	×	×	×	0.0231	0.0740	0.3716	0.5969	0.7395	0.7632
✓	✓	×	×	**0.0331**	0.0870	0.4411	0.6035	0.7362	0.7815
✓	✓	✓	×	0.0187	0.1016	0.3998	0.6169	**0.7574**	0.7935
✓	✓	✓	✓	0.0320	**0.1417**	**0.4671**	**0.6297**	0.7329	**0.7972**

of our proposed adaptive correction and regularization, respectively.

When evaluating ablation experiments, in addition to comparing the accuracy of the method in different cases, it is necessary to consider the number of times it achieved the best results in different cases, which is a more accurate indicator of the method's overall effect. The results of the ablation experiments indicate that the method without geographical supervision is not as effective as the method with geographical supervision, demonstrating the effectiveness of geographical supervision. However, the "GLC" and "Ens" methods achieve the best results in a minority of cases. After incorporating our proposed adaptive correction and regularization, the methods achieve the best results in the vast majority of cases, particularly for semantic segmentation and object detection tasks. This also demonstrates that using adaptive correction can indeed result in the network learning a more effective representation, although the corrected geographical scale may no longer be practical without the addition of regularization. Once regularization is incorporated, it is possible to ensure that the corrected geographical supervision is more realistic. The performance of downstream tasks can be enhanced based on the experimental results. This demonstrates that our method can indeed produce a better pre-trained model by preserving the land cover information during the pre-training stage.

In order to demonstrate the effect of regularization on geographical supervision correction intuitively, we use a histogram to compare the distribution of geographical supervision obtained with and without regularization terms, as illustrated in Fig. 10. Without regularization terms, the corrected geographical supervision varies dramatically and the proportion of some land covers is negative, which is inconsistent with their real physical significance. This is because, in the absence of regularization terms, the network may prioritize geographical supervision correction in order to promote loss function decline. It may result in overfitting the networks, which is detrimental to the pre-training model's migration to downstream tasks. Additionally, the advantage of using representation learning with geographical supervision is that it can be used to learn about land cover information, which is frequently associated with specific remote sensing image processing tasks. The regularization terms are intended to constrain the network in order to maximize pre-training effectiveness and to ensure that the corrected geographical supervision still corresponds to the original distribution of land covers.

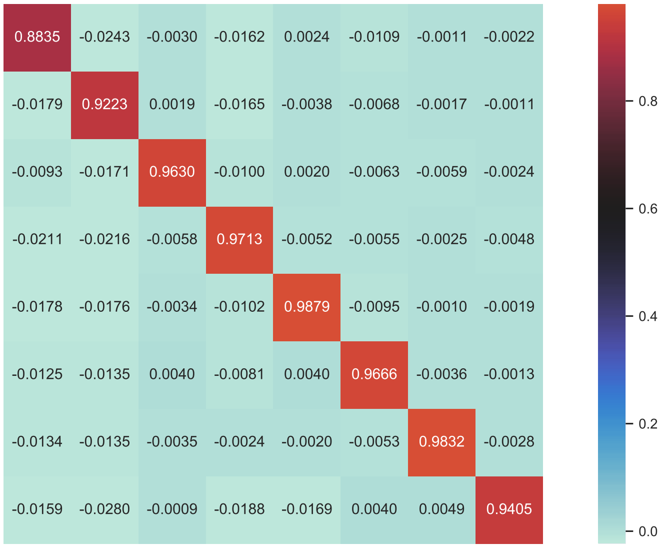


Fig. 9. The specific values of correction matrix.

Fig. 9 displays the specific values of correction matrix. Each column represents the correction coefficient of each land cover, corresponding to cultivated land, artificial surface, bare land, water bodies, forest, grassland, wetland and permanent snow & ice respectively. In each column, the diagonal element represents the proportional change at the corresponding position, while the other positions represent the proportion of other land covers that are incorrectly classified as this land cover.

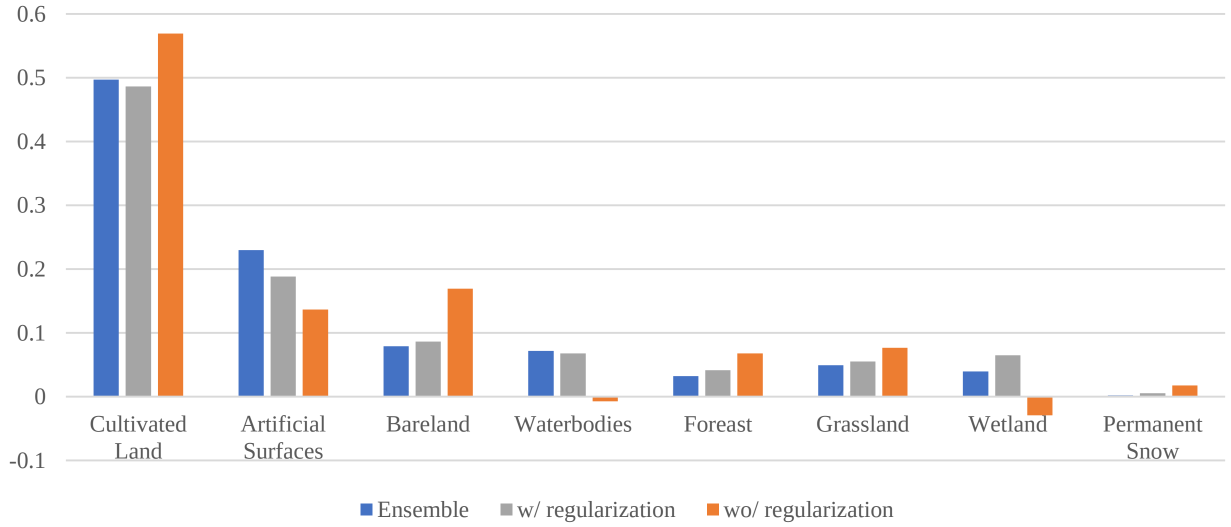


Fig. 10. A histogram to compare the distribution of geographical supervision obtained with and without regularization terms. For each land cover, bars marked with different colors from left to right are the uncorrected geographical supervision, the corrected geographical supervision with and without regularization terms.

Consistent with the proposed prior ”minor change”, it can be observed that the diagonal element is always the largest of all the elements. The smaller the value of the diagonal element, the greater the potential land cover deviation. Consequently, greater correction efforts are necessary. In conjunction with tables III and IV, we can observe that the diagonal element value of the correction matrix decreases as the potential deviations of land cover increase. For example, among all land covers, cultivated land and artificial surface have the greatest possible deviations, so their corresponding values in correction matrix are the smallest. This further proves that our correction method is effective.

The above experiments demonstrate that even though the correction matrix we employ is linear, it can still ensure the efficacy and rationality of correction. Nonetheless, it is still a very worthwhile research direction to investigate various correction matrix design schemes. We will investigate the effect of nonlinear correction methods and the method of embedding land cover priors in the nonlinear correction matrix in future research.

V. Conclusion

We propose an adaptive correction method of deviated geographical supervision to improve the performance of representation learning. Deviated geographical supervision, aroused by the disparity in producing times between GLC products and remote sensing images, may affect the performance of representation learning. We define a correction matrix that enables adaptive correction of deviated geographical supervision during network pre-training. The joint optimization process is designed to ensure that both the correction matrix and network parameters are updated in a timely and reasonable manner. We perform a systematic analysis of the generated geographical supervision and discover two prior terms, “minor changes” and “spatial aggregation”, that can restrict the correction of geographical supervision. According to the prior knowledge, two regularization terms are constructed. The regularization term L_T is deduced from prior knowledge “minor changes”. By bringing the correction matrix close to the identity matrix, it guarantee that the geographical supervision not change significantly. The other regularization term L_P is deduced from prior knowledge “spatial aggregation”. It can guide networks to produce uneven predictions and keep networks from producing excessively smooth results. We demonstrate our method’s effectiveness in scene classification, semantic segmentation, and object detection of remote sensing images. In comparison to three different types of comparison methods, our proposed method consistently outperformed the others across a range of downstream tasks and training scales. Experiments demonstrate that our proposed method can effectively eliminate impact of deviations and enhances the effect of representation learning.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, “Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[2] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, “More diverse means better: Multimodal deep learning meets remote-sensing imagery classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[3] Z. Zou and Z. Shi, “Ship detection in spaceborne optical image with svd networks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[5] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, “Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images,” *ISPRS Journal of*

Photogrammetry and Remote Sensing, vol. 146, pp. 182 – 196, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618302612

[6] P. Ding, Y. Zhang, W.-J. Deng, P. Jia, and A. Kuijper, "A light and faster regional convolutional neural network for object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 208 – 218, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618301382

[7] W. Li, Z. Zou, and Z. Shi, "Deep matting for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8490–8502, 2020.

[8] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 197 – 212, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271619300565

[9] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 60 – 77, 2018, deep Learning RS Data. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618301229

[10] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5367–5376, 2020.

[11] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sensing*, vol. 13, no. 21, p. 4441, 2021.

[12] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[13] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-

[14] S. Jenni and P. Favaro, "Self-supervised feature learning by learning to spot artifacts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2733–2742.

[15] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2547–2555.

[16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[17] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European conference on computer vision*. Springer, 2016, pp. 577–593.

[18] ——, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6874–6883.

[19] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1058–1067.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[23] 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," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[24] 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, 2020.

[25] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

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

[27] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[28] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

[29] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[30] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

[31] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8392–8401.

[32] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, and J. Shao, "Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2208–2217.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[34] J. Wang, H. Wang, J. Deng, W. Wu, and D. Zhang, "Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling," *arXiv preprint arXiv:2109.04699*, 2021.

[35] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.

[36] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190.

[37] O. Mañas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[38] J. Hall, K. Roth, and S. Kefauver, "Openstreetmap: Challenges

and opportunities in machine learning and remote sensing," *IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE*, vol. 9, no. 1, pp. 283–286, 2021.

[39] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[40] J. Chen, J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu *et al.*, "Global land cover mapping at 30 m resolution: A pok-based operational approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, pp. 7–27, 2015.

[41] X. Zhang, L. Liu, X. Chen, Y. Gao, S. Xie, and J. Mi, "Glc_fcs30: Global land-cover product with fine classification system at 30 m using time-series landsat imagery," *Earth System Science Data Discussions*, pp. 1–31, 2020.

[42] P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, S. Liu *et al.*, "Finer resolution observation and monitoring of global land cover: First mapping results with landsat tm and etm+ data," *International Journal of Remote Sensing*, vol. 34, no. 7, pp. 2607–2654, 2013.

[43] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "Clusterfit: Improving generalization of visual representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6509–6518.

[44] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[45] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1195–1204.

[46] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5051–5059.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[49] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

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

[51] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 87–104, 2021.

[52] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[53] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

Wenyuan Li received his B.S. degree from North China Electric Power University, Beijing, China in 2017. He is currently working toward his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include self-supervised learning and remote sensing image processing.

Keyan Chen received the B.S. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019. He is currently working toward the M.S. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include image processing, machine learning and pattern recognition.

Zhenwei Shi (M'13) received his Ph.D. degree in mathematics from Dalian University of Technology, Dalian, China, in 2005. He was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar in the Department of Electrical Engineering and Computer Science, Northwestern University, U.S.A., from 2013 to 2014. He is currently a professor and the dean of the Image Processing Center, School of Astronautics, Beihang University. His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Associate Editor for the *Infrared Physics and Technology*. He has authored or co-authored over 100 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Neural Networks, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters and the IEEE Conference on Computer Vision and Pattern Recognition. His personal website is http://levir.buaa.edu.cn/.