TOV: The Original Vision Model for Optical Remote Sensing Image Understanding via Self-Supervised Learning

Chao Tao, Member, IEEE, Ji Qi, Member, IEEE, Guo Zhang, Member, IEEE, Qing Zhu, Weipeng Lu, and Haifeng Li, Member, IEEE

Abstract—Are we on the right way for remote sensing image understanding (RSIU) by training models in a supervised data-dependent and task-dependent manner, instead of original human vision in a label-free and task-independent way? We argue that a more desirable RSIU model should be trained with intrinsic structure from data rather than extrinsic human labels to realize generalizability across a wide range of RSIU tasks. According to this hypothesis, we proposed The Original Vision (TOV) model in the remote sensing field. Trained by massive unlabeled optical data along a human-like self-supervised learning (SSL) path that is from general knowledge to specialized knowledge. TOV model can be easily adapted to various RSIU tasks, including scene classification, object detection, and semantic segmentation, and outperforms dominant ImageNet supervised pretrained method as well as two recently proposed SSL pretrained methods on the majority of 12 publicly available benchmarks. Moreover, we analyze the influence of two factors on the performance of building TOV model for RSIU, including the influence of using different data sampling methods and the selection of learning paths during self-supervised optimization. We believe that a general model which is trained in a label-free and task-independent way may be the next paradigm for RSIU and hope the insights distilled from this study can help to foster the development of an original vision model for RSIU.

Manuscript received 9 December 2022; revised 7 February 2023 and 2 April 2023; accepted 25 April 2023. Date of publication 28 April 2023; date of current version 7 June 2023. This work was supported in part by the Open Project of Xiangjiang Laboratory under Grant 22XJ03007, in part by the National Natural Science Foundation of China under Grant 42171376 and Grant 41771458, in part by the National Science Foundation for Distinguished Young Scholars of Hunan Province under Grant 2022J10072, in part by the National Science Foundation of Hunan under Grant 2021JJ30815, in part by the Inner Mongolia Science and Technology Plan under Grant 2022YFS0014, and in part by the High Performance Computing Center of Central South University. (Corresponding author: Haifeng Li.)

Chao Tao is with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China, and also with the Xiangjiang Laboratory, Central South University, Changsha 410205, China (e-mail: kingtaochao@csu.edu.cn).

Ji Qi and Haifeng Li are with the School of Geosciences and Info-Physics, Central South University, and also with Key Laboratory of Ecological Environment Protection of Space Information Application of Henan, Changsha 410083, China (e-mail: jameschi95@foxmail.com; lihaifeng@csu.edu.cn).

Guo Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: guozhang@whu.edu.cn).

Qing Zhu is with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: zhuq66@263.net).

Weipeng Lu is with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Kowloon, Hong Kong 100872, China (e-mail: weipeng.lu@connect.polyu.hk).

Digital Object Identifier 10.1109/JSTARS.2023.3271312

The source code is available at https://github.com/GeoX-Lab/G-RSIM/tree/main/TOV_v1.

Index Terms—Human vision, original vision model, pretraining models, remote sensing image understanding (RSIU), self-supervised learning (SSL).

I. INTRODUCTION

Human vision is a natural capability with which one can easily perform remote sensing image understanding (RSIU) from coarse (scene) to fine (object) without any task-oriented learning. Modern RSIU has achieved remarkable progress based on a machine vision model rather than human vision via teaching a machine to complete specific RSIU tasks, such as scene classification, object detection, and semantic segmentation [1], [2], [3], [4], [5], through supervised training a task-specific model with human-labeled task-specific data [6], [7], [8]. Thus, comparing what human vision does, a natural question is raised: Can this “teaching” method really solve the problem of RSIU? In fact, this “teaching” way has many limitations in constructing visual models for RSIU as follows.

1) Model training relies too much on large-scale, high-quality labeled data. When teaching machines to complete RSIU tasks, human labels can be regarded as a kind of knowledge. Therefore, the more knowledge learned, the better the model performance [9], [10], [11]. However, building a big remote sensing dataset is very challenging, as the accurate annotation of remote sensing images (RSIs) is tedious and requires rich experience and geographic knowledge. Moreover, the annotating approach used for RSIU tasks is extremely task dependent. For example, a scene classification task requires image-level annotation while a semantic segmentation task requires pixel-level annotation. Obviously, different labeling methods have different cost, which means a great number of effort need to be paid for constructing datasets in a task-dependent way.

2) More importantly, taking manual labels as supervised signals alone cannot learning the desired vision model itself because the only function of manual labels, as extrinsic supervised signals, is to guide a model fitting to given training data. On the opposite, the intrinsic information

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
hidden in massive remote sensing data should theoretically be much richer and more fundamental than the semantic information provided by human-labeled samples. Thus, human-labeled samples may be insufficient for annotating more complex scenes with multiple semantic meanings or with ambiguous semantic contents, which causes the problem of limited feature representation learning.

Unlike the machine vision that is “taught” by labeled data, human-like vision is achieved by holistic and joint models that can simultaneously solve real-world problems in an unsupervised way [12]. The key reason is that the human visual recognition system is not limited to a specific task or specific dataset, and human language-based labels are not the prerequisite for constructing the human visual system. For example, a person who has never purposely learned any remote sensing knowledge can easily identify common objects, such as farmland, vehicles, and parking lots from optical RSIs. Inspired by this, we believe that training a general model in a label-free and task-independent way may be the next paradigm for RSIU, which is closer to the human visual process.

Most recently, a new machine learning paradigm, self-supervised learning (SSL), has emerged in the field of natural language process (NLP) and computer vision (CV) [13], [14], [15], [16]. Its main idea is to use human-designed task-agnostic SSL signals to generate pseudolabels for massive unlabeled data, thereby replacing human label to guide the model learning. Since the model trained by SSL signals can be easily adapted to a wide range of downstream tasks, it can be considered a general model. For example, original models, such as BERT [13] and GPT-3 [14], have demonstrated significant effects on NLP. In the field of CV, researchers in Microsoft Research have built a fundamental or general vision model called Florence via unified image-text contrastive learning trained on web-scale image-text data [17] and showed that the model trained from massive unlabeled data in a task-independent way can adapt to a wide range of downstream tasks, such as classification, object detection, visual question answering, image caption, video retrieval, and action recognition.

In this article, we forge the concept, The Original Vision (TOV) model, in the two dimensions of data modality and task granularity. For the data modality, TOV can be served as the start point to be modified to generalize across various RSIU tasks and RS modalities. For the task granularity, TOV can be served as the start point to pretrain a very big deep neural network with unlabeled massive scale RSIs which are extremely easy to obtain by Earth observing system. Thus, as a new paradigm to overcome the transfer catastrophe problem, TOV as a model which is trained on broad and massive data via SSL at very large can be adapted to wide diverse application tasks with very limited labeled samples. Pioneer works [18], [19], [20], [21], [56], [57] have demonstrated surprising emergent capabilities on a wide range of downstream tasks and demonstrated the potential of SSL-based machine learning paradigm on RSIU tasks. Seen these results, we argue that TOV models as a growing paradigm shift, where many applications will be directly derived from TOV models. However, there are still two key questions that remain unclear as follows. 1) Are there general guidelines on creating benchmark datasets for TOV model learning and how to create them efficiently? 2) Under the SSL learning paradigm, how to train a high-performance TOV model from the perspective of model optimization?

This article focuses on constructing TOV model from the perspective of task granularity, that is, training TOV using optical RSIs in a task-independent way but can generalize across various RSIU tasks. Then, for the first question, though SSL method can be used to learn TOV model without labels, we experimentally find that the traditional grid sampling approach is not the best way on creating datasets for TOV model learning, since it may sample large amounts of semantically meaningless data. Besides, the classes in such datasets are severely imbalanced posing further challenges in SSL [22], [23], [24]. To solve this problem, we propose an automated data sampling and resampling mechanism guided by geographic data products, such as OSM and FROM-GLC10, to formulate a prototype for constructing a massive, scalable, and relatively balanced RSI dataset for training TOV model. In this way, we efficiently create two datasets collected from Google Earth, which contain 3 million class-imbalanced RSIs and 0.5 million relatively class-balanced RSIs, respectively. For the second question, though previous works have demonstrated the potential of using SSL for training TOV model in CV [25], there is a gap of model generalization compared to traditional supervised learning (e.g., ImageNet pretrained model) if it is applied in RSIU tasks. We argue that only using RSIs for TOV model learning may limit the discrimination and generalizability of the learned feature due to the characteristic of lower spatial resolution and nonobject centralization of RSI training samples. Thus, we propose a novel human-like SSL learning mechanism, which first learns general knowledge from web-scale natural images and then learns domain-relevant specialized knowledge from unlabeled RSIs. Experiments have shown that TOV model learned by the proposed method can adapt to various RSIU tasks (e.g., scene classification, object detection, and semantic segmentation) and achieves the state-of-the-art results in the majority of 12 publicly available RSIU benchmarks. In summary, our contributions are as follows.

1) We first define the TOV model for RSIU and analyze the influences of using different data sampling methods and the selection of learning paths during self-supervised optimization.

2) We release the benchmark dataset for training the TOV model as well as the pretrained model, which can help to foster the development of TOV model in the remote sensing community.

3) We propose a novel human-like SSL learning mechanism inspired by the following insights. Only using RSIs for TOV model learning may limit the discrimination and generalizability of the learned feature due to the characteristic of lower spatial resolution and nonobject centralization of RSI training samples.
Experiments have shown that the TOV model learned by the proposed method can adapt to various RSIU tasks (e.g., scene classification, object detection, and semantic segmentation) and achieves the state-of-the-art results in the majority of 12 publicly available RSIU benchmarks. Compared with publicly available TOV models, our model achieves the state-of-the-art results on various RSIU tasks.

The rest of this article is organized as follows. Section II gives the definition of TOV model for RSIU. The details of constructing TOV model are presented in Section III. The generalization performance of TOV model on various RSIU tasks is evaluated in Section IV. Finally, discussions and conclusions are presented in Sections V and VI, respectively.

II. Definition of TOV Model for RSIU

Although some original models, such as GPT-3 [14] and Florence [17], have demonstrated good performance in NLP and CV, the concept of TOV model is new in the field of remote sensing. Some scholars have paid attention to the potential of TOV models for RSIU and have proposed different construction solutions. With a novel one-million scale RSI dataset (Million-AID [11]) for scene classification, Wang et al. [26] pretrained a series of deep models in a supervised learning manner to benefit various downstream RSIU tasks. However, the supervised paradigm, which relies on high-quality labeled datasets, is significantly constrained by the sample volume and has only limited scalability. Therefore, most recent studies have been devoted to constructing pretraining models in a label-free and task-independent way [19], [21], [27]. For example, Manas et al. [19] constructed a one-million dataset, called SeCo, by automating the sampling of Sentinel-2 images of global urban areas and trained the TOV model with a SSL method based on SeCo. Further, Wang et al. [21] constructed an unlabeled dataset called SSL4EO-S12, which contains three million samples automated sampling from both Sentinel-1 and Sentinel-2 images and covers a wider spatio-temporal range than SeCo. Using SSL4EO-S12 and SSL methods (e.g., MoCov2 [28]), they pretrained a series of TOV models and demonstrated that TOV with good generalization performance can achieve comparable performance with fully supervised learning on several downstream task datasets. However, compared with the rapid development of TOV models in CV and NLP fields, the TOV model for RSIU remains underexplored, and its theoretical foundations are very limited. In particular, the definition of TOV model as a fundamental issue is still unclear. Thus, in this section, we first define TOV model for RSIU from the two dimensions of task granularity and data modality.

From the perspective of task granularity, RSIU tasks can be divided into three levels: scene-level, object-level, and pixel-level corresponding to understanding RSIs from coarse to fine. From the perspective of data modalities, RSIs can be divided into four modal data types, including optical (multispectral, hyperspectral), satellite video, and SAR, according to the imaging mechanism of remote sensing data and the difference in spectral sensing range. Thus, as shown in Fig. 1, RSIU tasks

![Diagram of RSIU containing task granularity and data modality dimensions.](image-url)
can be mapped to a problem space containing two dimensions of task granularity and data modality, and TOV model should be a general vision solution for all tasks in the abovementioned problem space. With this motivation, we define TOV model for RSIU to be a pretrained model trained in a task-independent and modality-independent way but can easily adapt to (e.g., fine-tuned) a wide range of RSIU tasks and data modalities.

Considering the complexity of the problem, this article focuses on constructing TOV model from the perspective of task granularity.

III. METHODOLOGY

A. Overall Framework for Constructing TOV Model for RSIU

Constructing TOV model for RSIU includes three stages: data acquisition, model pretraining, and task adaptation, as shown in Fig. 2.

Data acquisition: Rich and high-quality data are important for training TOV model. Although this model can be trained in a label-free SSL way, it may hardly learn valuable remote sensing visual knowledge from an unlabeled dataset that contains large amounts of semantically meaningless content or has severely class-imbalanced distribution. To address this problem, we propose an automatic RSIs sampling and resampling method (see Section III-B) guided by public geographic data products, which can automatically collect relative class-balanced RSI samples with rich semantic content over a global scale at low cost.

Model pretraining: Visual representation that generalizes well to different RSIU tasks should be both discriminative enough to train a strong classifier and invariance to significantly varying imaging conditions [29]. To achieve this, we propose a novel human-like SSL learning mechanism under the contrastive learning framework [30] for training TOV, which first learns general knowledge from web-scale natural image dataset and then learns domain-relevant specialized knowledge from the constructed RSI dataset (see Section III-C).

Task adaptation: TOV is expected to be adaptable to various RSIU tasks. Therefore, the learned representations are first stored as parameters in TOV model. Then, we adapt the learned general representation from TOV model to various RSIU tasks by adding task-specific adapters following the backbone of TOV model. Specifically, a simple fully connected layer is used as an adapter for scene-level understanding tasks; region proposal network and ROIHead in faster RCNN (i.e., a popular two-stage object detection model) [31] are jointly used as adapters for object-level understanding tasks; the common decoder in fully convolutional networks (FCN) [32] is used as an adapter for pixel-level understanding tasks. In this way, TOV can effectively adapt to RSIU tasks with few labeled data.
B. Data Acquisition for TOV Model Learning

Compared with supervised learning methods, SSL methods can learn visual knowledge from an unlabeled dataset, which provides a way to build TOV model for RSIU at a low cost. But whether models can learn valuable visual knowledge depends on the semantic content richness and class balance of the samples in the unlabeled dataset [18], [33]. Therefore, the unlabeled RSI dataset used for TOV model learning should have the abovementioned two key properties.

Without too much human involvement, constructing an RSI dataset with these two attributes is a challenging task. Although the traditional grid sampling approach is simple and straightforward, it may sample large amounts of semantically meaningless data. Besides, the classes in such datasets are usually severely imbalanced. The main reasons are as follows. First, geographic elements are naturally severely unbalanced in quantity at a global scale, as shown in Fig. 3 [34]. Second, the geographic elements shown in RSIs usually cannot be completely divided by the uniform grid. As a result, many samples contain mixed geographic elements making it even more impossible to control the class balance of the dataset. To solve these problems, we propose an automatic RSI sampling and resampling method guided by the geographic data product to efficiently build a large-scale RSI sample dataset.

Geographical elements can be divided into natural scene elements (e.g., forests, meadows, and water bodies) and man-made scene elements (e.g., residential areas, industrial areas, schools, and parking lots). These two types of geographic elements are significantly different in spatial distribution, temporal distribution, and the corresponding geographic data product forms (e.g., global land cover mapping product with the spatial resolution of 10 m (FROM-GLC10 [35])). Thus, we first design two sampling methods to, respectively, collect samples of these two categories (as shown in Fig. 4).

Then, the resulting dataset is resampled using the noise label provided by the geographic data product to obtain a relatively class-balanced dataset.

1) Data Sampling for Natural Scene Elements: Considering the characteristics of low-scale variance and slow temporal change of natural scene elements in RSIs, we use the global land cover mapping product with the spatial resolution of 10 m (FROM-GLC10 [35]) to guide the sampling of natural geographical elements shown in Table I. The sampling process consists of the following two steps.

Step 1: Automatic collection of candidate samples. To avoid the redundant and invalid sampling existing in traditional grid sampling methods, in this step, we utilize the region proposal mechanism that is commonly used in object detection, where the classic selective search algorithm [36] is used to select candidate regions from the input image for collecting candidate samples. Specifically, each image $I_i$ in the dataset $H = \{I_1, I_2, \ldots, I_{N_{img}}\}$ is first oversegmented into superpixels by a graph-based image segmentation method [37]. Then, adjacent small superpixels are continuously grouped based on the similarity metric established by color, texture, and shape features. In this way, image $I_i$ can

Then, the resulting dataset is resampled using the noise label provided by the geographic data product to obtain a relatively class-balanced dataset.

Table I: Natural Scene Categories

| Name       | Description                                                                 |
|------------|-----------------------------------------------------------------------------|
| Forest     | A large area of land that is thickly covered with trees.                    |
| Grassland  | A large area of open land covered with wild grass.                          |
| Shrubland  | Land on which shrubs are the dominant vegetation.                           |
| Cropland   | Land that is suited to or used for crops.                                   |
| Wetland    | Land covered by swamps or marshes.                                         |
| Water      | An area of water, especially a lake, river, sea or ocean.                   |
| Tundra     | Land where no trees grow and the soil below the surface is always frozen.   |
| Bareland   | Land covered with no vegetation or buildings.                               |
| Snow/Ice   | Land covered by snow or ice.                                                |

[2] [Online]. Available: https://www.webmap.cn/commres.do?method=globeDetails&type=GeographicalStatistics
be segmented into \( n_i \) segments. Finally, \( N_1 = \sum_{i=1}^{N_{\text{img}}} n_i \) candidate samples are collected by taking the minimum bounding rectangle of each segment.

**Step 2:** Sample selection guided by FROM-GLC10. To facilitate the subsequent resampling process (see Section III-B3), we prefer each sample to contain a single type of natural scene element to avoid category confusion. Therefore, we establish the following metric to evaluate the category homogeneity of each candidate sample:

\[
S_i = 1 + \frac{1}{\log(C_{\text{nature}})} \sum_{c=1}^{C_{\text{nature}}} p_{i,c} \log(p_{i,c})
\]  

where, larger \( S_i \) means that the proportion of a type of natural scene element contained in this sample is much higher than other natural scene elements. \( p_{i,c} \) denotes the percentage of pixels of class \( c \) in the \( i \)th candidate sample. And \( C_{\text{nature}} \) represents the number of categories as shown in Table I. We score the candidate samples by (1), and drop those with scores less than \( T \). The value of \( T \) is empirically set as 0.2 to tradeoff the category richness and homogeneity of the obtained samples.

Through the abovementioned steps, we obtain a natural scene dataset \( S_{\text{nature}} \) with \( N_{\text{nature}} \) samples.

2) **Automatic Sampling Method for Man-Made Scene Elements:** Considering the high scale variance and rapid temporal change of man-made scene elements in RSIs, we adopt frequently updated and fine-grained Open Street Map (OSM)\(^3\) to guide the sampling of man-made geographical elements. The sampling process has the following two steps.

**Step 1:** Associating OSM categories with man-made scene categories. OSM contains many redundant categories and invalid categories. Redundant categories have similar semantic contents (e.g., “college” and “university”), which prevents obtaining a balanced category distribution during resampling (see Section III-B3). Invalid categories (e.g., “phone” and “advice”) may be unrelated to geographical elements, which may prevent providing proper guidance for collecting desired samples. To address these issues, we first construct a man-made scene classification system with \( C_{\text{man-made}} \) categories [1], [2], [38], [39], [40]. Then, we define a mapping rule shown in Table II to associate OSM categories to man-made scene categories.

\(^3\) [Online]. Available: https://www.openhistoricalmap.org
TABLE II
RULE TO ASSOCIATE OSM CATEGORIES TO MAN-MADE SCENE CATEGORIES

| Scene categories       | OSM categories                                                                 |
|------------------------|---------------------------------------------------------------------------------|
| Airport                | aerodrome, airfield, apron, security, aerohangar, waiting area, terminal, hangar, bridge, cross river road, pontoon, weighbridge, footbridge, |
| Bridge                 |                                                                                |
| Commercial area        | shopping center, retail, marketplace, wholesale, commercial, shopping mall, parking, disused parking, parking space, car park, parking, carpooling, |
| Parking                |                                                                                |
| Residential area       | residential, apartment, terrace, terrace house, townhouse, neighborhood, university, college, school, secondary school, education, |
| School                 |                                                                                |
| Sports center          | volleyball, sport, playground, netball, court, playing field, recreation ground, |

Step 2: Automatic sampling guided by OSM. Considering that the man-made scene elements may change rapidly over time, we first retrieve OSM data $M_i$ that matches the input image $I_i$ in space and time as much as possible for sampling. Then, we parse $n_i$ man-made scene elements from $I_i$ based on the crowd-sourced annotations in $M_i$ and the association rules defined in step 1. Afterward, we determine the sampling region of each element in $I_i$ by geographic coordinate transformation, and finally, collect $n_i$ samples by taking the minimum bounding rectangle of each region.

Through abovementioned step, we obtain a man-made scene dataset $S_{man-made}$ with the size of $N_{man-made} = \sum_{i=1}^{N_{img}} n_i$.

3) Class-Balanced Oriented Resampling Strategy: Many recent studies have investigated SSL in the context of class-imbalanced and consistently observed the undesired performance of existing SSL algorithm [41]. Thus, constructing a class-balanced dataset is very important for promoting the performance of TOV model for RSIU even if it is trained with a label-free dataset.

After the abovementioned two sampling methods, we construct an RSI dataset with 3 million samples of 31 categories and name it TOV-RS-imbalanced. However, this dataset is not class balanced since the distribution of geographical elements is naturally uneven in space. To alleviate this problem, we develop a class-balanced oriented resampling strategy, which contains the following two steps.

1) Finding the class with the least sample data from the natural scene feature dataset $S_{nature}$, and let $n_k$ represent the number of samples in this class. Then, we randomly choose $n_k$ samples from each class in $S_{nature}$, and construct a relatively class-balanced subset $S'_{nature}$ containing $n_k \times C_{nature}$ samples.

2) Similarly, a relatively class-balanced subset $S'_{man-made}$ containing $n'_k \times C'_{man-made}$ man-made scene samples ($n'_k = \frac{n_k}{C_{man-made}} \times C'_{man-made}$), is obtained from $S_{man-made}$. Finally, combing $S'_{nature}$ and $S'_{man-made}$, we obtain a relatively class-balanced dataset, TOV-RS-balanced, with 0.5 million samples of 31 categories. Details about the dataset TOV-RS-balanced are shown in Fig. 5.

To sum up, the proposed automated data sampling and resampling method can construct a dataset with three notable characteristics, which contribute to training TOV model for RSIU.

1) Diversity: The samples in TOV-RS-balanced are collected in the categories of natural geographic elements and man-made geographic elements, with spatial resolutions between 1 and 20 m, regions covering more than 100 countries and time phases from 2019–2021, so they are diverse in the category, spatial resolution, illumination, and background.

2) Relatively class-balanced: TOV-RS-balanced can maintain the balance of samples from natural scenes and
man-made scenes because we use two different methods to sample such two kinds of scenes. But it can only achieve relative class balance in subcategories due to the noisy annotations in geographic data products.

3) **Scalable**: TOV-RS-balanced can expand in category, quantity, and diversity since the proposed RSIs sampling and resampling are automated. Moreover, its data source can also extend to other data modalities of RSI.

C. Training TOV Model for RSIU Based on a Human-Like SSL Mechanism

Recent studies have demonstrated that TOV model trained by contrastive SSL with mass unlabeled nature images has impressive generalizability, which performs comparably well or even better than supervised learning methods across various CV tasks [12], [17], [25]. However, we experimentally find that directly using this pipeline to train TOV model for RSIU cannot obtain desired results. The main reasons could be as follows. 1) Compared with natural images, RSIs have lower spatial resolution and blurrier texture details, from which learning low-level visual knowledge (e.g., texture features and edge features) is difficult. 2) Most RSI samples contain a complex scene consisting of multiple ground objects, which makes self-supervised optimization more difficult.

Humans usually learn knowledge along a path from easy to difficult and from general to specialized. We refer to such a path as the human-like learning path. For example, a person who has never purposely learned any remote sensing knowledge can identify common geographic elements in the real world, such as buildings, roads and vehicles, but he may need long-term remote sensing knowledge learning to identify some complex geographic elements such as tundra, coniferous forest, and broadleaf forest. These phenomena inspire us that: 1) common knowledge exists between RSIs and natural images, and the general low-level visual representations that are difficult to learn from RSIs can be learned from natural images; 2) since the knowledge for RSIU is more complex and specialized than that of natural image understanding, it is better to build TOV model for RSIU along an easy to difficult and general to a specialized learning path, similar as what human do. With this motivation, we propose a human-like SSL mechanism to learn TOV model for RSIU. In the following, we first briefly introduce the contrastive SSL framework and then detail the proposed human-like SSL learning path.

1) **Self-Supervised Contrastive Learning**: Contrastive learning is a typical SSL method, and it outperforms other unsupervised learning methods in learning general feature representation [30], [42], [43]. Contrastive learning methods bring different augmented views (positive sample pairs) of the same image closer and separate views (negative sample pairs) of different images, to learn both invariant and distinguishable visual representation. Specifically, it consists of the following two steps.

**Step 1**: Given a training set \( X = \{x_1, x_2, \ldots, x_n\} \) containing \( n \) unlabeled samples, each sample \( x_i \) is augmented by \( T(\cdot) \) to create two views as a pair of positive samples \((x_i^1, x_i^2)\). In contrast, any two augmented views of different samples are treated as a negative sample pair \((x_i^1, x_i^2)\). Here, \( T(\cdot) \) is a stochastic set of augmentations including random crop, random flip, color distortion, and Gaussian blur.

**Step 2**: Training a model to distinguish the positive and negative samples by embedding them to a proper feature space \( f(\cdot) \) using the loss function defined as (2) shown at the bottom of this page, where \( g(\cdot) \) is a multilayer projection head with two fully connected layers, which is widely used in SSL to compress the extracted features for contrasting. \( \tau = 0.5 \) is a temperature scalar. By minimizing (2), positive samples are pulled closer while negative samples are pushed apart in the learned feature space, which enhances the invariance and distinguishability of the learned visual representation.

2) **Training TOV Model for RSIU Along a Human-Like SSL Mechanism**: The proposed human-like SSL mechanism consists of the following two stages.

**Stage 1**: Learning general visual knowledge from the natural image dataset. Since natural images have higher resolution and richer texture details than RSIs, we first perform self-supervised contrastive learning by using large-scale unlabeled natural images for learning general low-level visual features, such as textures and edges. Specifically, we first construct a web-crawled natural image dataset containing 1 million samples. Rather than directly using a manually constructed natural image dataset with a limited size, we aim to minimize human intervention in the sample acquisition process to enhance the scalability and reduce the cost of the proposed framework for constructing the TOV model. Then we train TOV model using the loss function defined by (2) in an SSL way. The learned model is denoted as \( f_1(X; W_B) \), where the learned general low-level visual knowledge is stored in weights \( W_B \).

**Stage 2**: Learning specialized visual knowledge from the constructed RSI dataset. A common idea to achieve this is to initialize the model using the parameters learned in Stage 1 and perform secondary learning by using the RSI dataset. However, this approach may suffer the problem of catastrophic forgetting [44], which occurs specifically when the network is trained sequentially on different datasets or tasks. To better connect these two learning stages, we designed a simple memory retention strategy, which fixes the weights of the shallow and middle layers of the network learned in Stage 1 to keep the memory of low-level visual representations and then use the RSIs dataset to optimize weights of other layers of the network.

\[
\min_{\{w|w \notin W_s\& w \in W_e\}} \mathcal{L} \tag{3}
\]

\[
\mathcal{L} = -E_X \left[ \log \frac{\exp \left( g \left( f(x_1^1)^T f(x_2^2) / \tau \right) \right)}{\exp \left( g \left( f(x_1^1)^T f(x_2^2) / \tau \right) \right) + \sum_{j=1}^{n-1} \exp \left( g \left( f(x_1^1)^T f(x_j^2) / \tau \right) \right)} \right] \tag{2}
\]
Similar to SimCLR, MoCov2 is also represents the weights fixed in Stage 2, \( W_b \) represents the weights needed to be optimized and \( W_b + W_b' = W_f \). The selection of \( W_b \) and \( W_b' \) is based on the studies about the properties of the learned visual representation of each layer in deep convolutional neural network [45], [46], that is, the layers close to the input layer tends to learn the general low-level feature representation, while the layers distant from input layer tends to learn specific high-level feature representation.

### IV. Experiments

#### A. Datasets for Pretraining TOV Model

To pretrain TOV model along a learning path from general knowledge to specialized knowledge, we construct two datasets TOV-NI and TOV-RS (Table III), respectively.

TOV-NI is a web-crawled natural image dataset of 1 million samples. It is used for learning general knowledge in TOV model. To cover a set of visual concepts as broadly as possible, we automatically search images from the internet using 10 000 text queries from Wordnet [47]. Then, we choose at most 100 images per query to approximately keep the class balance of the resulting dataset.

TOV-RS is an RSI dataset constructed by the proposed data sampling method, which is used for learning specialized knowledge in TOV model. It has two versions. One is TOV-RS-imbalanced containing 3 million class-imbalanced samples and the other is TOV-RS-balanced containing 0.5 million class-balanced samples.

#### B. Comparison Experiments and Baseline

In the experiments, we evaluate the performance and generalization capabilities of the pretrained TOV model by three kinds of downstream RSIU tasks on 12 publicly available benchmarks, including scene classification, object detection, and semantic segmentation. We also compared the proposed method with six representative related works, including two commonly used model initialization methods, two recently proposed SSL methods and two publicly available TOV models as follows.

1) **Model initialization methods:**
   - **Random initialization:** Directly train a Resnet-50 model with random initialization parameters for each downstream RSIU task.
   - **ImageNet pretraining:** Directly train a Resnet-50 model initialized by the parameters pretrained on ImageNet [48] for each downstream RSIU task.

2) **SSL pretraining methods:**
   - **SimCLR** [30]: In this SSL method, different transformed instances obtained by applying artificial augmentations to the same sample are considered positive instances, and that of different samples in a training batch are regarded as negative instances. The artificial augmentations include random cropping, color perturbation, etc. The model learns visual knowledge by enhancing the similarity of positive instances and the difference between negative instances.
   - **MoCov2** [28]: Similar to SimCLR, MoCov2 is also based on contrastive learning, but it focuses on obtaining negative instances with a size far beyond the batch size to learn more discriminative visual knowledge. Therefore, a dynamic queue is proposed to save the features of negative instances, and a momentum update encoder is proposed to avoid the consistency problem of the representations of negative instances from the rapid change of the encoder.

3) **Publicly available TOV models:**
   - **SeCo** [19]: A TOV model that uses 1 million RSIs for representation learning in a self-supervised manner. First, SeCo constructed the pretraining dataset by collecting paired Sentinel-2 images (with only RGB bands) of different seasons at about 200 K locations distributed over global urban areas. Unlike SimCLR and MoCo, which use only artificial augmentations to obtain positive instances, SeCo regards sample pairs of different seasons as positive instances obtained by natural temporal augmentations for learning time-invariant features. Second, instead of projecting all possible and negative instances into a common feature space for contrastive learning as in SimCLR and MoCo, SeCo projects artificial-augmented positive instances, temporal-augmented positive instances, and artificial-temporal-augmented positive instances into three different feature subspaces for contrastive learning. By this way pretrained model can learn both time-varying and invariant features.
   - **SSL4EO** [21]: Similar to SeCo, SSL4EO constructs a pretraining dataset of 3 million in size and includes additional SAR images from Sentinel-1 in addition to the optical RSIs from Sentinel-2. For a fair comparison, we select the pretrained TOV model provided by SSL4EO using MoCov2 on 1 million Sentinel-2 RGB images as the comparison.
TABLE IV
DATASETS USED FOR SCENE CLASSIFICATION EXPERIMENTS

| Dataset                  | High-resolution RSIs datasets | Multispectral RSIs datasets |
|--------------------------|--------------------------------|-----------------------------|
|                          | AID  | NR  | RSD46 | PatternNet | UC Merced | EuroSAT | NaSC-TG2 |
| Number of categories     | 30   | 45  | 46    | 38         | 21        | 10      | 10       |
| Number of samples        | 10,000 | 31,500 | 117,000 | 30,400    | 2,100     | 27,000  | 20,000   |
| Spatial resolution (m)   | 0.5~8 | 0.2~30 | 0.5~2  | 0.062~4.693 | 0.3       | 10~60   | 100      |
| Image sizes              | 600×600 | 256×256 | 256×256 | 256×256   | 256×256   | up to 64×64 | 128×128 |

TABLE V
SCENE CLASSIFICATION RESULTS OF THE SEVEN METHODS FOR SEVEN DATASETS. OA IS USED AS THE EVALUATION INDEX

| Method                  | AID  | RSD46 | PatternNet | UC Merced | EuroSAT | NaSC-TG2 | Mean   |
|-------------------------|------|-------|------------|-----------|---------|----------|--------|
| Five samples            |      |       |            |           |         |          |        |
| Random initialization   | 36.50 | 25.44 | 11.08      | 33.40     | 17.71   | 43.81    | 45.82  | 30.54 |
| ImageNet pre-training   | 71.60 | 58.13 | 33.34      | 72.06     | 66.57   | 65.61    | 84.05  | 64.48 |
| SimCLR                  | 69.95 | 55.59 | 27.91      | 62.24     | 52.19   | 70.31    | 82.20  | 60.06 |
| MoCoV2                  | 73.30 | 63.16 | 30.82      | 66.04     | 58.19   | 72.44    | 83.02  | 63.85 |
| SeCo                    | 50.00 | 38.35 | 20.39      | 41.76     | 34.10   | 62.07    | 55.95  | 43.23 |
| SSLABO                  | 66.65 | 53.37 | 25.37      | 56.60     | 51.43   | 78.17    | 81.67  | 59.04 |
| Our method              | 78.55 | 62.89 | 33.76      | 73.68     | 61.52   | 74.70    | 85.57  | 67.24 |
| 20 samples              |      |       |            |           |         |          |        |
| Random initialization   | 57.85 | 47.27 | 20.35      | 45.58     | 33.14   | 65.39    | 64.65  | 47.75 |
| ImageNet pre-training   | 81.65 | 71.92 | 47.36      | 77.50     | 75.24   | 79.70    | 92.67  | 75.15 |
| SimCLR                  | 78.65 | 70.41 | 37.83      | 67.02     | 64.76   | 83.15    | 89.10  | 70.13 |
| MoCoV2                  | 83.05 | 76.92 | 42.98      | 72.85     | 68.00   | 84.02    | 91.22  | 74.15 |
| SeCo                    | 61.35 | 59.89 | 30.21      | 55.46     | 46.29   | 78.78    | 65.30  | 56.24 |
| SSLABO                  | 76.75 | 68.46 | 37.96      | 64.64     | 65.90   | 89.26    | 90.47  | 70.49 |
| Our method              | 84.95 | 76.60 | 45.75      | 76.02     | 70.48   | 87.35    | 92.27  | 76.17 |
| 50 samples              |      |       |            |           |         |          |        |
| Random initialization   | 65.05 | 68.48 | 26.82      | 59.90     | 54.57   | 73.87    | 84.12  | 61.83 |
| ImageNet pre-training   | 86.10 | 78.98 | 52.72      | 84.16     | 85.90   | 84.61    | 95.47  | 81.13 |
| SimCLR                  | 81.70 | 78.38 | 44.12      | 76.71     | 77.90   | 85.93    | 92.15  | 76.70 |
| MoCoV2                  | 85.40 | 83.33 | 49.13      | 79.80     | 79.81   | 86.98    | 94.05  | 79.79 |
| SeCo                    | 68.55 | 63.87 | 38.34      | 63.87     | 72.19   | 82.15    | 77.55  | 66.65 |
| SSLABO                  | 81.10 | 77.00 | 45.40      | 74.59     | 79.71   | 91.74    | 94.60  | 77.73 |
| Our method              | 88.55 | 83.57 | 52.96      | 85.38     | 83.81   | 89.54    | 95.05  | 82.69 |

learning path while two compared SSL methods are trained only using the TOV-RS dataset. For all three methods, we use the Adam optimizer with a batch size of 1024. The learning rate was initially set as 0.75 and was reduced in a cosine manner within 800 epochs. All experiments were implemented in PyTorch environment under the CentOS 7.5 platform with 8 NVIDIA Tesla A100 (memory 32 GB).

D. Results and Analysis

1) Scene Classification. Task and dataset description: Seven datasets shown in Table IV were used to evaluate the generalization capabilities of TOV model on scene classification task: Aerial image dataset (AID) [40], NWPU-RESISC45 (NR) [1], RSD46 [49], PatternNet [6], UC Merced [50], EuroSAT [51], and NaSC-TG2 [52]. For all the datasets, only RGB channels were used in the experiment. The overall accuracy (OA) was used to assess the performance.

Fine-tuning settings: For the proposed method and two compared SSL methods, we add a simple fully connected layer at the end of TOV model as a scene classification adapter and then fine tune the model using 5, 20, and 50 labeled samples per category, respectively. During training, we used the Adam optimizer with a batch size of 32. The learning rate was initially set to 0.001 and was reduced in a cosine manner within 200 epochs.

Results and analysis: Table V shows the experiment results. The best results are marked in bold. From the results, we can get the following two findings.

First, our method consistently outperforms all compared methods in an average of seven test datasets no matter how many samples are used for fine tuning. For example, when using five samples per category for fine tuning, the proposed method achieved a 4.3% performance improvement on average compared to the second-best method. Moreover, for each task dataset, our base model achieves the best or second-best results in almost all cases, while the two TOV models (SeCo and SSL4EO) pretrained on Sentinel-2 RSIs perform well only on the EuroSAT and NaSC-TG2 datasets, which also use low- and medium-resolution RSIs. This result indicates the impressive generalization capabilities of our TOV model on the scene classification task.
Second, training a general model in a label-free and task-independent way is more effective and robust for RSIU tasks. The ImageNet pretrained model was obtained at a high cost, i.e., 1.28 million labeled samples for supervised learning, but its advantage over the two SSL methods in terms of average performance was not obvious. For example, when using five samples per category for fine tuning, MoCov2 achieves comparable results to the ImageNet pretraining method, with a slight decrease of less than 1% in average OA compared to the latter.

2) Object Detection. Task and dataset description: Two datasets DOTA [4] and Levir [53] were used to evaluate the generalization capabilities of TOV model on object detection task, and the mean average precision (mAP50) is used to assess the performance.

- **DOTA** consists of RGB images and grayscale images. The RGB images are from Google Earth and CycloMedia while the grayscale images are from the panchromatic band of GF-2 and JL-1 satellite images. This dataset contains 188,282 objects from 15 categories.

- **Levir** is collected from Google Earth and consists of over 22,000 images with a size of $800 \times 600$ and the resolution of 0.2 ~ 1.0 m/pixels. It has three categories: airplane, ship, and oil tank.

Fine-tuning settings: For the proposed method and two compared SSL methods, we joined use region proposal network and ROIHead in faster RCNN\(^6\) as object detection adapter and then fine tune the model using 0.5%, 1.0%, 5.0% labeled samples of the whole dataset. During the training, we use the SGD optimizer with a batch size of 4 for fine tuning. The learning rate was initially set at 0.0025 and was reduced in a cosine manner within 200 epochs.

Results and analysis: Table VI shows the experiment results. The best results are marked in bold. A similar result can also be seen that our method consistently outperforms all compared methods in an average of two test datasets no matter how many samples are used for fine tuning. Moreover, we can observe that the performance of MoCov2 is better than the ImageNet pretraining method when using 1.0% and 5.0% labeled data for fine tuning, which further suggests the advantage of training TOV model in an SSL mechanism for RSIU. Besides, since SSL4EO and MoCoV2 were trained on different RSI datasets using the same SSL method, the most possible reason for the significant advantage of the latter is that we constructed the pretrained dataset with higher resolution and presents richer details of diverse ground objects.

3) Semantic Segmentation. Task and dataset description: Three datasets DLRSD [3], DGLCC [54], and Potsdam dataset [55] were used to evaluate the generalization capabilities of the proposed TOV model on semantic segmentation task, and the mean intersection over union (MIoU) was used to assess the performance.

- **DGLCC** is collected from DeepGlobe satellite and contains 803 images with a size of $2448 \times 2448$ and a resolution of 0.5 m. The dataset is annotated in seven classes.

- **DLRSRD** is a densely labeled dataset that consists of 2100 RGB images with the size of $256 \times 256$ and the resolution of 0.3 m. The dataset is annotated in 17 classes.

- **Potsdam** contains 38 UAV images with a size of $6000 \times 6000$ and the spatial resolution of 0.05 m. The dataset is annotated in seven classes.

Fine-tuning settings: For the proposed method and two compared SSL methods, we add the decoder in FCN [32] at the end of the encoder part of TOV model as an adapter for semantic segmentation tasks, and then fine tune the model using the 0.5%, 1.0%, 5.0% labeled samples of the whole dataset. During the training, we use the Adam optimizer with a batch size of 32. The learning rate was initially set as 0.005 and was reduced in a cosine manner within 200 epochs.

Results and analysis: Table VII shows the experiment results. The best results are marked in bold. Although results in segmentation tasks are different from those in classification and detection tasks, our method is still very competitive, for example, the best results are obtained on the Potsdam dataset with 0.5% and 5% samples for fine tuning. Moreover, we can observe that the ImageNet pretraining method does not significantly improve the semantic segmentation accuracy compared with the random initialization, while the SSL learning method, such as SimCLR, can outperform ImageNet pretraining in most cases.

---

\(^6\) Faster RCNN was implemented by using MM Detection ([Online]. Available: https://github.com/open-mmlab/mmdetection).
The main reason could be that ImageNet pretrained model is learned in a task-dependent way (i.e., scene classification task). Since the task of scene classification is not directly related to the task of semantic segmentation, resulting the generalization ability of learned features is not as good as that learned in a task-independent way.

V. DISCUSSION

In this section, we conduct a series of comparative experiments to further analyze the effect of two key factors on the performance of building TOV model, including the influence of using different data sampling methods and the selection of learning paths during self-supervised optimization.

A. Data Sampling Methods

Rich and high-quality data are important to training TOV model for RSIU. Although this model can be trained in a label-free SSL way, it can hardly learn valuable remote sensing visual knowledge from an unlabeled dataset that contains large amounts of semantically meaningless content or has severely class-imbalanced distribution. To figure out how different sampling methods affect the performance of the constructed TOV models, we repeat the process of building TOV model by employing one of the following sampling methods each time.

1) Grid sampling method. Give the set $H = \{I_1, I_2, \ldots, I_{5000}\}$ of RSIs, we meshed each RSI $I_i$ in $H$ into $n_i$ nonoverlapping patches with a size of $600 \times 600$ pixels, and then randomly sample 600 patches. Finally, a dataset TOV-RS-gridsampling containing 3 million samples was obtained.

2) The proposed sampling method without resampling strategy. We use the data sampling method described in Section III-B1 and III-B2 for obtaining a class-imbalanced dataset, TOV-RS-imbalanced, containing 3 million samples.

3) The proposed sampling method with resampling strategy. We use the data sampling method described in Section III-B for obtaining a relative class-balanced dataset, TOV-RS-balanced, containing 0.5 million samples.

We evaluated the generalization capabilities of TOV model trained by different datasets on the scene classification task. As shown in Table VIII, though the size of dataset TOV-RS-gridsampling and TOV-RS-imbalanced is much larger than TOV-RS-balanced, TOV model learned from the TOV-RS-balanced dataset significantly outperforms those learned from other two datasets for all seven datasets, with an average OA improvement of 17.1% and 13.3%. This result suggests that it is crucial to choose an appropriate sampling method to obtain a high-quality dataset for training TOV model for RSIU. This result may be from two reasons as follows.

First, the grid sampling approach may sample large amounts of semantically meaningless data, which confuses the feature representation learning of TOV model.

Second, the problem of data imbalance existing in the first and second data sampling methods poses challenges in training TOV using the contrastive SSL method. The idea of contrastive SSL learning is to encourage a model to learn invariance features by distinguishing between positive and negative samples. Since there is no annotation information in SSL, false negative samples (i.e., samples belonging to the same class) are more likely to be sampled in class-imbalanced datasets. As a result, the more imbalanced the classes are, the more false negative samples are sampled. This phenomenon potentially lets TOV model push features that belong to the same class farther away, and thus hurts the model’s performance.
B. Selection of Learning Paths

During self-supervised optimization, we design a human-like learning path, which first learns general knowledge from web-scale natural images and then learns domain-relevant specialized knowledge from unlabeled RSIs. To figure out how learning path selection affects the performance of TOV model for RSIU, we designed comparison experiments, as shown in Table IX, where $\langle D_{NI} \rangle$ and $\langle D_{RS} \rangle$ means only using natural image dataset TOV-NI and the RSI dataset TOV-RS-balanced, respectively, for learning TOV model, $\langle D_{NI}, D_{RS} \rangle$ represents using both TOV-NI and TOV-RS-balanced along a general to a specialized learning path for learning TOV model. From the results shown in Table IX, we found two phenomena.

First, training TOV model along a learning path from general knowledge to specialized knowledge can improve model performance greatly, with an average OA improvement of 6.63% and 11.95% compared with that only using natural image dataset or RSI dataset, respectively. The main reason could be that different types of datasets can provide complementary knowledge. Besides, the model trained only using RSI dataset is even worse than the model trained only using the nature image dataset, which confirms that remote sensing data are not sufficient to support TOV model to learn generic visual representations.

Second, designing a memory maintenance strategy for continual learning on different kinds of datasets is important, because TOV model has the problem of catastrophic forgetting which occurs frequently when the network is trained sequentially on different datasets. If training TOV model without using the memory maintenance strategy, it may totally forget what it has learned during training and get similar results as only using one dataset. (See row 2 and row 3 of Table IX).

VI. CONCLUSION

In this study, we give the definition of TOV model for RSIU and investigate a new paradigm for training TOV. Moreover, we perform a comprehensive comparative study by analyzing two key factors on the performance of building TOV model for RSIU, including the influence of using different data sampling methods and the selection of learning paths during self-supervised optimization. By combining our findings, our TOV model has shown impressive generalization capabilities across various RSIU tasks and outperforms the dominant ImageNet supervised pretrained method as well as two recently proposed SSL pretrained methods on the majority of 12 publicly available benchmarks. Our future work aims at building TOV 2.0 model for RSIU considering both task granularity and data modality. We expect TOV 2.0 to be broadly adaptable to multiple RSIU tasks and data modalities, such as hyperspectral image, SAR, and even video data, which can potentially pave the way for building general intelligence in the remote sensing field [58], [59].

ACKNOWLEDGMENT

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

REFERENCES

[1] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[2] H. Li et al., “RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data,” Sensors, vol. 20, no. 66, 2020, Art. no. 1594.
[3] Z. Shao, K. Yang, and W. Zhou, “Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling database,” Remote Sens., vol. 10, no. 6, 2018, Art. no. 964.
[4] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE Comput. Vis. Pattern Recognit., 2018, pp. 3974–3983.
[5] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, “A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data,” Remote Sens. Environ., vol. 270, no. 1, 2022, Art. no. 112830.
[6] W. Zhou, S. Newsam, C. Li, and Z. Shao, “PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” ISPRS J. Photogrammetry Remote Sens., vol. 145, pp. 197–209, 2018.
C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, “Spatial information inference net: Road extraction using road-specific contextual information,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 155–166, 2019.

Z. Zheng et al., “HyNet: Hyper-scale object detection network framework for multiple spatial resolution mapping imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 1–14, 2020.

C. Sun, A. Shirivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.

K. He, R. Girshick, and P. Dollar, “Rethinking imagenet pre-training,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4917–4926.

Y. Long et al., “On creating benchmark dataset for aerial image interpretation: Reviews, guidelines, and million-aid,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

J. Shao et al., “INTERN: A new learning paradigm towards general vision,” 2021, arXiv:2111.08687.

T. Chen et al., “M2W3:01988,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

T. Brown et al., “Language models are few-shot learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: a survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

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

L. Yuan et al., “Florence: A new foundation model for computer vision,” 2021, arXiv:2111.07232.

C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, “Remote sensing image scene classification with self-supervised paradigm under limited labeled samples,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 0084005.

O. Manas, A. Lacoste, X. Giro-i Nieto, D. Vázquez, and P. Rodríguez, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9394–9403.

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

Y. Wang, N. A. A. A. Braham, C. M. Albrecht, Z. Xiong, C. Liu, and X. X. Zhu, “SSL4EO-S12: A large-scale multimodal multitemporal dataset for self-supervised learning in Earth observation,” 2022, arXiv:2211.07044.

J. Li, P. Zhou, C. Xiong, and S. Hoi, “Prototypical contrastive learning of unsupervised representations,” in *Proc. Int. Conf. Learn. Representations*, 2021.

M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, “Partially view-aligned representation learning with noise-robust contrastive loss,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1134–1143.

Z. Zhan, J. Wang, X. Mei, C. Tao, and H. Li, “FALSE: False negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6518505.

P. Goyal et al., “Self-supervised pretraining of visual features in the wild,” 2021, arXiv:2103.01988.

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, to be published, doi: 10.1109/TGRS.2022.3176603.

X. Sun et al., “RingMo: A remote sensing foundation model with masked image modeling,” *IEEE Trans. Geosci. Remote Sens.*, 2022, to be published, doi: 10.1109/TGRS.2022.3194732.

X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, arXiv:2003.04297.

G.-J. Qi and J. Luo, “Small data challenges in Big Data era: A survey of recent progress on unsupervised and semi-supervised methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2168–2187, Apr. 2022.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Int. Mach. Learn.*, 2020, vol. 119, pp. 1597–1607.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, “When does contrastive visual representation learning work?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 01–10.
Chao Tao (Member, IEEE) received the B.S. degree in mathematics from the School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent systems from the Institution of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, in 2012.

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

Dr. Tao has frequently served as a Reviewer for several international journals including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, Photogrammetric Engineering and Remote Sensing, and ISPRS Journal of Photogrammetry and Remote Sensing. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China.

Ji Qi received the B.E. degree in remote sensing science and technology from the School of Geosciences and Info-Physics, Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing. His research interests include remote sensing and computer vision, machine learning, deep learning, and their applications in remote sensing.

Guo Zhang (Member, IEEE) received the B.E. and Ph.D. degrees in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2000 and 2005, respectively. His doctoral dissertation concerned the rectification for high-resolution remote sensing image under lack of ground control points.

Since 2005, he has been with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, where he became a Professor in 2011. His research interests include space photogrammetry (including the processing and applications for optical load, SAR load, video load, and laser load).

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

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

Weipeng Lu received the B.E. degree in surveying and mapping engineering from the China University of Petroleum (East China), Qingdao, China, in 2019, and the M.E. degree in surveying and mapping science and technology from Central South University, Changsha, China, in 2022. He is currently working toward the Ph.D. degree in geomatics with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Kowloon, Hong Kong.

His research interests include remote sensing image interpretation, urban remote sensing, computer vision, machine learning, and deep learning.

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

He is currently a Professor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. In 2011, he was a Research Associate with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, and from 2013 to 2014, a Visiting Scholar with the University of Illinois at Urbana-Champaign, Urbana, IL, USA. He has authored more than 30 journal papers. His research interests include geospatial remote sensing big data, machine/deep learning, and artificial/intelligent intelligence.

Dr. Li is a Reviewer for many journals.