MTP: Advancing Remote Sensing Foundation Model via Multitask Pretraining

Di Wang, *Member, IEEE*, Jing Zhang, *Senior Member, IEEE*, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, *Student Member, IEEE*, Haonan Guo, *Student Member, IEEE*, Bo Du, *Senior Member, IEEE*, Dacheng Tao, *Fellow, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*—Foundation models have reshaped the landscape of remote sensing (RS) by enhancing various image interpretation tasks. Pretraining is an active research topic, encompassing supervised and self-supervised learning methods to initialize model weights effectively. However, transferring the pretrained models to downstream tasks may encounter task discrepancy due to their formulation of pretraining as image classification or object discrimination tasks. In this study, we explore the multitask pretraining (MTP) paradigm for RS foundation models to address this issue. Using a shared encoder and task-specific decoder architecture, we conduct multitask supervised pretraining on the segment anything model annotated remote sensing segmentation dataset, encompassing semantic segmentation, instance segmentation, and rotated object detection. MTP supports both convolutional neural networks and vision transformer foundation models with over 300 million parameters. The pretrained models are finetuned on various RS downstream tasks, such as scene classification, horizontal, and rotated object detection, semantic segmentation, and change detection. Extensive experiments across 14 datasets demonstrate the superiority of our models over existing ones of similar size and their competitive performance compared to larger state-of-the-art models, thus validating the effectiveness of MTP.**

Manuscript received 20 March 2024; revised 8 May 2024; accepted 22 May 2024. Date of publication 4 June 2024; date of current version 1 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62225113, in part by the Innovative Research Group Project of Hubei Province under Grant 2024AFA017, and in part by the National Key Research and Development Program of China under Grant 2022YFB3903405 and Grant 2023YFC2705700. *(Corresponding authors: Jing Zhang; Minqiang Xu; Bo Du; Liangpei Zhang.)*

Di Wang and Bo Du are with the School of Computer Science, Institute of Artificial Intelligence, National Engineering Research Center for Multimedia Software, Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China (e-mail: wd74108520@gmail.com; dubo@whu.edu.cn).

Jing Zhang is with the School of Computer Science, Faculty of Engineering, The University of Sydney, Camperdown, NSW 2050, Australia (e-mail: jing.zhang1@sydney.edu.au).

Minqiang Xu, Lin Liu, Dongsheng Wang, and Erzhong Gao are with the National Engineering Research Center of Speech and Language Information Processing, iFlytek Company Ltd., Hefei 230088, China (e-mail: mqxu7@iflytek.com; linliu@iflytek.com; dswang7@iflytek.com; ezgao@iflytek.com).

Chengxi Han, Haonan Guo, and Liangpei Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: chengxihan@whu.edu.cn; haonan.guo@whu.edu.cn; zlp62@whu.edu.cn).

Dacheng Tao is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: dacheng.tao@gmail.com).

The codes and pretrained models will be released at https://github.com/ViTAE-Transformer/MTP.

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2024.3408154, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2024.3408154

***Index Terms*—Change detection, foundation model, multitask pretraining (MTP), object detection, remote sensing (RS), scene classification, semantic segmentation.**

I. Introduction

REMOTE sensing (RS) image is one of the most important data resources for recording ground surfaces and land objects. Precisely understanding RS images is beneficial to many applications, including urban planning [1], environmental survey [2], disaster assessment [3], etc.

Utilizing its inherent capability to automatically learn and extract deep features from objects, deep learning methods have found widespread application in the RS domain, particularly for tasks, such as scene classification, land use and land cover classification, and ship detection. Typically, ImageNet pretrained weights are employed in training deep networks for RS tasks due to their extensive representational ability. However, these weights are derived from pretraining models on natural images, leading to domain gaps between natural images and RS images. For instance, RS images are captured from a bird's-eye view, lack the vibrant colors of natural images, and possess lower spatial resolution. These disparities may impede the model's finetuning performance [4], [5]. Moreover, relying solely on limited task-specific data for training restricts the model size and generalization capability of current RS deep models due to the notorious overfitting issue.[1]

To tackle these challenges, the development of RS vision foundation models is imperative, which should excel in extracting representative RS features. However, the RS domain has long grappled with a scarcity of adequately large annotated datasets, impeding related investigations. Until recently, the most expansive RS scene labeling datasets were fMoW [6] and BigEarthNet [7], boasting 132 716 and 590 326 unique scene instances [8], respectively—yet still falling short of benchmarks set by natural image datasets, such as ImageNet-1K [9]. Long et al. [8] addressed this gap by introducing MillionAID, a large-scale RS scene labeling dataset with a closed sample capacity of 1 000 848 compared to ImageNet-1 K, igniting interest in supervised RS pretraining [5], [10]. These studies show the feasibility of pretraining RS foundation models on large-scale RS datasets. Nonetheless, supervised pretraining of RS foundation models

[1] https://github.com/ViTAE-Transformer/MTP

may not be the most preferable choice due to the expertise and substantial time and labor costs associated with labeling RS images.

Constructing large-scale RS annotation datasets is challenging due to the high complexity and cost of labeling. Despite this challenge, the advancement of Earth observation technologies grants easy access to a vast amount of unlabeled RS images. Efficiently leveraging these unlabeled RS images is crucial for developing robust RS foundation models. In the realm of deep learning, unsupervised pretraining has emerged as a promising approach for learning effective knowledge from massive unlabeled data [14], [15], [16], [17]. Typically, unsupervised pretraining employs self-supervised learning (SSL) to learn effective feature representation. SSL encompasses two primary techniques: contrastive-based [18], [19], [20] and generative-based learning [21], [22], [23]. Contrastive learning aims to bring similar samples closer while maximizing distances between dissimilar samples through the object discrimination pretext task. When applied to the RS domain, data characteristics such as geographic coordinates [24], [25], [26] and temporal information [27], [28], [29] are usually leveraged in formulating the pretext task. However, designing these pretext tasks and gathering requisite data can be inefficient, especially for training large-scale models. Generative-based learning, exemplified by masked image modeling (MIM), circumvents this challenge by enhancing network representation through reconstructing masked regions. Many RS studies leverage MIM initialization for its efficiency [30], [31], [32], [33], [34], [35], [36], [37]. Recent approaches have attempted to combine contrastive-based and generative-based learning techniques to pretrain more powerful models [38], [39], [40].

However, existing research usually resorts to a single data source. For instance, the authors in [5] and [30] utilized RGB aerial images from MillionAID, while the authors in [31] and [34] utilized Sentinel-2 multispectral images. Despite recent advancements in RS multimodal foundation models [41], [42], [43], [44], which are beginning to incorporate more diverse imagery, such as SAR, they still remain within the realm of in-domain data, namely pretraining with RS data. However, restricting pretraining solely to RS images may limit model capabilities since understanding RS objects requires specialized knowledge [12]. Can RS foundation models benefit from incorporating information from other data sources? [5] suggests that traditional ImageNet pretraining aids in learning universal representations, whereas pretraining on RS data is particularly beneficial for recognizing RS-related categories. To address this, Huang et al. [45] developed a teacher–student framework that integrates ImageNet supervised pretraining and RS unsupervised pretraining simultaneously, while Mendieta et al. [46] employed representations from ImageNet to enhance the learning process of MIM for improving RS foundation models. In addition, the authors in [12] and [11] sequentially pretrain models on natural images and RS images using contrastive SSL or MAE [23], respectively, as illustrated in Fig. 1(a).

While previous RS foundation models have shown remarkable performance across various RS tasks, a persistent challenge remains: the *task discrepancy* between pretraining and finetuning, which often dictates the effectiveness of migrating pretrained models to downstream tasks. Research has highlighted the impact of representation granularity mismatch between pretraining and finetuning tasks [5]. For instance, models pretrained on scene-level classification tasks perform favorably when finetuned on similar tasks but falter on pixel-level segmentation tasks. To address this issue, recent work [13] has explored the segmentation pretraining (SEP) paradigm, as shown in Fig. 1(b), yielding improved finetuning results. This suggests that enhancing model representation capability through additional pretraining, particularly on tasks demanding finer representation granularity, such as pixel-level segmentation, could be beneficial. Motivated by these findings, we ask: *can we significantly enhance RS foundation models' representation ability through additional pretraining incorporating multiple tasks with diverse representation granularity?* To this end, we investigate the multitask pretraining (MTP) paradigm to bridge the gap between upstream and downstream tasks and obtain more powerful RS foundation models, as shown in Fig. 1(c). Importantly, MTP is designed to be applied to any existing pretraining models, irrespective of whether trained on RS or natural images.

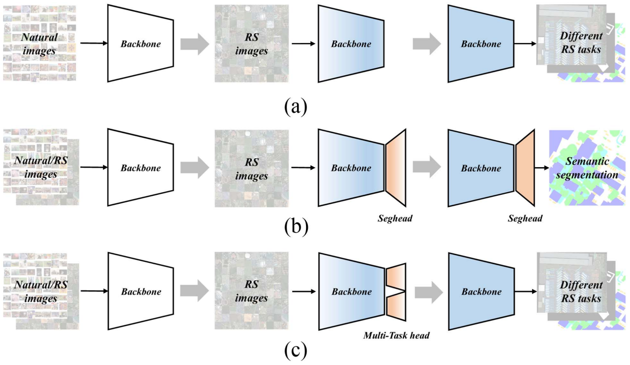


Fig. 1. Comparison of various pretraining methods. (a) The authors in [11] and [12] sequentially pretrains a foundational model on both natural and RS images. (b) Wang et al. [13] employs a two-stage pretraining strategy to initialize task-specific decoders (e.g., segmentation) using existing foundational models pretrained on either natural or RS images, preserving the decoder during subsequent finetuning. We extend (b) by incorporating multitask decoders to enhance the representation capacity of the foundational model, facilitating easy transferability across diverse tasks during finetuning, as depicted in (c).

Implementing MTP to bridge upstream-downstream task discrepancy necessitates the utilization of a similar or the same pretraining task as the downstream one, such as SEP for RS segmentation tasks [13]. Therefore, to cover the common task types in typical downstream applications, MTP tasks should encompass dense prediction tasks, such as object detection and semantic segmentation. Hence, MTP requires a pretraining dataset with labels for these tasks, ideally, each sample encompassing all task labels. However, existing RS datasets often lack annotations for segmentation and rotated object detection. Fortunately, recent work [13] introduces **S**egment **A**nything **M**odel annotated **R**emote Sensing **S**egmentation (SAMRS), a large-scale segmentation dataset derived from existing RS rotated object detection datasets via the segment anything model

(SAM) [47]. SAMRS provides both detection and segmentation labels, facilitating MTP across RS semantic segmentation, instance segmentation, and rotated object detection tasks. Utilizing SAMRS, we demonstrate MTP's efficacy in enhancing RS foundation models, including both convolutional neural networks (CNNs) and vision transformer foundation models with over 300 million parameters.

The main contributions of this article are three-fold as follows.

1) We address the discrepancy between upstream pretraining and downstream finetuning tasks by introducing a stagewise MTP approach to enhance the RS foundation model.
2) We utilize MTP to pretrain representative CNN and vision transformer foundation models with over 300 M parameters on the SAMRS dataset, encompassing semantic segmentation, instance segmentation, and rotated object detection tasks in a unified framework.
3) Extensive experiments demonstrate that MTP significantly advances the representation capability of RS foundation models, delivering remarkable performance across various RS downstream tasks such as scene classification, semantic segmentation, object detection, and change detection.

The rest of this article is organized as follows. Section II introduces the existing works related to supervised, multistage, and multitask RS pretraining. Section III presents the details of MTP, where the used SAMRS dataset and vision foundation models are also briefly introduced. Experimental results and corresponding analyses are depicted in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Supervised Pretraining for RS Foundation Model

Before the rise of SSL-based RS foundation models, researchers have already delved into pretraining deep models using labeled RS datasets. Tong et al. [48] pretrained an ImageNet-pretrained ResNet-50 [49] using images from the GID dataset [48] to derive pseudolabels for precise land-cover classification on high-resolution RS images. Recognizing the challenge of labeling large-scale RS images, others sought alternatives to RS annotation datasets. For instance, Li et al. [50] utilized the global land cover product Globeland30 [51] as supervision for RS representation learning. They adopted a mean-teacher framework to mitigate random noise stemming from inconsistencies in imaging time and resolution between RS images and geographical products. Moreover, they incorporated additional geographical supervisions, such as change degree and spatial aggregation, to regularize the pretraining process [52]. Long et al. [10] subsequently demonstrated the effectiveness of various CNN models (including AlexNet [53], VGG-16 [54], GoogleNet [55], ResNet-101 [49], and DenseNet-121/169 [56]) pretrained from scratch on the MillionAID dataset. Their models outperformed traditional ImageNet pretrained models in scene classification tasks, indicating the potential of leveraging large-scale RS datasets for pretraining. Later, Wang et al. [5] pretrained typical CNN models and vision transformer models, including Swin-T [57] and ViTAEv2 [58], all randomly initialized, on the MillionAID. They conducted a comprehensive empirical study comparing finetuning performance using different pretraining strategies (MillionAID vs. ImageNet) across four types of RS downstream tasks: scene recognition, semantic segmentation, rotated object detection, and change detection. Their results demonstrated the superiority of vision transformer models over CNNs on RS scenes and validated the feasibility of constructing RS foundation models via supervised pretraining on large-scale RS datasets. Bastani et al. [59] introduced the larger Satlas dataset for RS supervised pretraining. Very recently, SAMRS [13] introduced supervised semantic SEP to enhance model performance on the segmentation task. Inspired by Wang et al. [13], this article revisits the supervised learning approach by integrating it with existing pretraining strategies, such as ImageNet pretraining, and exploring MTP to construct distinct RS foundation models.

B. Multistage Pretraining for RS Foundation Model

Given the domain gap between RS images and natural images or between various RS modalities, it is reasonable to conduct multiple rounds of pretraining. Gururangan et al. [60] demonstrated that unsupervised pretraining on in-domain or task-specific data enhances model performance in natural language processing (NLP) tasks. Building on this insight, Zhang et al. [11] devised a sequential pretraining approach, initially on ImageNet followed by the target RS dataset, employing MIM for pretraining. Similarly, Tao et al. [12] proposed a strategy inspired by human-like learning, first performing contrastive SSL on natural images, then freezing shallow layer weights and conducting SSL on an RS dataset. Contrary to Gururangan et al.'s [60] work, Dery et al. [61] introduced stronger end-task-aware training for NLP tasks by integrating auxiliary data and end-task objectives into the learning process. Similarly, Wang et al. [13] introduced additional SEP using common segmenters (e.g., UperNet [62] and Mask2Former [63]) and the SAMRS dataset, enhancing model accuracy in RS segmentation tasks. Notably, our objective diverges from Wang et al.'s [13] work in applying stagewise pretraining. While Wang et al. [13] retains the segmentor after SEP to enhance segmentation performance, we aim to enhance the representation capability of RS foundation models via stagewise pretraining, preserving only the backbone network after pretraining to facilitate transfer to diverse RS downstream tasks.

C. MTP for RS Foundation Model

Applying multitask learning to enhance the RS foundation model is an intuitive idea. Li et al. [64] introduced multitask SSL representation learning, combining image inpainting, transform prediction, and contrast learning to boost semantic segmentation performance in RS images. However, it was limited to finetuning a pretrained model solely on semantic segmentation tasks, constrained by model size and pretraining dataset capacity. RSCoTr [65] constructs a multitask learning framework to simultaneously achieve classification, segmentation, and detection tasks. Unfortunately, the network can only be optimized by one task in each iteration during training due to lacking of multilabel datasets. The aspiration to consolidate multiple tasks into a single model has been a longstanding pursuit [15], [17], [42], [58], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], aligning with the original goals of the foundation model exploration. Bastani et al. [59] devised a multitask model by integrating Swin-Base [57] with seven heads from existing networks (e.g., Faster-RCNN [76] and UNet [77]), facilitating training on the multitask annotated Satlas dataset. However, their approach lacked incorporation of typical RS rotated object tasks, focusing solely on transferring the model to RS classification datasets. Inspired by these pioneering efforts, this article constructs a unified MTP framework that supports multiple datasets, where the sample in each dataset simultaneously possesses multitask labels, which are uniformly processed in a data loader pipeline. During the training, the model can be simultaneously optimized through multiple tasks when in each iteration. Based on this framework, we pretrain RS foundation models with over 300 M parameters, encompassing semantic segmentation, instance segmentation, and rotated object detection tasks using the SAMRS dataset. After pretraining, the backbone network is further finetuned on various RS downstream tasks.

III. Multitask Pretraining

We utilize semantic segmentation, instance segmentation, and rotated object detection annotations from the SAMRS dataset for MTP. Advanced CNN and vision transformer models serve as the backbone networks to thoroughly investigate MTP. This section begins with an overview of the SAMRS dataset, followed by a brief introduction to the selected models. Subsequently, we present the MTP framework and implementation details.

A. SAMRS Dataset

SAMRS dataset [13] is a large-scale RS segmentation database, comprising 105 090 images and 1 668 241 instances from three datasets: SOTA, SIOR, and FAST. These datasets are derived from existing large-scale RS object detection datasets, namely DOTA-V2.0 [78], DIOR [79], and FAIR1M-2.0 [80], through transforming the bounding box annotations using the SAM [47]. SAMRS inherits the categories directly from the original detection datasets, resulting in a capacity exceeding that of most existing RS segmentation datasets by more than tenfold (e.g., ISPRS Potsdam[2] and LoveDA [81]). The image sizes for the three sets are 1024 × 1024, 800 × 800, and 600 × 600, respectively. Despite being primarily intended for large-scale pretraining exploration rather than benchmarking due to its automatically generated labels, SAMRS naturally supports instance segmentation and object detection. This versatility extends its utility to investigating large-scale MTP.

B. Backbone Network

In this research, we adopt **R**otated **V**aried-**S**ize window **A**ttention (RVSA) [30] and InternImage [82] as the representative vision transformer-based and CNN-based foundation models.

1) RVSA: This model is specially designed for RS images. Considering the various orientations of RS objects caused by the bird's-eye view, this model extends the varied-size window attention in Zhang et al.'s [83] work by additionally introducing a learnable angle factor, offering windows that can adaptively zoom, translate, and rotate.

Specifically, given an input feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ (C, H, W are the number of channel, height, and width in $\mathbf{X}$), which is evenly divided into different windows, where the feature of each window can be formulated as $\mathbf{X}_w \in \mathbb{R}^{C \times s \times s}$ (s is the window size), obtaining $\frac{H}{s} \times \frac{W}{s}$ windows totally. Then, three linear layers are used to generate the query feature, and initial key and value features, which are separately represented as $\mathbf{Q}_w$, $\mathbf{K}_w$, and $\mathbf{V}_w$. We use $\mathbf{X}_w$ to predict the variations of the window

$$S_w, O_w, \Theta_w = \text{Linear}(\text{LeakyReLU}(\text{GAP}(\mathbf{X}_w))) \quad (1)$$

where GAP is global average pooling, $S_w = \{s_x, s_y \in \mathbb{R}^1\}$ and $O_w = \{o_x, o_y \in \mathbb{R}^1\}$ are the scale factor and offset in the X- and Y-axis, while $\Theta_w = \{\theta \in \mathbb{R}^1\}$ is the rotation angle. Taking an example using the corner points of a window

$$\begin{bmatrix} x_{l/r} \\ y_{l/r} \end{bmatrix} = \begin{bmatrix} x^c \\ y^c \end{bmatrix} + \begin{bmatrix} x^r_{l/r} \\ y^r_{l/r} \end{bmatrix} \quad (2)$$

where x_l, y_l, x_r, y_r are the coordinates of the upper left and lower right corners of the initial window, x_c, y_c are the coordinates of the window center point. Therefore, $x^r_l, y^r_l, x^r_r, y^r_r$ are the distances between the corner points and the center in horizontal and vertical directions. The transformation of the window can be implemented using the obtained scaling, translation, and rotation factors

$$\begin{bmatrix} x'_{l/r} \\ y'_{l/r} \end{bmatrix} = \begin{bmatrix} x^c \\ y^c \end{bmatrix} + \begin{bmatrix} o_x \\ o_y \end{bmatrix} + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x^r_{l/r} \cdot s_x \\ y^r_{l/r} \cdot s_y \end{bmatrix} \quad (3)$$

x'_l, y'_l, x'_r, y'_r are the coordinates of the corner points of the transformed window. Then, new key and value feature $\mathbf{K}_{w'}$ and $\mathbf{V}_{w'}$ can be sampled from the obtained window, and the self-attention (SA) can be operated by the following formula:

$$\mathbf{F}_w = SA(\mathbf{Q}_w, \mathbf{K}_{w'}, \mathbf{V}_{w'}) = \text{softmax}\left(\frac{\mathbf{Q}_w \mathbf{K}^T_{w'}}{\sqrt{C'}}\right) \mathbf{V}_{w'}. \quad (4)$$

In this formula, $\mathbf{F}_w \in \mathbb{R}^{s^2 \times C'}$ is the feature output by one SA of one window, $C = hC'$, where h is the number of SA. The shape of final output features in RVSA can be recovered by concatenating the features from different SAs in the channel dimension and merging features from different windows along the spatial dimension.

[2] [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-labelpotsdam.aspx

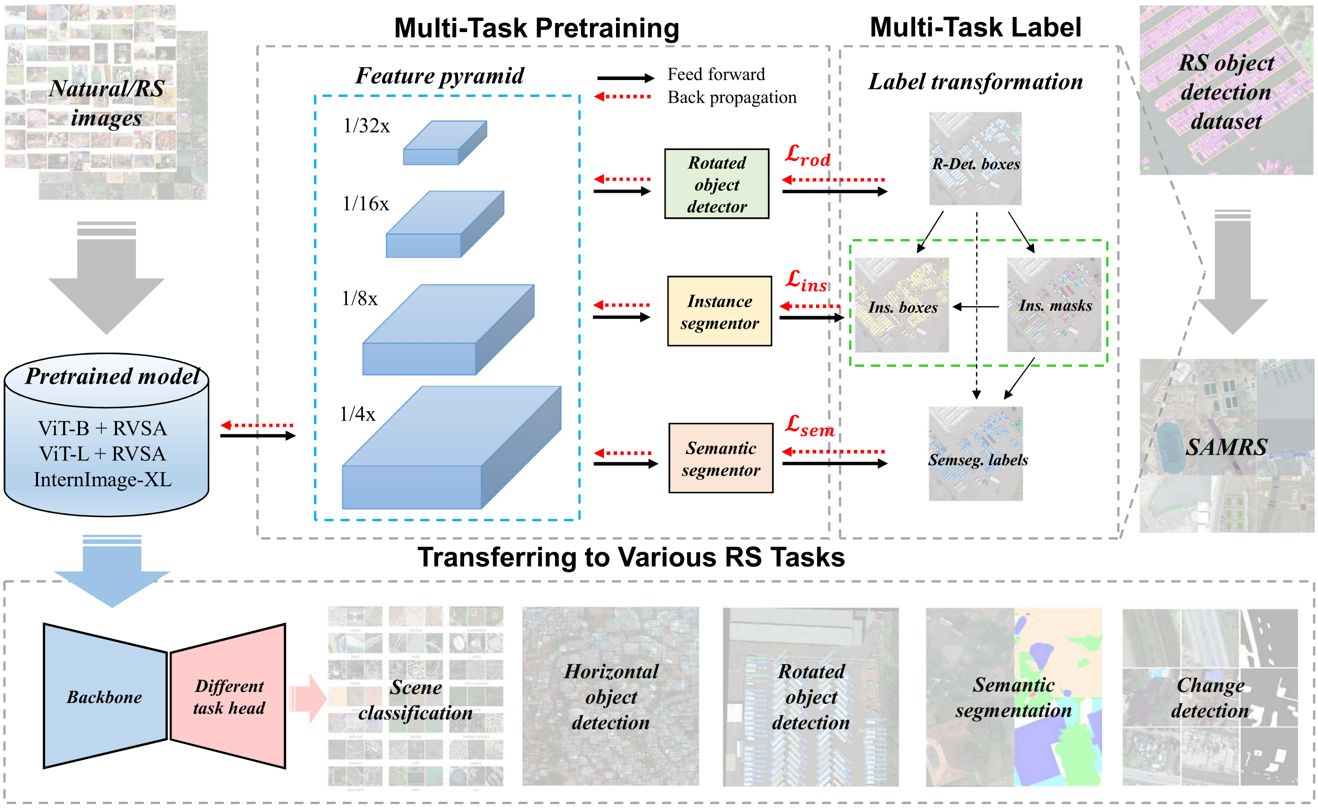


Fig. 2. Overall pipeline of MTP. Inside MTP, the feature pyramid from the backbone network is fed into multiple decoders for various tasks, including rotated object detection, instance segmentation, and semantic segmentation. These tasks are supervised by diverse labels in the SAMRS dataset. Following MTP, the pretrained model is transferred to different RS tasks for finetuning.

TABLE I
DETAILED CONFIGURATIONS OF DIFFERENT RVSA MODELS

Backbone	ViT-B + RVSA	ViT-L + RVSA
Depth	12	24
Embedding dim	768	1024
Head	12	16
Full attention index	[3, 6, 9, 12]	[6, 12, 18, 24]
Feature pyramid index	[4, 6, 8, 12]	[8, 12, 16, 24]

RVSA is used to replace the original multihead full attention in original vision transformers. To achieve a tradeoff between accuracy and efficiency, following Li et al.'s [84] work, only the full attention in 1/4 depth layer is preserved. In the original paper [30], RVSA is separately used on ViT [85] and ViTAE [72], whereas ViTAE is a CNN-Transformer hybrid model. In this article, we employ the ViT-based version to investigate the impact of MTP on a plain vision transformer. In addition, the RVSA model in the original paper is limited to the base version of vision transformers, i.e., ViT-B + RVSA. To pretrain larger models, we further apply RVSA to ViT-Large, obtaining ViT-L + RVSA. Their detailed configurations are presented in Table I.

2) InternImage: This model integrates the strengths of recent vision transformers and large kernels into CNNs via dynamic sparse kernels, combining long-range context capture, adaptive spatial information aggregation, and efficient computation. It extends deformable convolution [86], [87] with depthwise and multihead mechanisms and incorporates modern transformer designs, such as layer normalization [88], feedforward networks [89], and GELU activation [90]. We evaluate its performance on diverse RS downstream tasks, showcasing its potential beyond its initial design for natural images. Furthermore, this choice facilitates investigating the impact of MTP on CNN-based models. Here, we employ the XL version to match the model size of ViT-L + RVSA.

C. *Multitask Pretraining*

We examine MTP using three models: ViT-B + RVSA, ViT-L + RVSA, and InternImage-XL. As the original RVSA research [30] focuses solely on the base version, we independently pretrain ViT-L on MillionAID similar to ViT-B + RVSA. These pretrained weights will be publicly accessible. Fig. 2 shows the overall pipeline of MTP. Technically, we further train the pretrained model on the SAMRS dataset, encompassing various annotations, such as semantic segmentation, instance segmentation, and rotated object detection tasks concurrently. We employ well-established classical networks, including UperNet [62], Mask-RCNN [91], and Oriented-RCNN [92], as segmentors or detectors. These networks utilize feature pyramids and are supervised with different labels. To illustrate this process, we depict the label transformation when generating SAMRS. Initially, rotated detection boxes (R-Det. boxes) are transformed into binary masks using SAM, serving as instance-level mask annotations. Subsequently, the minimum circumscribed horizontal rectangle of the binary mask is derived as instance-level box annotations, with categories inherited from rotated boxes. These

TABLE II
Training Costs of Implementing MTP Using Different Models

Backbone	#Param.(M)	#GPU	Time (days)
ViT-B + RVSA	86	16	3.0
ViT-L + RVSA	305	32	6.3
InternImage-XL	335	32	6.3

TABLE III
OA(%) of Different Model Pretraining Strategies on EuroSAT

Model	MAE	MTP	OA
ViT-B + RVSA	✔		98.54
ViT-B + RVSA		✔	98.24
ViT-B + RVSA	✔	✔	98.76

instance-level annotations are utilized for instance segmentation. Semantic segmentation labels are then obtained by assigning rotated box categories to the masks. The losses stemming from these labels are $\mathcal{L}_{\text{rod}}$, $\mathcal{L}_{\text{ins}}$, and $\mathcal{L}_{\text{sem}}$, employed for the respective tasks. Notably, the instance segmentation loss comprises two components: the box annotation loss $\mathcal{L}_{\text{ins\_}b}$ and the binary mask loss $\mathcal{L}_{\text{ins\_}m}$. The overall loss for MTP is

$$\mathcal{L} = \mathcal{L}_{\text{rod}} + \mathcal{L}_{\text{ins\_}b} + \mathcal{L}_{\text{ins\_}m} + \mathcal{L}_{\text{sem}}. \tag{5}$$

Since SAMRS contains three sets, we have

$$\mathcal{L} = \sum_{i=1}^{3} \mathcal{L}^{i}_{\text{rod}} + \mathcal{L}^{i}_{\text{ins\_}b} + \mathcal{L}^{i}_{\text{ins\_}m} + \mathcal{L}^{i}_{\text{sem}} \tag{6}$$

where i indexes the three subsets: SOTA, SIOR, and FAST. The other settings of the loss follow the original papers [62], [91], [92]. In practice, we implement the overall framework based on MMSegmentation,[3] MMDetection,[4] and MMRotate.[5] However, all these packages only support a single task. So we integrate the key components from these packages, such as the dataloader, model structure, loss function, and metric calculator, into a unified pipeline, to realize the MTP.

D. Implementation Details

The pretraining is conducted on NVIDIA V100 GPUs. All models are trained for 80 K iterations using the AdamW optimizer [93]. The base learning rates of RVSA and InternImage are 0.00006 and 0.00002, respectively, with a weight decay of 0.05. We adopt an iterationwise cosine annealing scheduler to adjust the learning rate. The layer decay rates of RVSA models and InternImage are 0.9 and 0.94, following original papers [30], [82]. For ViT-B + RVSA, the batch size and input image size are set to 48 and 224, which are doubled for training larger models. Table II lists the training costs of implementing MTP using different models.

IV. Finetuning Experiments

In this section, we thoroughly evaluate MTP's performance by finetuning pretrained models across four classical RS tasks: scene classification, object detection, semantic segmentation, and change detection, where the most representative and widely used benchmarks in the literature for different downstream tasks are employed for comparison. We also investigate the characteristics of MTP-based RS foundation models, examining the relationships between adopted datasets, hyperparameters, and finetuning performances, measuring accuracy variations with reduced training samples, and visualizing the predicted results.

A. Scene Classification

We first evaluate the pretrained models on the scene classification task. It does not need any extra decoder and can reflect the overall representation capability of the pretrained model.

1) Dataset: We adopt two classical datasets: EuroSAT [94] and RESISC-45 [95] for scene classification.

1) EuroSAT: This dataset is captured by Sentinel-2 from Europe for land use and land cover classification. It has ten classes, a total of 27 000 images with a resolution of 64 × 64. We adopt the public train/val split [96] by following [29] and[31].
2) RESISC-45: This is a commonly-used dataset. It contains 31 500 images in a size of 256 × 256 across 45 categories, where each category possesses 700 samples. Following [5], [30], [32], [42], and [45], we randomly select 20% of the data for training and 80% of the data for testing.

2) Implementation Details: In the implementation, all models are trained with a batch size of 64. The training epochs for EuroSAT and RESISC-45 are set to 100 and 200, respectively. The AdamW optimizer is used, where the base learning rate for RVSA and InterImage are 0.00006 and 0.00002, respectively, with a weight decay of 0.05. In the first five epochs, we adopt a linear warming-up strategy, where the initial learning rate is set to 0.000001. Then, the learning rate is controlled by the cosine annealing scheduler. The layer decay rates are 0.9 and 0.94 for RVSA and InternImage models, respectively. For classification, a global pooling layer and a linear head are used after the backbone network. To avoid overfitting, we adopt multiple data augmentations, including random resized cropping, random flipping, RandAugment [97], and random erasing. Since the original image size of EuroSAT is too small, before feeding into the network, we resize the image to 224 × 224. The overall accuracy (OA) is used as the evaluation criterion. All experiments are implemented by MMPretrain.[6]

3) Ablation Study of Stagewise Pretraining: As aforementioned, MTP is implemented based on existing pretraining models since it tries to address the task-level discrepancy. So an interesting question naturally arises: what about conducting MTP from scratch? To this end, we experiment by exploring different pretraining strategies using ViT-B + RVSA on EuroSAT, and the results are shown in Table III. It can be seen that, without using pretrained weights, MTP cannot achieve the ideal performance and even performs worse than

[3] [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[4] [Online]. Available: https://github.com/open-mmlab/mmdetection

[5] [Online]. Available: https://github.com/open-mmlab/mmrotate

[6] [Online]. Available: https://github.com/open-mmlab/mmpretrain

TABLE IV
OA (%) OF FINETUNING DIFFERENT PRETRAINED MODELS ON EUROSAT AND RESISC-45 DATASETS

Method	Model	EuroSAT	RESISC-45
GASSL [27]	ResNet-18	89.51	-
SeCo [29]	ResNet-18	93.14	-
SatMAE [31]	ViT-L	95.74	94.10
SwiMDiff [98]	ResNet-18	96.10	-
GASSL [27]	ResNet-50	96.38	93.06
GeRSP [45]	ResNet-50	-	92.74
SeCo [29]	ResNet-50	97.34	92.91
CACo [28]	ResNet-50	97.77	91.94
TOV [12]	ResNet-50	-	93.79
RSP [5]	ViTAEv2-S	-	95.60
RingMo [32]	Swin-B	-	95.67
SatLas [59]	Swin-B	-	94.70
CMID [38]	Swin-B	-	95.53
GFM [46]	Swin-B	-	94.64
CSPT [11]	ViT-L	-	95.62
Usat [99]	ViT-L	98.37	
Scale-MAE [36]	ViT-L	98.59	95.04
CtxMIM [37]	Swin-B	98.69	-
SatMAE++ [100]	ViT-L	99.04	-
SpectralGPT$^+$ [34]	ViT-B	99.21*	-
SkySense [42]	Swin-L	-	95.92*
SkySense [42]	Swin-H	-	96.32*
MAE	ViT-B + RVSA	98.54	95.49
MAE + MTP	ViT-B + RVSA	**98.76**	**95.57**
MAE	ViT-L + RVSA	98.56	95.46
MAE + MTP	ViT-L + RVSA	**98.78**	**95.88**
IMP	InternImage-XL	**99.30***	95.82
IMP + MTP	InternImage-XL	99.24*	**96.27***

Here IMP means pretraining on imagenet-22k. Bold indicates the better accuracy of pretrained models with or without MTP. * represents the first three methods among all comparison methods, where the first, second, and third places are emphasized by black, red and green colors, respectively.

MAE pretraining. These results demonstrate the importance of performing stagewise pretraining.

4) Finetuning Results and Analyses: Table IV shows the finetuning results. It can be seen that MTP can improve existing foundation models on scene classification tasks, especially for the RVSA series. It helps the model achieve state-of-the-art performances compared to other pretraining models that have comparable sizes. With the help of MTP, on the RESISC-45 dataset, InternImage-XL surpasses Swin-L-based SkySense [42], which is pretrained on a tremendously large dataset that has more than 20 million multimodal RS image triplets involving RGB high-resolution images and multitemporal multispectral and SAR sequences. MTP boosts the performance of InternImage-XL close to the Swin-H-based SkySense (96.27 versus 96.32), which has more parameters. We also notice the accuracy of IMP-InternImage-XL is decreased marginally in EuroSAT after MTP. We will investigate this phenomenon later. Nevertheless, the obtained model still outperforms SpectralGPT$^+$, which is pretrained with 1 million multispectral images, where each sample can be regarded as containing multiple groups of trispectral images, similar to RGB channels.

B. Horizontal Object Detection

After completing the scene-level task of recognition, we focus on the object-level tasks, i.e., horizontal and rotated object detection. Here, we first consider the horizontal detection task.

1) Dataset: We use two public datasets Xview [101] and DIOR [79] for evaluation. The details are as follows.

1) Xview: This dataset is from the DIUx xView 2018 Detection Challenge [101]. It collects Worldview-3 satellite imagery beyond 1400 km^2 in a ground resolution of 0.3m, involving 60 classes over 1 million object instances. Due to only the 846 images (beyond 2000 × 2000 pixels) in the training set are available, following [27] and [37], we randomly select 700 images as the training set and 146 images for testing.
2) DIOR: This dataset consists of 23 463 images with resolutions ranging from 0.5 to 30 m, including 192 472 instances. The images have been clipped to 800 × 800 for the convenience of model training and testing. It involves 20 common object categories. The training set, validation set, and testing set contain 5862, 5863, and 11 738 samples, respectively. In this article, we jointly use the training set and the validation set to finetune models and conduct the evaluation on the testing set.

2) Implementation Details: For Xview, we train a RetinaNet [102] by following [27] and [37] with the pretrained model for 12 epochs, with a batch size of 8. While Faster-RCNN [76] is adopted when finetuning on DIOR with the same settings except for a batch size of 4. We also apply a linear warming-up strategy with an initial learning rate of 0.000001 at the beginning of 500 iterations. We keep the same layer decay rates as the scene classification task. The basic learning rate, optimizer, and scheduler are the same as Wang et al.'s [30] work. Before input into the network, the large images are uniformly clipped to 416 × 416 pixels. The data augmentation only includes random flipping with a probability of 0.5. We use MMDetection to implement the finetuning, where the AP_{50} is used as the evaluation metric for the comparison of different models.

3) Finetuning Results and Analyses: The experimental results are shown in Table V. We can find that the MTP enhances the performance of all pretrained models, especially for ViT-L + RVSA. On Xview, the performance of MAE pretrained ViT-L + RVSA is not as good as InterImage-XL, even worse than the smaller ResNet-50-based models. After utilizing MTP, the performance of ViT-L + RVSA has been greatly improved. It outperforms CtxMIM [37] and achieves the best. On DIOR, with the help of MTP, ViT-B + RVSA has outperformed all existing methods, including the recently distinguished method SkySense [42] that employs a larger model. In addition, MTP also greatly enhances ViT-L + RVSA, setting a new state-of-the-art. *Here, we emphasize that despite the pretraining dataset SAMRS includes the samples of DIOR [79]. To avoid unfair comparison, following Wang et al.'s [13] work, the images of the testing set in DIOR have not been used for MTP. This rule also applies to other datasets that form the SAMRS, involving DOTA-V1.0 [106], DOTA-V2.0 [78], DIOR-R [107], and FAIR1M-2.0 [80].* It should also be noted that the RVSA model is initially proposed by considering the diverse orientations of RS objects, which are related to the rotated object detection task. Nevertheless, the models after MTP demonstrate an excellent capability in detecting horizontal boxes.

TABLE V
AP_{50} (%) OF FINETUNING DIFFERENT PRETRAINED MODELS WITH RETINANET ON XVIEW AND DIOR DATASETS

Method	Backbone	Xview	DIOR
Random Init.	ResNet-50	10.8	-
Sup. Lea. w IN1K	ResNet-50	14.4	-
Sup. Lea. w IN1K	Swin-B	16.3	-
GASSL [27]	ResNet-50	17.7	67.40
SeCo [29]	ResNet-50	17.2	-
CACO [28]	ResNet-50	17.2	66.91
CtxMIM [37]	Swin-B	18.8*	-
MSFCNet [103]	ResNeSt-101 [104]	-	70.08
TOV [12]	ResNet-50	-	70.16
SATMAE [31]	ViT-L	-	70.89
CSPT [11]	ViT-L	-	71.70
FSoDNet [105]	MSENet	-	71.80
GeRSP [45]	ResNet-50	-	72.20
GFM [46]	Swin-B	-	72.84
Scale-MAE [36]	ViT-L	-	73.81
SatLas [59]	Swin-B	-	74.10
CMID [38]	Swin-B	-	75.11
RingMo [32]	Swin-B	-	75.90
SkySense [42]	Swin-H	-	78.73*
MAE	ViT-B + RVSA	14.6	75.80
MAE + MTP	ViT-B + RVSA	**16.4**	**79.40***
MAE	ViT-L + RVSA	15.0	78.30
MAE + MTP	ViT-L + RVSA	**19.4***	**81.10***
IMP	InternImage-XL	17.0	77.10
IMP + MTP	InternImage-XL	**18.2***	**78.00**

The "Sup. Lea. W IN1K" means supervised learning with imagenet-1k. Random init. Means the backbone network is randomly initialized. Bold indicates the better accuracy of pretrained models with or without MTP. * represents the first three methods among all comparison methods, where the first, second, and third places are emphasized by black, red and green colors, respectively.

C. Rotated Object Detection

We then investigate the impact of MTP on the rotated object detection task, which is a typical RS task distinguished from natural scene object detection because of special overhead views. This is also one of the motivations to implement MTP using SAMRS.

1) Dataset: We adopt the four most commonly used datasets for this task: DIOR-R [107], FAIR1M-2.0 [80], DOTA-V1.0 [106], and DOTA-V2.0 [78].

1) DIOR-R: This is the extended oriented bounding box version of DIOR [79]. It has 23 463 images and 192 158 instances over 20 classes. Each image in this dataset has been cropped into 800 × 800. Following [30], [35], and [42], we merge the original training and validation sets for training, while the testing set is used for evaluation.
2) FAIR1M-2.0: This is a large-scale RS benchmark dataset, including more than 40 000 images and 1 million instances for fine-grained object detection. It collects samples with resolutions ranging from 0.3 to 0.8 m and image sizes ranging from 1000 × 1000 to 10 000 × 10 000 from various sensors and platforms. It contains 37 subcategories belonging to five classes: ship, vehicle, airplane, court, and road. In this article, we use the more challenging version of 2.0, which additionally incorporates the 2021 Gaofen Challenge dataset. The training and validation sets are together adopted for training.
3) DOTA-V1.0: This is the most popular dataset for RS rotated object detection. It comprises 2806 images spanning from 800 × 800 to 4000 × 4000 ×, where 188 282 instances from 15 typical categories are presented. We adopt classical train/test split, that is, the original training and validation sets are together for training, while the original testing set is used for evaluation.
4) DOTA-V2.0: The is the enhanced version of DOTA V1.0. By additionally collecting larger images, adding new categories, and annotating tiny instances, it finally contains 11 268 images, 1 793 658 instances, and 18 categories. We use the combination of training and validation sets for training, while the test-dev set is used for evaluation.

2) Implementation Details: Since the large-size image is not suitable for training, we first perform data cropping. For DOTA-V2.0, we adopt single-scale training and testing by following Xu et al.'s [130] work, where the images are cropped to patches in size of 1024 × 1024 with an overlap of 200. For DOTA-V1.0 and FAIR1M-2.0, we implement the multiscale training and testing, i.e., the original images are scaled with three ratios: (0.5, 1.0, 1.5). Then, the DOTA-V1.0 images are cropped to 1024 × 1024 patches but with an overlap of 500, while FAIR1M-2.0 images adopt a patch size of 800 and an overlap of 400. The batch sizes are set to 4, 16, 4, and 4 for the DIOR-R, FAIR1M, DOTA-V1.0, and DOTA-V2.0 datasets, respectively. The other settings during training are the same as horizontal object detection. We adopt the Oriented-RCNN network implemented in MMRotate. During training, input data is augmented by random flipping and random rotation. The mean average precision (mAP) is adopted as the evaluation metric.

3) Finetuning Results and Analyses: Table VI shows the finetuning results. Except for DIOR-R, we find the MTP pretrained models cannot always demonstrate obvious advantages compared to their counterparts. Since the volumes of FAIR1M-2.0, DOTA-V1.0, and DOTA-V2.0 are much larger than DIOR-R, we speculate that after long-time finetuning, the benefit of MTP becomes diminished. We will further explore this issue in later sections. Nevertheless, owing to the excellent structure, RVSA-L outperforms the ViT-G-based foundation model [35] with over 1 billion parameters on DOTA-V2.0. Compared to the powerful SkySense model [42], our models achieve better performance on the DIOR-R. While on FAIR1M-2.0, except SkySense, our models surpass all other methods by a large margin. Generally, our models have comparable representation capability as SkySense, although it has over 600 M parameters and utilizes 20 million images for pretraining. We also notice the performances of our models still have gaps compared with the current advanced method STD [135] on DOTA-V1.0. It may be attributed to the adopted classical detector Oriented-RCNN [92], which limits the detection performance.

D. Semantic Segmentation

We further consider finetuning the pretrained models on the finer pixel-level tasks, e.g., the semantic segmentation task. It is one of the most important RS applications for the extraction and recognition of RS objects and land covers.

TABLE VI
MAP (%) OF FINETUNING DIFFERENT PRETRAINED MODELS ON THE DIOR-R, FAIR1M-2.0, DOTA-V1.0, AND DOTA-V2.0 DATASETS

Method	**Backbone**	**MS**	**DIOR-R**	**FAIR1M-2.0**	**DOTA-V1.0**	**DOTA-V2.0**
RetinaNet-OBB [102]	ResNet-50	-	57.55	-	-	46.68
Faster RCNN-OBB [76]	ResNet-50	✗	59.54	-	69.36	47.31
FCOS-OBB [108]	ResNet-50	-	-	-	-	48.51
ATSS-OBB [109]	ResNet-50	✗	-	-	72.29	49.57
SCRDet [110]	ResNet-101	✗	-	-	72.61	-
Gilding Vertex [111]	ResNet-50	✗	60.06	-	75.02	-
ROI Transformer [112]	ResNet-50	✓	63.87	-	74.61	52.81
CACo [28]	ResNet-50	-	64.10	47.83	-	-
RingMo [32]	Swin-B	-	-	46.21	-	-
R^{3}Det [113]	ResNet-152	✓	-	-	76.47	-
SASM [114]	ResNeXt-101	✓	-	-	79.17	44.53
AO2-DETR [115]	ResNet-50	✓	-	-	79.22	-
S^{2}ANet [116]	ResNet-50	✓	-	-	79.42	49.86
ReDet [117]	ReResNet-50	✓	-		80.10	
R^{3}Det-KLD [118]	ResNet-50	✓	-	-	80.17	47.26
R^{3}Det-GWD [119]	ResNet-152	✓	-	-	80.23	-
R^{3}Det-DEA [120]	ReResNet-50	✓	-	-	80.37	-
AOPG [107]	ResNet-50	✓	64.41	-	80.66	-
DODet [121]	ResNet-50	✓	65.10	-	80.62	-
PP-YOLOE-R [122]	CRN-x [123]	✓			80.73	
GASSL [27]	ResNet-50	-	65.65	48.15	-	-
SatMAE [31]	ViT-L	-	65.66	46.55	-	-
TOV [12]	ResNet-50	-	66.33	49.62	-	-
Oriented RepPoints [124]	ResNet-50	✗	66.71	-	75.97	48.95
GGHL [125]	DarkNet-53 [126]	✗	66.48	-	76.95	57.17
CMID [38]	Swin-B	✗	66.37	50.58	77.36	-
RSP [5]	ViTAEv2-S	✗	-	-	77.72	-
Scale-MAE [36]	ViT-L	-	66.47	48.31	-	-
SatLas [59]	Swin-B	-	67.59	46.19	-	-
GFM [46]	Swin-B	-	67.67	49.69	-	-
PIIDet [127]	ResNeSt-101	✓	70.35	-	80.21	-
Oriented RCNN [92]	ResNet-50	✓	-	-	80.87	53.28
R^{3}Det-KFIoU [128]	ResNet-152	✓	-	-	81.03	-
RTMDet-R [129]	RTMDet-R-l	✓	-	-	81.33	-
DCFL [130]	ReResNet-101 [117]	✗	71.03	-	-	57.66
SMLFR [131]	ConvNeXt-L [132]	✗	72.33	-	79.33	-
ARC [133]	ARC-R50	✓	-	-	81.77*	-
LSKNet [134]	LSKNet-S	✓	-	-	81.85*	-
STD [135]	ViT-B	✓	-	-	81.66	-
STD [135]	HiViT-B [136]	✓	-	-	82.24*	-
BillionFM [35]	ViT-G12X4	-	73.62*	-	-	58.69*
SkySense [42]	Swin-H	-	74.27*	54.57*	-	-
RVSA [30] †	ViT-B + RVSA	✓	70.67	-	81.01	-
MAE	ViT-B + RVSA	✓	68.06	51.56	**80.83**	55.22
MAE + MTP	ViT-B + RVSA	✓	**71.29**	**51.92**	80.67	**56.08**
MAE	ViT-L + RVSA	✓	70.54	**53.20***	81.43	**58.96***
MAE + MTP	ViT-L + RVSA	✓	**74.54***	53.00*	**81.66**	58.41*
IMP	InternImage-XL	✓	71.14	50.67	80.24	54.85
IMP + MTP	InternImage-XL	✓	**72.17**	**50.93**	**80.77**	**55.13**

MS indicates whether the accuracy on DOTA-V1.0 is obtained from the multi-scale training and testing. †: the feature pyramid is formed by upsampling and downsampling the last layer feature of the backbone network by following the strategy of ViTDet [84].
Bold indicates the better accuracy of pretrained models with or without MTP. * represents the first three methods among all comparison methods, where the first, second, and third places are emphasized by black, red and green colors, respectively.

1) Dataset: We separately take into account both single-class geospatial target extraction and multiclass surface element perception through two RS semantic segmentation datasets: SpaceNetv1 [137] and LoveDA [81]. The details are as follows.

1) SpaceNetv1: This dataset is provided by SpaceNet Challenge [137] for extracting building footprints. It is made up of the DigitalGlobe WorldView-2 satellite imagery with a ground sample distance of 0.5m photoed during 2011–2014 over Rio de Janeiro. It covers about 2544 km^2, including 382 534 building instances. Since only the 6940 images in the original training set are available, following [31] and [37], we randomly split these images into two parts, where the first part containing 5000 images being used as the training set, and another part will be used for testing.
2) LoveDA: This is a challenging dataset involving both urban and rural scenes. It collects 0.3 m spaceborne imagery from Google Earth, where the images were obtained in July 2016, covering 536.15 km^2 of Nanjing, Changzhou, and Wuhan. It has 5987 images in size of 1024 × 1024, involving seven types of common land covers. We merge the official training and validation sets for training and conduct evaluation using the official testing set.

2) Implementation Details: Except that the models are trained with 80 K iterations with a batch size of 8, and the warming up stage in the parameter scheduler lasts 1500

TABLE VII
mIOU (%) OF FINETUNING DIFFERENT PRETRAINED MODELS WITH UPERNET ON THE SPACENETV1 AND LOVEDA DATASETS

Method	Backbone	SpaceNetv1	LoveDA
PSANet [138]	ResNet-50	75.61	-
SeCo [29]	ResNet-50	77.09	43.63
GASSL [27]	ResNet-50	78.51	48.76
SatMAE [31]	ViT-L	78.07	-
CACo [28]	ResNet-50	77.94	48.89
PSPNet [139]	ResNet-50	-	48.31
DeeplabV3+ [140]	ResNet-50	-	48.31
FarSeg [141]	ResNet-50	-	48.15
FactSeg [142]	ResNet-50	-	48.94
TOV [12]	ResNet-50	-	49.70
HRNet [143]	HRNet-W32	-	49.79
GeRSP [45]	ResNet-50	-	50.56
DCFAM [144]	Swin-T	-	50.60
UNetFormer [145]	ResNet-18	-	52.40
RSSFormer [146]	RSS-B	-	52.43
UperNet [62]	ViTAE-B + RVSA [30]	-	52.44
Hi-ResNet [147]	Hi-ResNet	-	52.50
RSP [5]	ViTAEv2-S	-	53.02
SMLFR [131]	ConvNext-L	-	53.03
LSKNet [134]	LSKNet-S	-	54.00
CtxMIM [37]	Swin-B	79.47	-
AerialFormer [148]	Swin-B	-	54.10
BillionFM [35]	ViT-G12X4	-	54.40*
MAE	ViT-B + RVSA	79.56*	51.95
MAE + MTP	ViT-B + RVSA	**79.63***	**52.39**
MAE	ViT-L + RVSA	**79.69***	53.72
MAE + MTP	ViT-L + RVSA	79.54	**54.17***
IMP	InternImage-XL	79.08	53.93*
IMP + MTP	InternImage-XL	**79.16**	**54.17***

Bold indicates the better accuracy of pretrained models with or without MTP. * represents the first three methods among all comparison methods, where the first, second, and third places are emphasized by black, red and green colors, respectively.

iterations, most of the optimization settings are similar to the scene classification section. We use the UperNet [62] as the segmentation framework, where the input image sizes during training are 384 × 384 and 512 × 512 for SpaceNetv1 and LoveDA, respectively, through random scaling and cropping. We also adopt random flipping for data augmentation. All experiments are implemented by MMSegmentation, where the mean value of the intersection over union (mIOU) is adopted as the evaluation metric.

3) Finetuning Results and Analyses: The results presented in Table VII demonstrate that MTP is also useful for enhancing the models' performance on semantic segmentation tasks. Compared to SpaceNetv1, the improvements on the classical land cover classification dataset: LoveDA, are even more significant. As a result, on this dataset, our models surpass all previous methods except the BillionFM [35], which utilizes a model with over 1 billion parameters. On the SpaceNetv1, our models set new state-of-the-art accuracy. Nonetheless, probably due to overfitting, the results of SpaceNetv1 also indicate that the performances on simple extraction tasks do not improve as increasing model capacity. We have also noticed the performance of ViT-L + RVSA on SpaceNetv1 is decreased when adopting MTP. We will conduct further exploration in later sections.

E. Change Detection

Finally, we pay attention to the change detection task, which can be regarded as a special type of segmentation by extracting the changed area between the RS images taken at different times in the same location. Here, we mainly consider the most representative bitemporal change detection.

1) Dataset: We conduct the finetuning on the datasets of different scales: Onera satellite change detection dataset (OSCD) [190], Wuhan University Building Change Detection Dataset (WHU) [191], the learning, vision, and remote sensing change detection dataset (LEVIR) [192], and the season-varying change detection dataset (SVCD) [193], which is also called "CDD."

1) OSCD: This is a small-scale dataset. It contains 24 pairs of Sentinel-2 multispectral images involving all bands and in an average size of 600 × 600. These images are obtained during 2015–2018 to record urban changes. We follow the same train/val split as Daudt et al.'s [190] work, where training and validation sets include 14 and 10 pairs, respectively.
2) WHU: This dataset is used for detecting building changes in a single view. It contains two large-scale images with a ground resolution of 0.3 m and in size of 32 507 × 15 354. They are collected in 2012 and 2016, containing 12 796 and 16 077 instances, respectively. Since there is no official data split, the 70%, 10%, and 20% patches of the cropped images are randomly selected as training, validation, and testing sets as suggested by Zhao et al. [186].
3) LEVIR: This dataset contains 637 pairs of 1024 × 1024 images with a spatial resolution of 0.5m. These images are acquired between 2002 and 2018 from 20 different regions in Texas, USA. It contains 31 333 change instances. We adopt the official split, where training, validation, and testing sets contain 445, 64, and 128 pairs, respectively.
4) SVCD/CDD: This dataset focuses on seasonal variations. It initially contains 11 pairs of images obtained from Google Earth in different seasons, with spatial resolutions ranging from 0.03 to 1m. It now has been cropped to 16 000 pairs of patches in size of 256 × 256 by Ji et al.'s [191] work. The 10 000/3000/3000 pairs are separately used as training, validation, and testing sets.

2) Implementation Details: Following [29] and [42], we crop the OSCD images to 96 × 96 patches with no overlapping, obtaining 827/385 pairs for training/testing. However, the training is difficult to converge due to the extremely small input size, thus we rescale the image to 224 × 224 before inputting it into the network. For the WHU dataset, we separately have 5334, 762, and 1524 images for training, validation, and testing, after cropping the image to patches in size of 256 × 256 without overlaps. A similar operation is conducted for LEVIR, generating training, validation, and testing sets containing 7120, 1024, and 2048 samples, respectively. The training epochs on OSCD, WHU, LEVIR, and CDD are separately set to 100, 200, 150, and 200. The batch size of all datasets is uniformly set to 32. We adopt the same optimization strategy as the scene classification task. To fully leverage the feature pyramid produced by foundation models, we adopt a UNet [77] to process the differences between different temporal features. The training is implemented through Open-CD,[7] where the data augmentation includes random rotation, random flipping, random exchange

[7] [Online]. Available: https://github.com/likyoo/open-cd

TABLE VIII
F1 SCORE (%) OF FINETUNING DIFFERENT PRETRAINED MODELS WITH UNET ON THE OSCD, WHU, LEVIR, AND SVCD/CDD DATASETS

Method	Backbone	OSCD	WHU	LEVIR	SVCD/CDD
GASSL [27]	ResNet-50	46.26	-	-	-
SeCo [29]	ResNet-50	47.67	-	90.14	-
FC-EF [149]	-	48.89	-	62.32	77.11
SwiMDiff [98]	ResNet-18	49.60	-	-	-
CACo [28]	ResNet-50	52.11	-	-	-
SatMAE [31]	ViT-L	52.76	-	-	-
SNUNet [150]	-	-	83.49	88.16	96.20
BIT [151]	ResNet-18	-	83.98	89.31	-
SRCDNet [152]	ResNet-18	-	87.40	-	92.94
CLNet [153]	-	-	-	90.00	92.10
HANet [154]	-	-	-	90.28	-
RECM [155]	ViT-S	-	-	90.39	-
ChangeFormer [156]	MiT-B2 [157]	-	-	90.40	-
AERNet [158]	ResNet-34	-	-	90.78	-
ESCNet [159]	-	-	-	-	93.54
DSAMNet [160]	ResNet-18	-	-	-	93.69
GCD-DDPM [161]	-	-	92.54	90.96	94.93
CDContrast [162]	-	-	-	-	95.11
DDPM-CD [163]	UNet [77]	-	92.65	90.91	95.62
DeepCL [164]	EfficientNet-b0 [165]	-	-	91.11	-
DMNet [166]	ResNet-50	-	-	-	95.93
ChangeStar [167]	ResNeXt-101 [168]	-	-	91.25	-
RSP [5]	ViTAEv2-S [58]	-	-	90.93	96.81
SAAN [169]	ResNet-18	-	-	91.41	97.03
SiamixFormer [170]	MiT-B5 [157]	-	-	91.58	97.13
TransUNetCD [171]	ResNet-50	-	93.59	91.11	97.17
RDPNet [172]	-	-	-	90.10	97.20
SDACD [173]	-	-	-	-	97.34
Siam-NestedUNet [174]	UNet++ [175]	-	-	91.50	-
Changen [176]	MiT-B1 [157]	-	-	91.50	-
HCGMNet [177]	VGG-16	-	-	91.77	-
CEECNet [178]	-	-	-	91.83	-
RingMo [32]	Swin-B	-	-	91.86	-
CGNet [179]	VGG-16	-	-	92.01	-
TTP [180]	SAM [47]	-	-	92.10	-
Changer [181]	ResNeSt-101 [104]	-	-	92.33	-
WNet [182]	ResNet-18 + DAT [183]	-	91.25	90.67	97.56
SpectralGPT$^+$ [34]	ViT-B	54.29	-	-	-
C2FNet [184]	VGG-16	-	-	91.83	-
MATTER [24]	ResNet-34	59.37*	-	-	-
GFM [46]	Swin-B	59.82*	-	-	-
FMCD [185]	EfficientNet-b4 [165]	-	94.48	-	-
SGSLN/256 [186]	-	-	94.67	91.93	96.24
P2V-CD [187]	-	-	92.38	91.94	98.42*
ChangeCLIP [188]	CLIP [17]	-	94.82*	92.01	97.89
BAN [189]	InternImage-XL [82]	-	-	91.94	-
BAN [189]	ViT-L [82]	-	-	91.96	-
BAN [189]	ChangeFormer [156]	-	-	92.30	-
SkySense [42]	Swin-H	60.06*	-	92.58*	-
MAE	ViT-B + RVSA	50.28	93.77	92.21	97.80
MAE + MTP	ViT-B + RVSA	**53.36**	**94.32**	**92.22**	**97.87**
MAE	ViT-L + RVSA	54.04	94.07	92.52	97.78
MAE + MTP	ViT-L + RVSA	**55.92**	**94.75**	**92.67***	**97.98**
IMP	InternImage-XL	51.61	95.33*	92.46	**98.37***
IMP + MTP	InternImage-XL	**55.61**	**95.59***	**92.54***	98.33*

Bold indicates the better accuracy of pretrained models with or without MTP. * represents the first three methods among all comparison methods, where the first, second, and third places are emphasized by black, red and green colors, respectively.

temporal, and color jitters that randomly adjust brightness, contrast, hue, and saturation of images. The F1 score of the changed class is adopted as the evaluation metric.

3) Finetuning Results and Analyses: To comprehensively assess the finetuning performance of pretrained models, we conduct the comparison by collecting existing advanced change detection methods, as shown in Table VIII. It should be noted that, since the original WHU dataset does not provide an official train/test split, various split strategies are adopted in different methods. Therefore, on this dataset, we only list the accuracy value of the methods that employ the same settings as us or training with more images. It can be seen that MTP effectively improves the performances of pretrained models on these datasets. Especially, our models perform well on three large-scale datasets: WHU, LEVIR, and SVCD/CDD. Even if adopting simple UNet [77] and the RVSA model of the base version, the finetuning performances have been competitive and surpassed many advanced approaches. When utilizing larger models, the performance can be further boosted. Finally, they achieve the best accuracy on the WHU and LEVIR datasets by outperforming almost all existing methods, including the recent SkySense [42] that builds a larger change detection network with

TABLE IX
DETAILED HYPERPARAMETER SETTINGS IN FINETUNING PRETRAINED MODELS ON DIFFERENT DATASETS

Dataset	*Scene classification*		*Horizontal detection*		*Rotated object detection*			
	EuroSAT	RESISC-45	Xview	DIOR	DIOR-R	FAIR1M-2.0	DOTA-V1.0	DOTA-2.0
Training image number (N_{TrIm})	16 200	6300	20 084	11 725	11 725	288 428	133 883	31 273
Training epoch number (N_{TrEp})	100	200	12	12	12	12	12	40
Total sample number (N_{ToSa})	1 620 000	1 260 000	241 008	140 700	140 700	3 461 136	1 606 596	1 250 920
Batch size (S_B)	64	64	8	4	4	16	4	4
Total iteration number (N_{ToIt})	25 312	19 688	30 126	35 175	35 175	216 321	401 649	312 730
Training image size (N_{TrIm})	224	224	416	800	800	800	1024	1024
Class number (N_C)	10	45	60	20	20	37	15	18
Average pixel per class (AP$_C$)	36 288 000	6 272 000	167 0988	5 628 000	5 628 000	74 835 373	109 676 954	71 163 449
Average iteration per class (AI$_C$)	2531	438	502	1759	1759	5847	26 777	17 374
ViT-B + RVSA	✔	✔	✔	✔	✔	✔	✘	✔
ViT-L + RVSA	✔	✔	✔	✔	✔	✘	✔	✘
InternImage-XL	✘	✔	✔	✔	✔	✔	✔	✔
Dataset	*Semantic segmentation*		*Bitemporal change detection*					
	SpaceNetv1	LoveDA	OSCD	WHU	LEVIR	SVCD/CDD		
Training image number (N_{TrIm})	5000	4191	827	5334	7120	10 000		
Training epoch number (N_{TrEp})	128	153	100	200	150	200		
Total sample number (N_{ToSa})	640 000	640 000	82 700	106 6800	1 068 000	2 000 000		
Batch size (S_B)	8	8	32	32	32	32		
Total iteration number (N_{ToIt})	80 000	80 000	2584	33 338	33 375	62 500		
Training image size (N_{TrIm})	384	512	224	256	256	256		
Class number (N_C)	2	7	2	2	2	2		
Average pixel per class (AP$_C$)	122 880 000	46 811 429	9 262 400	136,550,400	136 704 000	256 000 000		
Average iteration per class (AI$_C$)	40 000	11 429	1292	16 669	16 688	31 250		
ViT-B + RVSA	✔	✔	✔	✔	✔	✔		
ViT-L + RVSA	✘	✔	✔	✔	✔	✔		
InternImage-XL	✔	✔	✔	✔	✔	✘		

" ✔ " and " ✘ " indicate whether the MTP is useful for improving performance compared to the setting without MTP.

over 600 M parameters, ChangeCLIP [188] that uses CLIP [17] to obtain additional knowledge from language modalities, and the newly proposed adapter BAN [189], where the ability of existing foundation model and change detection approaches can be exploited. Different from large-scale scenes, on the small-scale dataset OSCD, although MTP is still useful, the performances of our models have relatively large gaps compared to current works. We attribute the reason to data discrepancy and image size. Specifically, our models are pretrained on high-resolution RS images, which are similar to another three change detection datasets. However OSCD images are captured by multispectral sensors with lower resolutions. In addition, OSCD images are only cropped to 96 × 96 during training, which may be unsuitable for feature extraction, especially for nonhierarchical vision transformers. In the comparison methods, MATTER [24], SkySense [42], and GFM [46] use pyramid feature networks. Among them, MATTER and SkySense adopt Sentinel-2 multispectral image in pretraining, while GFM crops the OSCD image to a larger size, i.e., 192 × 192. In contrast, a relatively small image size (128 × 128) restricts the performances of ViT in SpectralGPT [34]. These results suggest that it is necessary to conduct further explorations to enhance the model finetuning performance on out-of-domain datasets with small volumes and input sizes.

F. Further Investigations and Analyses

Besides evaluating the performances of pretrained models, we conduct further investigations to obtain deeper insights into the characteristics of MTP, including the influence factors of MTP, finetuning with fewer samples, and parameter reusing of decoders.

1) Influence Factors of MTP: Up to now, to comprehensively assess the impact of MTP, we have finetuned three types of foundation models, on five RS downstream tasks, involving a total of fourteen datasets. From the finetuning results (see Tables IV–VIII) we find that MTP improves these foundation models in most cases. But there are still some datasets, on which MTP does not perform well as expected, i.e., not all accuracies of three models are increased. To figure out the reason, we explore the influence factors related to the performance of MTP, as shown in Table IX. Intuitively, we suppose MTP may be affected by the characteristics of finetuning datasets and consider a series of variables, including "training image number" (N_{TrIm}), "training epoch number" (N_{TrEp}), "Batch Size" (S_B), and "training image size" (S_{TrIm}). The "training image number" means: for each dataset, the number of images used for training. For example, the N_{TrIm} of DIOR is 11 725 since the original training and validation sets are together used for training. While "training image size" represents the image size after data augmentation and preprocessing. Theoretically, we have

$$N_{\text{ToIt}} = \frac{N_{\text{TrIm}} \cdot N_{\text{TrEp}}}{S_B} \tag{7}$$

where N_{ToIt} means the number of training iterations for model parameter updating under the minibatch optimization strategy. In Table IX, we observe that as N_{ToIt} increases, there is a tendency for MTP to have a negative impact on finetuning performance for a given task. However, this trend is not universally applicable, as evidenced by varying results among pretrained models in segmentation tasks, all with the same N_{ToIt}. In addition, we account for dataset difficulty by considering the number of classes N_C and use the "average iteration per class" (AI$_C$) to represent each dataset as

$$\mathrm{AI}_C = \frac{N_{\mathrm{ToIt}}}{N_C}. \tag{8}$$

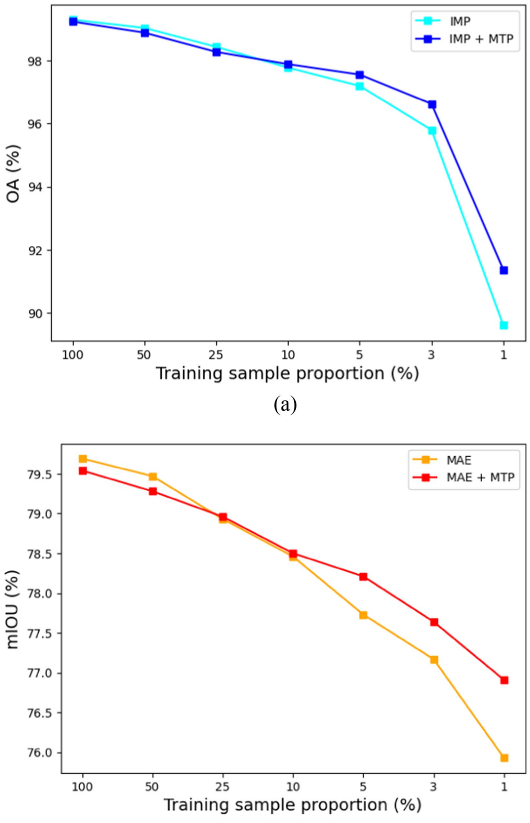
Surprisingly, Table IX reveals a notable trend: a relatively large AI_C corresponds to a negative impact of MTP for the same task. This suggests that, over extended finetuning periods, MTP models lose their advantage compared to conventional pretrained models. We propose a bold conjecture regarding this internal mechanism: *the benefits of MTP diminish gradually due to excessive network optimization*. This discovery prompts a reconsideration of the tradeoff between longer training times for accuracy gains and the benefits of pretraining when finetuning models. However, determining the critical point of AI_C remains challenging due to limited experimentation, necessitating further investigation. It is important to note that this phenomenon differs from overfitting, as our models continue to outperform existing methods at this stage.

In addition to training duration, we consider dataset capacity, introducing the index "average pixels per class" AP_C, which can be formulated by

$$\mathrm{AP}_C = \frac{N_{\mathrm{ToSa}} \cdot S_{\mathrm{TrIm}}}{N_C} \tag{9}$$

where $N_{\mathrm{ToSa}} = N_{\mathrm{TrIm}} \cdot N_{\mathrm{TrEp}}$ denotes the quantity of images processed during training. Consequently, AP_C approximately reflects the data volume encountered during finetuning. Table IX reveals that AP_C exhibits similar trends to AI_C, yet the correlation between AP_C and MTP performance is less discernible compared to AI_C, possibly due to the presence of redundant pixels in RS images.

2) Fewer Sample Finetuning: The efficacy of SEP has been demonstrated in scenarios with limited samples [13]. While MTP represents an extension of SEP, it is reasonable to anticipate that MTP could excel in analogous contexts. Moreover, as noted earlier, MTP primarily addresses the discrepancy between upstream pretraining and downstream finetuning tasks. This encourages us to consider that fewer downstream training samples might better showcase MTP's efficacy in facilitating efficient transfer from pretraining models. To explore this, we finetune InterImage-XL on EuroSAT and ViT-L + RVSA on SpaceNetv1, respectively, progressively reducing training samples. The results are depicted in Fig. 3. Initially, MTP's performance is slightly inferior to its counterparts when the training sample proportion is 100%, as illustrated in Tables IV and VII. However, as training samples decrease, the performance curves converge until the training sample proportion is 10%, at which point MTP's impact is minimal. Subsequent reductions in training samples lead to decreased accuracies across all models, yet the distances between the curves progressively widen. This trend suggests that the benefits of MTP are beginning to emerge, becoming increasingly significant. These findings validate our hypotheses, underscoring the benefit of MTP for finetuning foundational models on limited training samples.



(a)

(b)

Fig. 3. Finetuning accuracy of different pretrained models with varying training sample sizes. (a) InternImage-XL on EuroSAT. (b) ViT-L + RVSA on SpaceNetv1.

TABLE X
ACCURACIES OF FINETUNING VIT-B + RVSA ON VARIOUS DATASETS WITH AND WITHOUT REUSING PRETRAINED DECODER WEIGHTS

Model	SpaceNetv1	LoveDA	DIOR-R
w/o DPR	**79.63**	**52.39**	71.29
w DPR	79.54	51.83	**71.94**
Model	**FAIR1M-2.0**	**DOTA-V1.0**	**DOTA-V2.0**
w/o DPR	51.92	**80.67**	55.22
w DPR	**52.19**	80.54	**55.78**

Bold indicates the better accuracy of pretrained models with or without MTP.

3) Decoder Parameter Reusing: MTP utilizes task-specific decoders for segmentation and detection tasks. Hence, reusing these decoder weights during finetuning seems a natural choice, and we conduct experiments accordingly using a backbone of ViT-B + RVSA. Specifically, aside from the backbone network, we also initialize the corresponding decoders with pretrained weights during finetuning. However, only semantic segmentation and rotated detection decoders are eligible for reuse, as per the segmentor or detector used in existing methods. Therefore, we performed the experiment on the corresponding six datasets, and the results have been presented in Table X. Among four detection datasets, decoder parameter reusing (DPR) proves beneficial in three scenarios, which are actually employed in performing MTP. Nevertheless, on another classical dataset

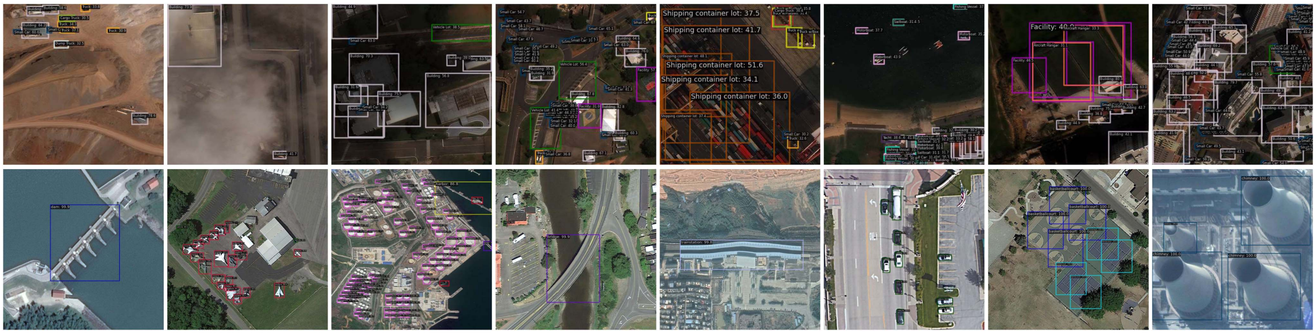

Shipping container lot: 37.5
Shipping container lot: 41.7
Shipping container lot: 51.6
Shipping container lot: 34.1
Shipping container lot: 36.0
Facility: 40.

Fig. 4. Visualization of the horizontal object detection predictions of MAE + MTP pretrained ViT-L + RVSA. The images of the first and the second rows are from Xview and DIOR testing sets, respectively.

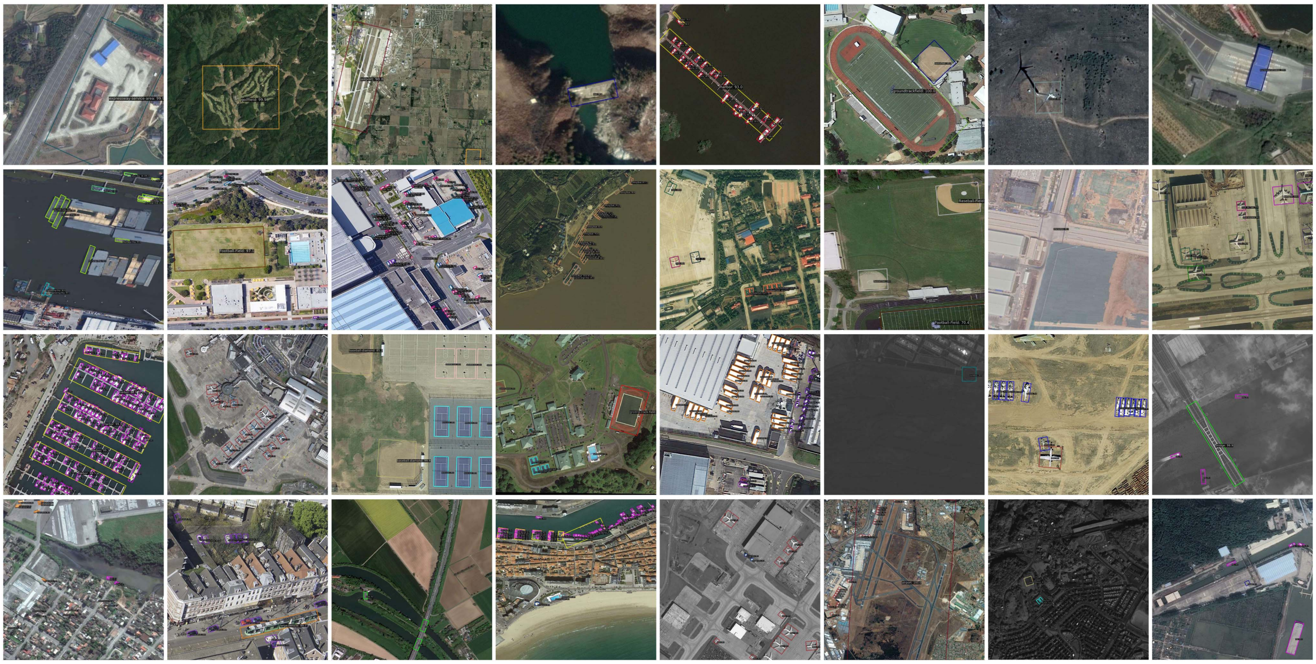


Fig. 5. Visualization of the rotated object detection predictions of MAE + MTP pretrained ViT-L + RVSA. The images in four rows are from the testing sets of DIOR-R, FAIR1M-2.0, DOTA-V1.0, and DOTA-V2.0, respectively.

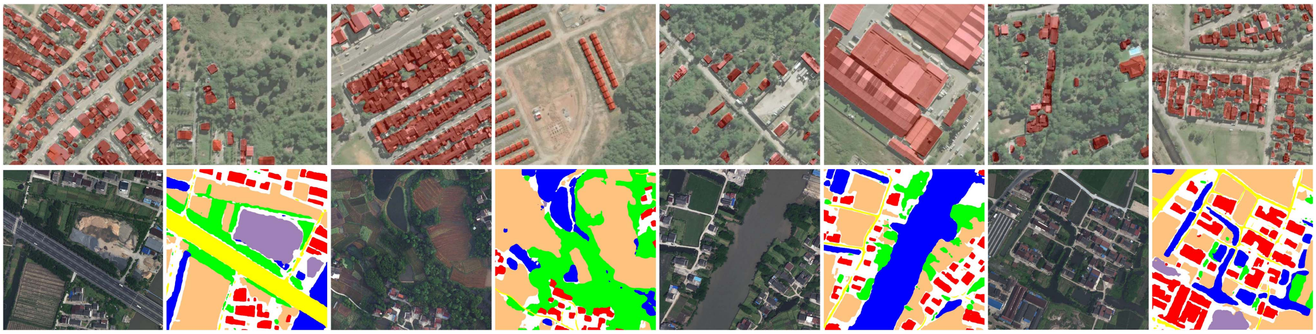


Fig. 6. Visualization of the semantic segmentation predictions of MAE + MTP pretrained ViT-L + RVSA. The samples of the first and the second rows are from SpaceNetv1 and LoveDA testing sets, respectively.

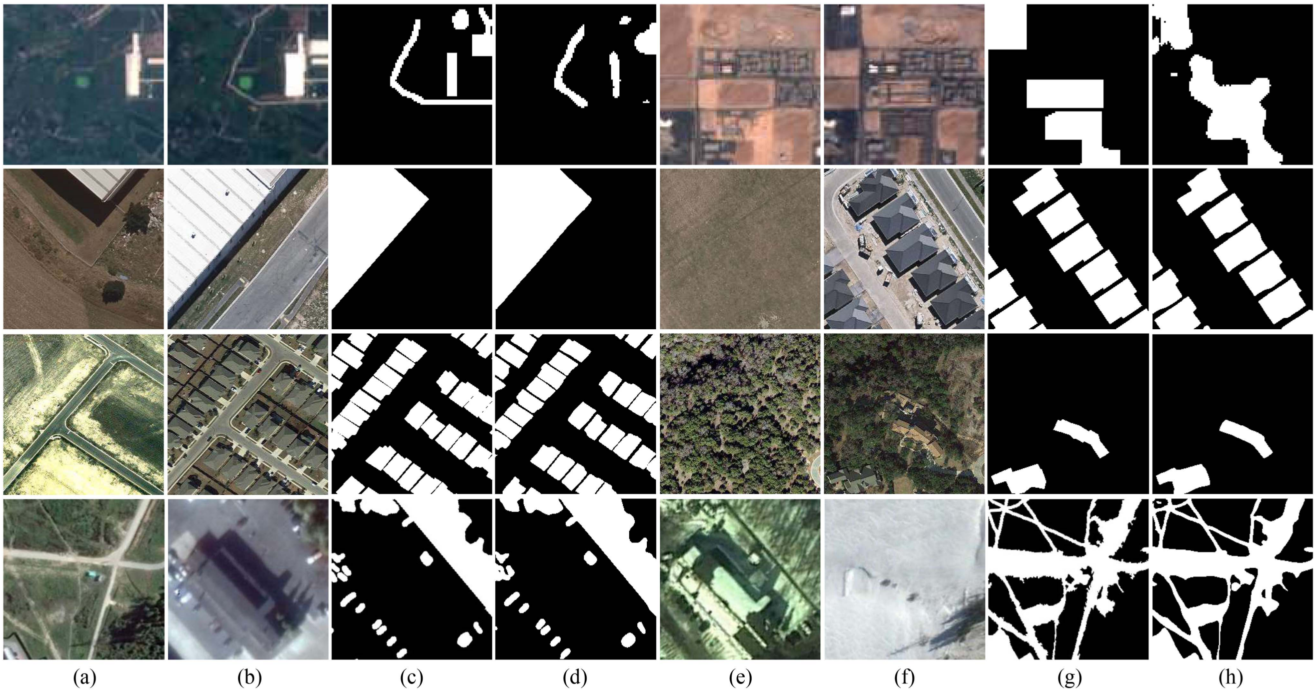


Fig. 7. Visualization of the bitemporal change detection predictions of MAE + MTP pretrained ViT-L + RVSA. The samples in four rows are from the testing sets of OSCD, WHU, LEVIR, and SVCD/CDD, respectively. (a), (b), (e), and (f) depict bitemporal images of different samples, with (c) and (g) representing corresponding ground truth labels. Our prediction maps are shown at (d) and (h).

DOTA-V1.0, which can be regarded as a subset of DOTA-V2.0, DPR is insufficient useful. On segmentation tasks, the performances of DPR models are decreased. We speculate that besides the domain differences compared to pretraining datasets, current models may be additionally affected by the segmentation labels in pretraining. This is because decoders typically encode task-specific information. However, given that the SAMRS dataset used for pretraining involves annotations generated by SAM [47], they inevitably contain errors, jeopardizing the quality of pretrained decoders. Finally, based on the above-mentioned considerations, we conclude that *after MTP, the performances of reusing pretrained decoder parameters in finetuning may depend on the similarity between pretraining and finetuning scenarios and the quality of pretraining annotations.*

G. Visualization

To further show the efficacy of MTP in enhancing RS foundation models, we present the predictions of MAE + MTP pretrained ViT-L + RVSA across detection, segmentation, and change detection tasks in Figs. 4–7. For detection, we demonstrate results across diverse scenes using horizontal or rotated bounding boxes. For segmentation, we display the original images alongside segmentation maps, highlighting building extraction masks in red. For change detection, we provide the bi-temporal images, ground truths, and predicted change maps. Our model accurately detects RS objects, extracts buildings, classifies land cover categories, and characterizes changes across diverse types. In summary, MTP enables the construction of an RS foundation model with over 300 parameters, which achieves superior representation capability for various downstream tasks.

V. Conclusion

In this article, we introduce the MTP approach for building RS foundation models. MTP utilizes a shared encoder and task-specific decoder architecture to effectively pretrain CNNs and vision transformer backbones on three tasks: semantic segmentation, instance segmentation, and rotated object detection in a unified supervised learning framework. We evaluate MTP by examining the finetuning accuracy of these pretrained models on 14 datasets covering various downstream RS tasks. Our results demonstrate the competitive performance of these models compared to existing methods, even with larger models. Further experiments indicate that MTP excels in low-data finetuning scenarios but may offer diminishing returns with prolonged finetuning on large-scale datasets. We hope this research encourages further exploration of RS foundation models, especially in resource-constrained settings. In addition, we anticipate the widespread application of these models across diverse fields of RS image interpretation due to their strong representation capabilities.

Appendix

We present detailed finetuning accuracies of the three models, i.e., ViT-B + RVSA, ViT-L + RVSA, and InternImage-XL, on the DIOR, DIOR-R, FAIR1M-2.0, DOTA-V1.0, DOTA-V2.0, and LoveDA datasets in Tables XI–XVI.

TABLE XI
DETAILED ACCURACIES OF DIFFERENT MODELS ON DIOR DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
airplane	68.2	87.5	76.5	93.8	65.0	69.0
airport	91.2	92.1	92.6	91.6	91.3	92.8
baseballfield	79.9	87.3	83.2	87.7	75.6	81.3
basketballcourt	88.0	89.4	90.7	92.1	89.3	90.1
bridge	53.6	58.0	58.8	64.6	59.3	59.1
chimney	82.1	83.7	84.1	85.9	84.9	84.8
expressway-service-area	90.6	92.9	92.3	94.3	92.8	93.9
expressway-toll-station	76.2	80.5	79.5	84.5	84.4	83.6
dam	78.2	82.0	79.3	81.4	80.2	82.0
golffield	84.9	88.1	85.7	87.3	86.2	83.9
groundtrackfield	83.9	85.6	85.3	86.7	85.9	87.4
harbor	56.8	62.4	60.4	64.5	62.2	63.3
overpass	67.4	69.8	70.6	72.0	68.8	68.6
ship	74.4	75.4	75.4	76.3	73.6	73.6
stadium	82.8	85.8	84.3	85.6	83.7	85.5
storagetank	61.4	62.5	65.3	62.6	59.6	57.7
tenniscourt	89.4	91.2	91.3	92.5	87.6	90.3
trainstation	76.1	80.3	76.6	79.7	77.7	77.7
vehicle	45.2	47.4	48.2	47.6	45.0	46.2
windmill	85.0	86.5	85.1	90.2	89.3	89.3
mAP	75.8	79.4	78.3	81.1	77.1	78.0

TABLE XII
DETAILED ACCURACIES OF DIFFERENT MODELS ON DIOR-R DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
airplane	72.1	89.6	81.2	90.7	72.0	72.3
airport	51.1	52.6	51.9	63.4	61.5	63.6
baseballfield	80.8	81.2	81.1	90.0	80.6	80.9
basketballcourt	81.3	87.8	90.1	90.1	90.0	90.1
bridge	44.9	48.6	48.1	56.4	53.5	54.7
chimney	72.7	77.2	78.2	81.5	81.5	81.5
expressway-service-area	87.5	89.1	88.4	89.4	89.9	89.7
expressway-toll-station	69.3	71.6	74.7	80.1	79.5	79.6
dam	35.5	43.3	39.8	39.9	43.0	45.9
golffield	78.4	79.0	79.4	79.3	80.0	79.3
groundtrackfield	81.9	84.2	84.3	85.1	85.2	85.3
harbor	43.3	51.3	46.4	56.0	54.8	55.2
overpass	60.1	60.9	60.5	67.2	64.2	65.8
ship	81.2	81.2	81.2	81.1	81.3	81.3
stadium	81.6	83.7	83.4	78.9	78.2	79.2
storagetank	70.5	71.2	71.4	71.4	62.7	62.8
tenniscourt	89.2	90.2	90.0	90.4	81.5	90.1
trainstation	65.6	66.6	65.2	73.9	66.7	67.5
vehicle	49.3	50.5	51.0	51.8	50.5	51.1
windmill	65.1	66.0	64.6	74.3	66.1	67.2
mAP	68.1	71.3	70.5	74.5	71.1	72.2

TABLE XIII
DETAILED ACCURACIES OF DIFFERENT MODELS ON FAIR1M-2.0 DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
Boeing737	47.17	44.78	48.91	43.21	41.36	47.48
Boeing747	93.31	93.73	95.00	94.95	95.76	95.67
Boeing777	41.48	45.01	36.86	36.20	50.87	44.57
Boeing787	60.01	57.03	64.70	61.42	63.66	66.78
C919	42.43	44.27	33.94	50.14	51.95	54.52
A220	53.64	52.23	56.51	52.70	54.10	58.17
A321	71.56	72.12	72.98	72.68	69.31	70.76
A330	61.97	64.22	60.78	64.01	68.53	64.66
A350	75.76	73.14	75.94	70.90	80.19	78.89
ARJ21	20.36	20.93	19.66	23.14	22.67	22.78
Passenger ship	20.12	21.29	20.38	19.80	15.76	15.55
Motorboat	71.17	71.73	72.72	73.07	64.86	66.28
Fishing boat	35.42	36.30	36.30	33.94	28.38	31.34
Tugboat	31.12	32.66	36.00	32.52	27.33	25.38
Engineering ship	22.75	25.90	29.02	28.63	20.16	21.00
Liquid cargo ship	48.73	49.30	52.80	49.31	46.91	46.32
Dry cargo ship	53.07	53.12	53.87	51.09	49.18	49.13
Warship	38.17	40.88	45.66	43.44	33.45	38.04
Small car	76.77	76.98	77.65	77.23	72.77	72.92
Bus	45.60	42.16	51.73	51.73	47.59	46.79
Cargo truck	59.64	59.87	61.56	60.53	56.15	56.32
Dump truck	61.73	61.85	63.08	60.95	57.69	57.48
Van	77.08	77.33	78.22	77.74	73.00	73.08
Trailer	19.74	19.34	23.63	22.48	14.48	17.55
Tractor	1.33	1.79	2.21	1.83	0.71	1.23
Excavator	23.38	25.03	27.58	29.12	21.49	18.68
Truck tractor	50.00	49.83	48.71	52.15	50.62	48.44
Basketball court	64.50	63.35	63.46	64.40	60.69	60.46
Tennis court	91.45	90.71	92.09	91.53	89.07	90.22
Football field	66.23	67.21	70.72	70.75	68.44	68.98
Baseball field	91.81	91.52	92.27	91.80	87.93	88.78
Intersection	63.65	64.73	64.94	66.22	64.30	63.28
Roundabout	26.13	25.43	28.36	31.00	30.71	27.07
Bridge	45.91	49.38	50.66	51.32	42.81	43.33
mAP	51.56	51.92	53.20	53.00	50.67	50.94

TABLE XIV
DETAILED ACCURACIES OF DIFFERENT MODELS ON DOTA-V1 DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
plane	88.42	88.91	88.52	88.33	88.91	88.96
baseball-diamond	85.03	84.07	85.36	86.59	86.78	85.93
bridge	60.86	60.76	61.55	63.38	59.93	60.62
ground-track-field	82.39	82.93	81.25	83.49	81.05	81.26
small-vehicle	80.70	80.41	80.69	81.06	80.80	80.08
large-vehicle	85.76	86.16	86.44	86.48	85.06	84.55
ship	88.58	88.64	88.51	88.53	88.38	87.96
tennis-court	90.88	90.87	90.87	90.87	90.81	90.86
basketball-court	86.61	86.21	85.80	86.26	86.27	86.37
storage-tank	86.88	86.76	86.84	85.80	86.19	86.50
soccer-ball-field	63.79	63.91	69.81	67.17	69.64	68.93
roundabout	72.52	71.00	72.51	71.84	71.50	73.11
harbor	78.52	79.04	84.82	84.94	79.21	78.76
swimming-pool	80.53	82.17	79.99	81.93	81.50	80.98
helicopter	81.00	78.34	78.51	78.31	67.57	76.63
mAP	80.83	80.68	81.43	81.66	80.24	80.77

TABLE XV
DETAILED ACCURACIES OF DIFFERENT MODELS ON DOTA-V2 DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
plane	77.86	78.30	79.16	78.57	78.52	70.98
baseball-diamond	48.99	52.58	48.16	45.54	50.96	50.82
bridge	46.55	48.42	47.97	49.72	43.21	43.60
ground-track-field	63.10	59.05	61.57	56.42	59.77	59.25
small-vehicle	43.55	43.65	43.74	43.78	43.58	43.55
large-vehicle	56.85	57.15	61.14	62.26	56.11	56.50
ship	61.09	61.08	68.60	68.76	61.38	61.41
tennis-court	76.90	77.83	78.45	74.89	77.61	78.23
basketball-court	54.57	56.32	61.97	64.17	58.20	61.41
storage-tank	58.55	59.23	59.62	58.70	58.55	51.30
soccer-ball-field	36.37	36.93	45.98	43.70	48.88	42.89
roundabout	50.39	51.26	54.89	49.46	50.17	50.60
harbor	56.34	56.97	62.03	63.32	57.86	56.84
swimming-pool	63.89	63.05	64.34	64.59	58.43	58.31
helicopter	65.43	66.40	70.23	72.71	58.16	59.55
container-crane	39.19	44.24	46.94	50.91	40.58	49.21
airport	79.90	87.57	87.64	87.64	77.39	84.73
helipad	14.43	9.33	18.92	16.27	7.91	13.09
mAP	55.22	56.08	58.96	58.41	54.85	55.13

TABLE XVI
DETAILED ACCURACIES OF DIFFERENT MODELS ON LOVEDA DATASET

Category	**ViT-B + RVSA w/o MTP**	**ViT-B + RVSA w MTP**	**ViT-L + RVSA w/o MTP**	**ViT-L + RVSA w MTP**	**InternImage-XL w/o MTP**	**InternImage-XL w MTP**
background	45.91	45.92	47.26	47.14	46.63	46.80
building	57.93	59.40	59.27	62.69	61.98	62.60
Road	56.08	56.15	59.54	58.00	58.25	58.96
water	79.72	80.66	81.45	81.43	82.14	82.25
barren	16.49	16.56	17.59	19.27	18.11	17.49
forest	46.03	46.38	47.39	46.82	47.99	47.63
agriculture	61.48	61.67	63.55	63.80	62.40	63.44
mIOU	51.95	52.39	53.72	54.17	53.93	54.17

ACKNOWLEDGMENT

The numerical calculations in this article are partly supported by the Dawning Information Industry Company Ltd.

REFERENCES

[1] Z. Zhu et al., "Understanding an urbanizing planet: Strategic directions for remote sensing," *Remote Sens. Environ.*, vol. 228, pp. 164–182, 2019.

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

[3] F. Dell'Acqua and P. Gamba, "Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives," *Proc. IEEE*, vol. 100, no. 10, pp. 2876–2890, Oct. 2012.

[4] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[5] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608020.

[6] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6172–6180.

[7] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.

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

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

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

[11] T. Zhang et al., "Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain," *Remote Sens.*, vol. 14, no. 22, 2022.

[12] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "TOV: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4916–4930, 2023.

[13] D. Wang et al., "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 8815–8827.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

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

[16] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640.

[17] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

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

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

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

[21] H. Bao, L. Dong, S. Piao, and F. Wei, “BEiT: BERT pre-training of image transformers,” in *Proc. Int. Conf. Learn. Representations*, 2022.

[22] Z. Xie et al., “SimMIM: A simple framework for masked image modeling,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9643–9653.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[24] P. Akiva, M. Purri, and M. Leotta, “Self-supervised material and texture representation learning for remote sensing tasks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8193–8205.

[25] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, “CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 23498–23515.

[26] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, “GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 8690–8701.

[27] K. Ayush et al., “Geography-aware self-supervised learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10161–10170.

[28] U. Mall, B. Hariharan, and K. Bala, “Change-aware sampling and contrastive learning for satellite images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5261–5270.

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

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

[31] Y. Cong et al., “SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 197–211.

[32] X. Sun et al., “RingMo: A remote sensing foundation model with masked image modeling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612822.

[33] F. Yao et al., “RingMo-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620821.

[34] D. Hong et al., “SpectralGPT: Spectral remote sensing foundation model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3362475.

[35] K. Cha, J. Seo, and T. Lee, “A billion-scale foundation model for remote sensing images,” 2023, *arXiv:2304.05215*.

[36] C. J. Reed et al., “Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4065–4076.

[37] M. Zhang, Q. Liu, and Y. Wang, “CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding,” 2023, *arXiv:2310.00022*.

[38] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, “CMID: A unified self-supervised learning framework for remote sensing image understanding,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607817.

[39] M. Tang, A. Cozma, K. Georgiou, and H. Qi, “Cross-scale MAE: A tale of multiscale exploitation in remote sensing,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 20054–20066.

[40] A. Fuller, K. Millard, and J. Green, “CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 5506–5538.

[41] Y. Wang, H. H. Hernández, C. M. Albrecht, and X. X. Zhu, “Feature guided masked autoencoder for self-supervised learning in remote sensing,” 2023, *arXiv:2310.18653*.

[42] X. Guo et al., “SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery,” 2023, *arXiv:2312.10115*.

[43] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, “DeCUR: Decoupling common & unique representations for multimodal self-supervision,” 2023, *arXiv:2309.05300*.

[44] Y. Feng et al., “A self-supervised cross-modal remote sensing foundation model with multi-domain representation and cross-domain fusion,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 2239–2242.

[45] Z. Huang, M. Zhang, Y. Gong, Q. Liu, and Y. Wang, “Generic knowledge boosted pretraining for remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5605913.

[46] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, “Towards geospatial foundation models via continual pretraining,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16760–16770.

[47] A. Kirillov et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[48] X.-Y. Tong et al., “Land-cover classification with high-resolution remote sensing images using transferable deep models,” *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

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

[51] C. Jun, Y. Ban, and S. Li, “Open access to Earth land-cover map,” *Nature*, vol. 514, no. 7523, pp. 434–434, 2014.

[52] W. Li, K. Chen, and Z. Shi, “Geographical supervision correction for remote sensing representation learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411520.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[54] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[55] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[57] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[58] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, “ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond,” *Int. J. Comput. Vis.*, vol. 131, pp. 1141–1162, 2023.

[59] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, “SatlasPretrain: A large-scale dataset for remote sensing image understanding,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16726–16736.

[60] S. Gururangan et al., “Don’t stop pretraining: Adapt language models to domains and tasks,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8342–8360.

[61] L. M. Dery, P. Michel, A. Talwalkar, and G. Neubig, “Should we be pre-training? An argument for end-task aware training as an alternative,” in *Proc. Int. Conf. Learn. Representations*, 2022.

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

[63] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1280–1289.

[64] W. Li, H. Chen, and Z. Shi, “Semantic segmentation of remote sensing images with self-supervised multitask representation learning,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6438–6450, 2021.

[65] Q. Li, Y. Chen, X. He, and L. Huang, “Co-training transformer for remote sensing image classification, segmentation, and detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5606218.

[66] L. Yuan et al., “Florence: A new foundation model for computer vision,” 2021, *arXiv:2111.11432*.

[67] C. Wu et al., “NÜWA: Visual synthesis pre-training for neural visual world creation,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 720–736.

[68] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[69] L. H. Li et al., “Grounded language-image pre-training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10955–10965.

[70] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “CoCa: Contrastive captioners are image-text foundation models,” *Trans. Mach. Learn. Res.*, 2022.

[71] W. Wang et al., “Image as a foreign language: Beit pretraining for vision and vision-language tasks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19175–19186.

[72] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, “ViTAE: Vision transformer advanced by exploring intrinsic inductive bias,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 28522–28535.

[73] Q. Zhang, J. Zhang, Y. Xu, and D. Tao, “Vision transformer with quadrangle attention,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3608–3624, May 2024.

[74] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, “RSGPT: A remote sensing vision language model and benchmark,” 2023, *arXiv:2307.15266*.

[75] C. Wu, B. Du, and L. Zhang, “Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.

[76] 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.

[77] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[78] J. Ding et al., “Object detection in aerial images: A large-scale benchmark and challenges,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

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

[80] X. Sun et al., “FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.

[81] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, “LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation,” in *Proc. NeurIPS Track Datasets Benchmarks*, 2021.

[82] W. Wang et al., “InternImage: Exploring large-scale vision foundation models with deformable convolutions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14408–14419.

[83] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, “VSA: Learning varied-size window attention in vision transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 466–483.

[84] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring plain vision transformer backbones for object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–296.

[85] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[86] J. Dai et al., “Deformable convolutional networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 764–773.

[87] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable convnets v2: More deformable, better results,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9300–9308.

[88] J. Lei Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.

[89] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[90] D. Hendrycks and K. Gimpel, “Gaussian error linear units (GELUs),” 2016, *arXiv:1606.08415*.

[91] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[92] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, “Oriented R-CNN for object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3500–3509.

[93] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[94] P. Helber, B. Bischke, A. Dengel, and D. Borth, “EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[95] 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.

[96] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, “In-domain representation learning for remote sensing,” 2019, *arXiv:1911.06721*.

[97] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3008–3017.

[98] J. Tian, J. Lei, J. Zhang, W. Xie, and Y. Li, “SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613213.

[99] J. Irvin et al., “USat: A unified self-supervised encoder for multi-sensor satellite imagery,” 2023, *arXiv:2312.02199*.

[100] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwar, S. Khan, and F. S. Khan, “Rethinking transformers pre-training for multi-spectral satellite imagery,” 2024, *arXiv:2403.05419*.

[101] D. Lam et al., “xView: Objects in context in overhead imagery,” 2018, *arXiv:1802.07856*.

[102] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[103] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, “Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608720.

[104] H. Zhang et al., “ResNeSt: Split-attention networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2022, pp. 2735–2745.

[105] G. Wang et al., “FSoD-Net: Full-scale object detection from optical remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602918.

[106] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[107] G. Cheng et al., “Anchor-free oriented proposal generator for object detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.

[108] A.-F. O. Detector, “FCOS: A simple and strong anchor-free object detector,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

[109] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9756–9765.

[110] X. Yang et al., “SCRDet: Towards more robust detection for small, cluttered and rotated objects,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8231–8240.

[111] Y. Xu et al., “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[112] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning RoI transformer for oriented object detection in aerial images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2844–2853.

[113] X. Yang, J. Yan, Z. Feng, and T. He, “R3Det: Refined single-stage detector with feature refinement for rotating object,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.

[114] L. Hou, K. Lu, J. Xue, and Y. Li, “Shape-adaptive selection and measurement for oriented object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 923–932.

[115] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, “AO2-DETR: Arbitrary-oriented object detection transformer,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2342–2356, May 2023.

[116] J. Han, J. Ding, J. Li, and G.-S. Xia, “Align deep features for oriented object detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[117] J. Han, J. Ding, N. Xue, and G.-S. Xia, “ReDet: A rotation-equivariant detector for aerial object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2785–2794.

[118] X. Yang et al., “Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18381–18394.

[119] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, “Rethinking rotated object detection with Gaussian Wasserstein distance loss,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.

[120] D. Liang et al., “Anchor retouching via model interaction for robust object detection in aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619213.

[121] G. Cheng et al., “Dual-aligned oriented detector,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111.

[122] X. Wang, G. Wang, Q. Dang, Y. Liu, X. Hu, and D. Yu, "PP-YOLOE-R: An efficient anchor-free rotated object detector," 2022, *arXiv:2211.02386*.

[123] S. Xu et al., "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.

[124] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1819–1828.

[125] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.

[126] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[127] T. Zhang et al., "Posterior instance injection detector for arbitrary-oriented object detection from optical remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5623918.

[128] X. Yang et al., "The KFIoU loss for rotated object detection," in *Proc. Int. Conf. Learn. Representations*, 2023.

[129] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[130] C. Xu et al., "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7318–7328.

[131] Z. Dong, Y. Gu, and T. Liu, "Generative ConvNet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5603816.

[132] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.

[133] Y. Pu et al., "Adaptive rotated convolution for rotated object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 6566–6577.

[134] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16748–16759.

[135] H. Yu, Y. Tian, Q. Ye, and Y. Liu, "Spatial transform decoupling for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 6782–6790.

[136] X. Zhang et al., "HiViT: A simpler and more efficient design of hierarchical vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2023.

[137] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[138] H. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 270–286.

[139] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[140] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[141] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4095–4104.

[142] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606216.

[143] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[144] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506105.

[145] L. Wang et al., "UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.

[146] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 1052–1064, 2023.

[147] Y. Chen, P. Fang, J. Yu, X. Zhong, X. Zhang, and T. Li, "Hi-ResNet: A high-resolution remote sensing network for semantic segmentation," 2023, *arXiv:2305.12691*.

[148] K. Yamazaki et al., "AerialFormer: Multi-resolution transformer for aerial image segmentation," 2023, *arXiv:2306.06842*.

[149] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[150] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805.

[151] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[152] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-Based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718.

[153] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 247–267, 2021.

[154] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[155] Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-supervised pretraining via multimodality images with transformer for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402711.

[156] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[157] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[158] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.

[159] H. Zhang, M. Lin, G. Yang, and L. Zhang, "ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 28–42, Jan. 2023.

[160] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[161] Y. Wen, X. Ma, X. Zhang, and M. Pun, "GCD-DDPM: A generative change detection model based on difference-feature guided DDPM," 2023, *arXiv:2306.03424*.

[162] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601816.

[163] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Remote sensing change detection using denoising diffusion probabilistic models," 2022, *arXiv:2206.11892*.

[164] H. Guo, B. Du, C. Wu, C. Han, and L. Zhang, "DeepCL: Deep change feature learning on remote sensing images in the metric space," 2023, *arXiv:2307.12208*.

[165] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[166] X. Li, L. Yan, Y. Zhang, and H. Zeng, "ESR-DMNet: Enhanced super-resolution-based dual-path metric change detection network for remote sensing images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5402415.

[167] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15173–15182.

[168] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.

[169] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "SAAN: Similarity-aware attention flow network for change detection with VHR remote sensing images," *IEEE Trans. Image Process.*, vol. 33, pp. 2599–2613, 2024.

[170] A. Mohammadian and F. Ghaderi, "SiamixFormer: A fully-transformer Siamese network with temporal fusion for accurate building detection and change detection in bi-temporal remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 12, pp. 3660–3678, 2023.

[171] Q. Li, R. Zhong, X. Du, and Y. Du, “TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[172] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, “RDP-Net: Region detail preserving network for change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635010.

[173] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, and B. Du, “An end-to-end supervised domain adaptation framework for cross-domain change detection,” *Pattern Recognit.*, vol. 132, 2022, Art. no. 108960.

[174] K. Li, Z. Li, and S. Fang, “Siamese nestedunet networks for change detection of high resolution satellite image,” in *Proc. 1st Int. Conf. Control Robot. Intell. Syst.*, 2021, pp. 42–48.

[175] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet : Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[176] Z. Zheng, S. Tian, A. Ma, L. Zhang, and Y. Zhong, “Scalable multi-temporal remote sensing change data generation via simulating stochastic change process,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21761–21770.

[177] C. Han, C. Wu, and B. Du, “HCGMNet: A hierarchical change guiding map network for change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5511–5514.

[178] F. I. Diakogiannis, F. Waldner, and P. Caccetta, “Looking for change? Roll the dice and demand attention,” *Remote Sens.*, vol. 13, no. 18, 2021.

[179] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, “Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, 2023.

[180] K. Chen et al., “Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection,” 2023, *arXiv:2312.16202*.

[181] S. Fang, K. Li, and Z. Li, “Changer: Feature interaction is what you need for change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111.

[182] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, “WNet: W-shaped hierarchical network for remote-sensing image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615814.

[183] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, “Vision transformer with deformable attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4784–4793.

[184] C. Han, C. Wu, M. Hu, J. Li, and H. Chen, “C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4702621.

[185] C. Zhao et al., “High-resolution remote sensing bitemporal image change detection based on feature interaction and multitask learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5511514.

[186] S. Zhao, X. Zhang, P. Xiao, and G. He, “Exchanging dual-encoder-decoder: A new strategy for change detection with semantic guidance and spatial localization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4508016.

[187] M. Lin, G. Yang, and H. Zhang, “Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images,” *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[188] S. Dong, L. Wang, B. Du, and X. Meng, “ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 208, pp. 53–69, 2024.

[189] K. Li, X. Cao, and D. Meng, “A new learning paradigm for foundation model-based remote-sensing change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610112.

[190] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, “Urban change detection for multispectral Earth observation using convolutional neural networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[191] S. Ji, S. Wei, and M. Lu, “Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery data set,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[192] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sens.*, vol. 12, no. 10, 2020.

[193] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. A. Knyaz, and A. Y. Rubis, “Change detection in remote sensing images using conditional adversarial networks,” *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, 2018.

Di Wang (Member, IEEE) received the B.E. degree in surveying and mapping from the China University of Petroleum (East China), Qingdao, China, in 2018, and the M.E. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2020. He is currently working toward the Ph.D. degree with the School of Computer Science, Wuhan University.

His research interests include deep learning, computer vision, and remote sensing.

Jing Zhang (Senior Member, IEEE) is currently a Research Fellow with the School of Computer Science, The University of Sydney, Camperdown, NSW, Australia. He regularly reviews for numerous prestigious journals and conferences. He has authored more than 80 papers in prestigious conferences and journals, including Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, NeurIPS, *International Conference on Learning Representations*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and *International Journal of Computer Vision*. His research interests include computer vision and deep learning.

Mr. Zhang is an Area Chair for ICPR, a Senior Program Committee Member for AAAI and IJCAI, and a Guest Editor for IEEE TRANSACTIONS ON BIG DATA.

Minqiang Xu received the Ph.D. degree in circuits and systems from University of Science and Technology of China, Hefei, China, in 2011.

He is currently the Principal Scientist of iFLYTEK Digital. During 2018–2010, he studied at the Image Formation and Processing Laboratory, University of Illinois at Urbana-Champaign (UIUC), where he was supervised under Prof. Thomas S. Huang, a pioneer scientist in the field of computer vision.

Dr. Xu was a recipient of the first place multiple times in international voiceprint recognition competitions including NIST SRE and VoxSRC. He participated in the development of the champion system (NEC-UIUC) for the ImageNet 2010 competition.

Lin Liu received the D.Eng. degree in electronic information from the University of Science and Technology of China, Hefei, China.

He is currently a Researcher with the National Engineering Research Center for Speech and Language Information Processing, University of Science and Technology of China. He has long been engaged in scientific research and industrialization in artificial intelligence technologies, such as speech recognition, remote sensing image recognition, and machine translation.

Dongsheng Wang received the B.E. degree from the North University of China, Taiyuan, China, in 2020, and the M.E. degree from Intelligent Image Analysis and Understanding Laboratory, Dalian University of Technology, Dalian, China, in 2023.

He is currently an AI Algorithm Engineer with iFLYTEK, Hefei, China. His research interests include computer vision and deep learning.

Erzhong Gao received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2006 and 2011, respectively.

He is currently a Technology Researcher with iFLYTEK, Hefei, China. His research interests include speech recognition, deep learning, computer vision, and remote sensing.

Chengxi Han (Student Member, IEEE) received the B.S. degree in remote sensing science and technology from the School of Geosciences and Info-Physics, Central South University, Changsha, China, in 2018, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

He was a Trainee with the United Nations Satellite Centre, United Nations Institute for Training and Research, Geneva, Switzerland. His research interests include deep learning and remote sensing image change detection.

Dr. Han has been the IEEE GRSS Wuhan Student Branch Chapter Chair since 2021. He was the recipient of the IEEE GRSS 2022 Student Chapter Excellence Award. Now, he has been a Student Chapter Coordinator in the IEEE GRSS Administrative Committee (AdCom) organization executive committee since January 2024.

Haonan Guo (Student Member, IEEE) received the bachelor's degree from Sun Yat-sen University, Guangzhou, China, in 2020. He is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include AI-empowered remote sensing image interpretation, including global mapping, dynamic monitoring, and disaster response.

Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010.

He is currently a Professor with the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University, where he is also the Director with the National Engineering Research Center for Multimedia Software. He has authored or coauthored more than 80 research articles in IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS), and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), and 13 of them are ESI hot papers or highly cited papers. His research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du was the recipient of the IJCAI Distinguished Paper Prize, the IEEE Data Fusion Contest Champion, and the IEEE Workshop on Hyperspectral Image and Signal Processing Best Paper Award in 2018. He was also a recipient of the Highly Cited Researcher 2019 by the Web of Science Group. He was the Area Chair for International Conference on Pattern Recognition (ICPR). He is a Senior PC Member of International Joint Conference on Artificial Intelligence (IJCAI) and Association for the Advancement of Artificial Intelligence. He is an Associate Editor for *Neural Networks*, *Pattern Recognition*, and *Neurocomputing*. He is a Reviewer for 20 Science Citation Index (SCI) magazines, including IEEE TPAMI, TCYB, TGRS, TIP, JSTARS, and GRSL.

Dacheng Tao (Fellow, IEEE) is currently a Distinguished University Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and more than 200 publications in prestigious journals and proceedings at leading conferences, with best paper awards, best student paper awards, and test-of-time awards. His publications have been cited over 112 K times and he has more than 160 h-index in Google Scholar.

Mr. Tao was the recipient of the 2015 and 2020 Australian Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a Fellow of the Australian Academy of Science, AAAS, and ACM.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently a "Chang-Jiang Scholar" Chair Professor with the Ministry of Education of China, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is a highly cited Author with the Institute for Scientific Information, Philadelphia, PA, USA. From 2011 to 2016, he was a Principal Scientist of the China State Key Basic Research Project appointed by the Ministry of National Science and Technology, Beijing, China, to lead the remote sensing program. He has authored or coauthored more than 700 research papers and five books. He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology (IET). He is also an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.