Brain-Inspired Remote Sensing Foundation Models and Open Problems: A Comprehensive Survey

Licheng Jiao, *Fellow, IEEE*, Zhongjian Huang, *Student Member, IEEE*, Xiaoqiang Lu, *Graduate Student Member, IEEE*, Xu Liu, *Member, IEEE*, Yuting Yang, *Graduate Student Member, IEEE*, Jiaxuan Zhao, *Graduate Student Member, IEEE*, Jinyue Zhang, Biao Hou, *Member, IEEE*, Shuyuan Yang, *Senior Member, IEEE*, Fang Liu, *Senior Member, IEEE*, Wenping Ma, *Senior Member, IEEE*, Lingling Li, *Senior Member, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, Puhua Chen, *Senior Member, IEEE*, Zhixi Feng, *Member, IEEE*, Xu Tang, *Senior Member, IEEE*, Yuwei Guo, *Senior Member, IEEE*, Dou Quan, *Member, IEEE*, Shuang Wang, *Senior Member, IEEE*, Weibin Li, Jing Bai, *Senior Member, IEEE*, Yangyang Li, *Senior Member, IEEE*, Ronghua Shang, *Senior Member, IEEE*, and Jie Feng, *Senior Member, IEEE*

***Abstract*—The foundation model (FM) has garnered significant attention for its remarkable transfer performance in downstream tasks. Typically, it undergoes task-agnostic pretraining on a large dataset and can be efficiently adapted to various downstream applications through fine-tuning. While FMs have been extensively explored in language and other domains, their potential in remote sensing has also begun to attract scholarly interest. However, comprehensive investigations and performance comparisons of these models on remote sensing tasks are currently lacking. In this survey, we provide essential background knowledge by introducing key technologies and recent developments in FMs. Subsequently, we explore essential downstream applications in remote sensing, covering classification, localization, and understanding. Our analysis encompasses over 30 FMs in both natural and remote sensing fields, and we conduct extensive experiments on more than 10 datasets, evaluating global feature representation, local feature representation, and target localization. Through quantitative assessments, we highlight the distinctions among various FMs and confirm that pretrained large-scale natural FMs can also deliver outstanding performance in remote sensing tasks. After that, we systematically presented a brain-inspired framework for remote sensing foundation models (RSFMs). We delve into the brain-inspired characteristics in this framework, including structure, perception, learning, and cognition. To conclude, we summarize 12 open problems in RSFMs, providing potential research directions. Our survey offers valuable insights into the burgeoning field of RSFMs and aims to foster further advancements in this exciting area.**

***Index Terms*—Brain modeling, deep learning, foundation model, image analysis, remote sensing.**

Manuscript received 1 August 2023; revised 25 August 2023; accepted 13 September 2023. Date of publication 18 September 2023; date of current version 14 November 2023. This work was supported in part by the Key Scientific Technological Innovation Research Project by the Ministry of Education, in part by the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61836009, in part by the National Natural Science Foundation of China under Grant U22B2054, Grant 62076192, Grant 62006177, Grant 61902298, Grant 61573267, Grant 61906150, and Grant 62276199, in part by 111 Project, in part by the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT 15R53, in part by the ST Innovation Project from the Chinese Ministry of Education, in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY03-06, in part by the National Science Basic Research Plan in Shaanxi Province of China under Grant 2022JQ-607, in part by China Postdoctoral Fund under Grant 2022T150506, in part by the Fundamental Research Funds for the Central Universities under Grant ZYTS23066, and in part by the Scientific Research Project of Education Department in Shaanxi Province of China under Grant 20JY023. *(Corresponding authors: Zhongjian Huang; Licheng Jiao.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: lchjiao@mail.xidian.edu.cn; huangzj@stu.xidian.edu.cn; luxiaoqiang5903@163.com; xuliu361@163.com; ytyang_1@stu.xidian.edu.cn; jiaxuanzhao@stu.xidian.edu.cn; jyzhang_2@stu.xidian.edu.cn; avcodec@163.com; syyang@xidian.edu.cn; f63liu@163.com; wpma@mail.xidian.edu.cn; llli@xidian.edu.cn; xrzhang@mail.xidian.edu.cn; phchen@xidian.edu.cn; zxfeng@xidian.edu.cn; tangxu128@gmail.com; yuweiguo18@126.com; dquan@stu.xidian.edu.cn; shwang.xd@gmail.com; weibinli@xidian.edu.cn; baijing@mail.xidian.edu.cn; yyli@xidian.edu.cn; rhshang@mail.xidian.edu.cn; jiefeng@xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3316302

I. Introduction

THE rapid advancements in data and model parameters have catalyzed the emergence of a new paradigm in artificial intelligence (AI) [1], [2], [3]. Through large-scale pretraining of neural networks, we witness the manifestation of novel characteristics, enabling a previously unprecedented level of understanding and reasoning [4]. The models trained on broad data can be adapted to a wide range of downstream tasks. These models are called foundation models (FMs) to underscore their critically central yet incomplete character [5].

Different from the nonFMs designed for a specific task or domain, FMs are a new paradigm that can be adapted to many different tasks and domains. As shown in Fig. 1, the main characteristics of FMs can be summarized into three aspects: data and model size, learning strategies, and adaptation.

1) *Data and model size:* FMs are trained on large amounts of unlabeled or weakly labeled data, such as text, images, audio, or video, that cover a broad range of topics and domains. For example, visual training datasets include ImageNet-22K [6] and JFT-300M [7], and multimodal datasets include Laion [8]. As for model size, flexible and

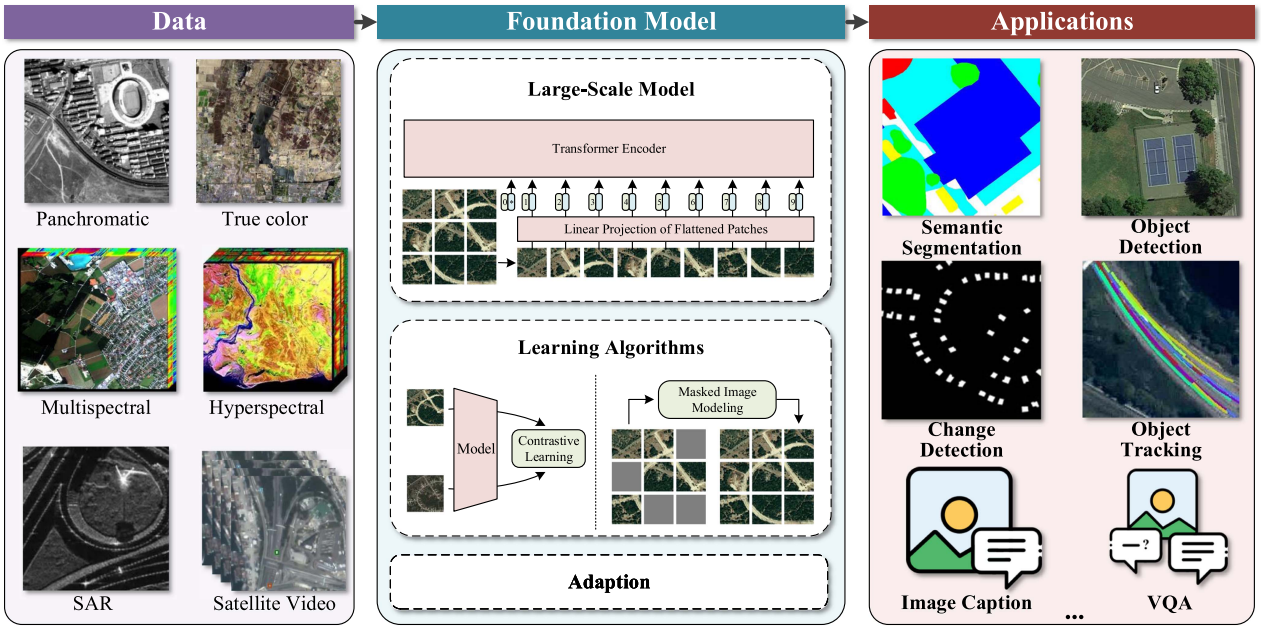


Fig. 1. Framework of the RSFMs. The models are trained with large-scale multimodality data and can be adapted to downstream applications.

expandable models, such as ViT, can be scaled up from ViT small with 50 million parameters to ViT-22B with 22 billion parameters [3].

2) *Learning strategies:* FMs use self-supervised or semisupervised learning to learn from the data without human supervision or with minimal human guidance. Numerous self-supervised algorithms are employed for pretraining, including contrast learning [9], generative masked image modeling [1], and multimodal contrast learning [10].
3) *Adaption:* FMs can be adapted or fine-tuned to various downstream tasks or domains by adding a small amount of task-specific data or parameters.

The interpretation of remote sensing (RS) is a crucial method for observing the Earth [11], [12], [13], and FMs have garnered significant attention and increasingly play a vital role in this domain [14], [15]. RS images, acquired through satellites, are generated at a substantial scale, reaching PB scale [16]. Due to the complexity of RS data and the requirement for professional knowledge, the labeled RS data are scarce. The pretraining approach of the FM can mine the value of RS data and enable the utilization of a significant amount of unlabeled data.

Inspired by the FMs developed for natural images, the field of RS has also seen the emergence of FMs, garnering attention [17], [18], [19], [20], [21], [22]. The typical remote sensing foundation model (RSFM) is pretrained using a substantial number of optical images, validating the feasibility of training FMs in the RS domain. In addition, scholars have considered factors, such as multispectral images, time-series images [20], and geographical resolution [22], to build more robust RS models.

The advancements in RSFMs have been impressive. However, there is still a noticeable gap between the scale of RS data and the models, especially when compared with natural FMs. Table I summarizes the basic information of FMs in both natural and RS domains, highlighting the disparity in dataset size and model parameters. RSFMs typically rely on data-driven approaches, training large-scale parameters from limited RS datasets, such as Million AID [40], which contains only 1 million images. In contrast, natural FMs benefit from much larger datasets, such as ImageNet-1K, containing millions of images.

Beside the scale of models, most RSFMs follow the paradigm of the nature of FMs. It has been demonstrated that natural FMs suffer from brittle, unchangeable structures. Model-based generation is prone to hallucinate unintended results [42]. These unstable results limit the application of the FMs in the field of RS, which requires high accuracy and robustness to guarantee security. To bridge this gap, brain-inspired RSFMs will be a new potential research direction [11]. Jiao et al. [43] have conducted systematic analyses of algorithms inspired by brain and biological mechanisms, including neural networks, natural computing, machine learning, and compression. Their work has provided valuable insights into how brain-inspired approaches can be applied to enhance the capabilities of FMs. Similarly, Schmidgall et al. [44] have explored the integration of more biologically plausible mechanisms into current brain-inspired learning representations, with the goal of further enhancing the capabilities of these networks. Zou et al. [45] focus on reviewing brain-inspired models with an emphasis on the spatiotemporal nature of visual signals.

In this article, we have drawn insights from brain characteristics to propose a brain-inspired framework for RSFMs. The exploration of brain-inspired algorithms in the context of RS holds great promise and offers exciting opportunities for future research and advancements in the field.

We investigate the progress of current RSFMs, as shown in Fig. 2. The rest of this article is organized as follows. In Section II, we describe the key technologies underlying these models, including the essential transformer structure of FMs and self-supervised pretraining methods. Furthermore, we introduce common methods for efficient parameter optimization, taking into account the application paradigm of the latest FMs.

TABLE I
SUMMARY OF EXISTING REPRESENTATIVE FMS IN NATURAL AND RS FIELDS

Field	Method	Modality	Visual encoder	Text encoder	Parameters	Training Dataset	Data Number
Natural	BYOL [23]	V	ResNet200	-	375M	IN1K	1.28M Images
	SimCLR v2 [24]	V	ResNet152	-	795M	IN1K	1.28M Images
	DINO [25]	V	ViT Base	-	85M	IN1K	1.28M Images
	MAE [1]	V	ViT Huge	-	632M	IN1K	1.28M Images
	SimMIM [26]	V	SwinV2 Large	-	197M	IN22K	14M Images
	Scaling ViT [2]	V	ViT Giant	-	1.8B	JFT-300M/3B	3B Images
	ViT 22B [3]	V	ViT 22B	-	22B	JFT-4B	4B Images
	CLIP [10]	VL	ViT Large	Transformer	307M/63M	WebImageText	400M ITPs
	ALBEF [27]	VL	ViT-B	BERT base	85.8M/123.7M	CC, SBU, COCO, VG	4M Images 14M ITPs
	CoCa [28]	VL	Transformer	Transformer	1B/1.1B	JFT-3B, ALIGN	
	BLIP-2 [29]	VL	ViT Giant	FlanT5-XXL	1.8B/11B	COCO, VG, CC3M, CC12M, SBU, subset of LAION400M	129M ITPs
	BEiT-v3 [30]	VL	Multiway Transformers	Multiway Transformers	1.9B	CC12M, CC3M, SBU, COCO, VG, IN21K, English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	21M ITPs, 14M Images, 160G documents
	SAM [31]	VL	ViT-Huge	-	636M	SA-1B	11M Images
RS	RSP [32]	V	Swin-Tiny ViTAEv2-S	-	27M	MillionAID	1M Images
	RVSA [17]	V	ViTAE-Base	-	89M	MillionAID	1M Images
	RingMo [18]	V	Swin-T Base	-	88M	self-collected	2M Images
	Geograph [33]	V	ResNet50	-	24M	fMoW, GeoImageNet	0.9M Images
	SatMAE [20]	V	ViT Large	-	307M	fMoW	0.7M Images
	Scale MAE [22]	V	ViT Large	-	307M	fMow	0.4M Images
	Billion [19]	V	ViT G12	-	2.4B	MillionAID	1M Images
	GFM [21]	V	Swin	-	80M	GeoPile	0.6M Images

NOTE: *V* and *VL* in *Modality* represent that the model is designed for vision and vision-language, respectively.
A single number shown in *Parameter* represent the overall parameter number of the model. The union like (307M/63M) means the model consist of a visual encoder with 307 million parameters and a text encoder with 63 millions.
The *ITP* shown in *Data Number* is the abbreviation of *Image-Text pair*.
The training dataset used for foundation models are ImageNet 1K (IN1K), ImageNet 22K (IN22K) [6], JFT [2], WebImageText [10], COCO [34], Conceptual 12M (CC12M) [35], Conceptual Captions (CC3M) [36], SBU Captions (SBU) [37], Visual Genome (VG) [38], ALIGN [39], LAION [8], Segment Anything 1B (SA-1B) [31], MillionAID [40], fMoW [41], GeoImageNet [21], GeoPile [21].
We summarize their modality, visual/text encoder, model parameters, training dataset, and data number.

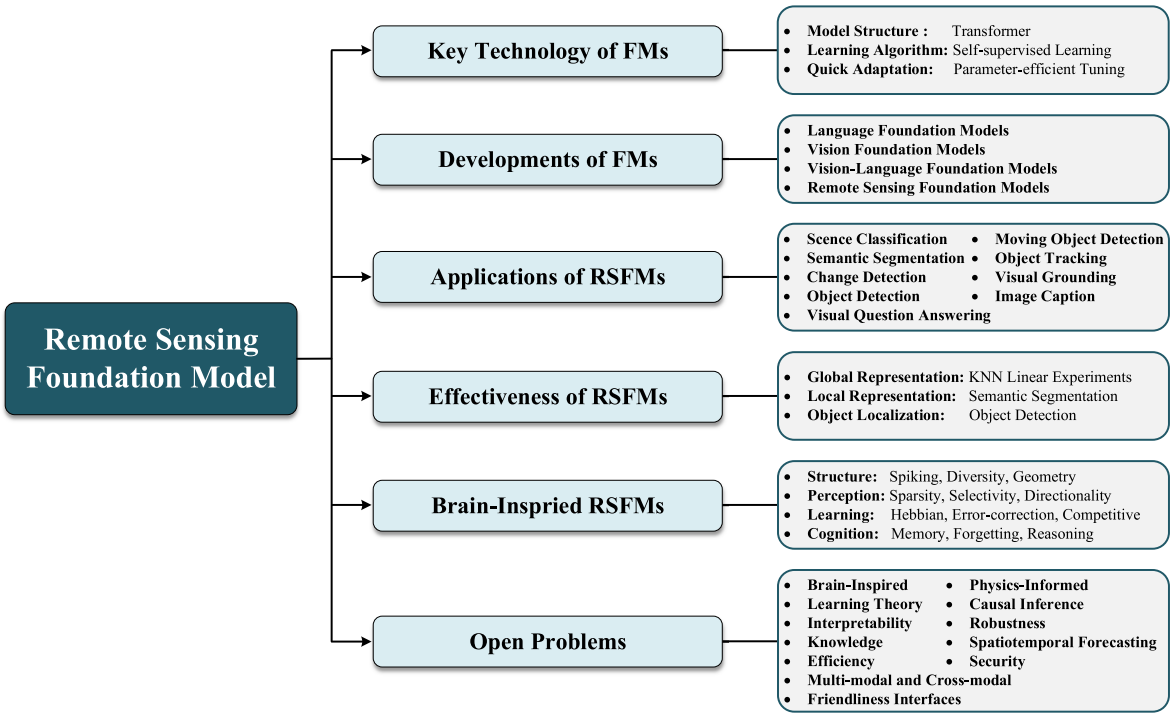


Fig. 2. Organizational structure of this survey.

Section III covers the latest developments in various FMs, including language FMs, visual FMs, visual-language FMs, and RSFMs. Section IV delves into several core applications of RS interpretation, focusing on classification, localization, and understanding tasks. To address the lack of systematic comparison between current RSFMs and natural FMs, experiments are conducted in RS interpretation from three perspectives: global representation, local representation, and object localization in Section V. These experiments provide a fair comparison of the proposed RSFMs. In Section VI, a framework of brain-inspired RSFM is proposed and the key characteristics of the brain are discussed. The 12 open problems of RS are discussed in Section VII. Finally, Section VIII concludes this article.

Our contributions can be summarized as follows.

1) We have comprehensively investigated the key technologies and latest advancements in FMs. This provides a comprehensive overview of FM research.
2) To the best of our knowledge, this is the first systematic summary and analysis of the performance of existing RSFMs compared with natural FMs. The experimental results can serve as a guide.
3) We propose a framework of brain-inspired RSFM and investigate the key characteristics of the brain. In addition, 12 open problems in the construction of RSFMs are discussed.

Overall, our work provides valuable insights into the FM landscape, offers performance comparisons, and highlights important characteristics and challenges in the realm of RSFMs.

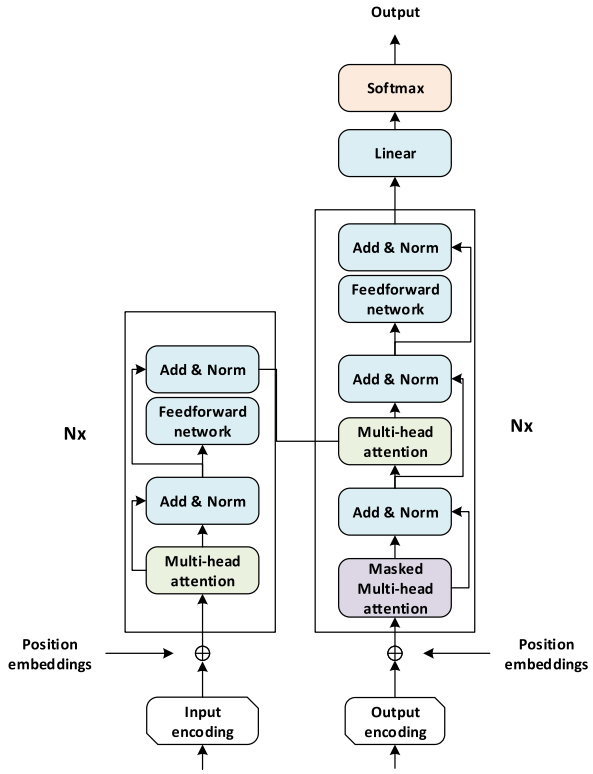


Fig. 3. Architecture of transformer.

II. Key Technology of FMs

The key technology of FMs consists of the model structures, learning algorithms, and fine-tuning. In this section, we first introduce the important structure, transformer. Then, the development of self-supervised learning (SSL) and parameter-efficient tuning are discussed.

A. Transformer

Transformer [46] is a neural network model based on a self-attention mechanism, which is often used in natural language processing tasks. Due to the nonlocality and the natural relationship of language, this long-term and self-attention behavior makes the transformer an effective tool [47].

The main idea of the transformer is to calculate the context-related representation through the self-attention mechanism. Convolutional neural networks (CNNs) of the traditional recurrent neural network (RNN) [48] have some difficulties in processing long sequence data when processing long sequence data [49]. The entire network structure of the transformer is composed of attention mechanisms, abandoning traditional CNN or RNN, and obtaining context information by calculating the correlation between each word and all other words, thereby avoiding the problem of traditional models [50].

The core component of the transformer includes a multihead self-attention mechanism and forward feedback network. In the multihead self-attention mechanism, the input text sequence will be split into multiple vectors. Then, a series of linear transformations, attention calculation, and concatenation operations will be followed to generate an output vector. This output vector contains the information in each position of the input sequence, and the information at each position is considered equally. Therefore, the transformer is more suitable for processing long text sequences compared with the circulating neural network [51], [52], [53].

1) Villian Transformer: The main structure is shown in Fig. 3. The core of the transformer consists of the encoder and decoder. It shows the overall encoder–decoder structure of the transformer model. The encoder converts the input sequence into the context vector, and the decoder uses the context vector to generate the output sequence. The encoder is mainly composed of two layers of the self-attention head and a two-layer feedforward neural network. There is also a decoder that has the self-attention layer and feedforward layer. In addition, there is also a self-attention layer between these two layers to pay attention to the relevant parts of the input sentence, which is similar to the attention of the Seq2Seq model [54].

Specifically, a transformer usually contains multiple continuous encoders. Each encoder consists of multiple layers. Each layer contains two sublayers: multihead attention and feedforward network. In the multihead self-attention mechanism, the input sequence is divided into multiple heads. Each head performs a self-attention calculation, and the attention weight weighted the input vector to obtain the relationship of each word with other words. And then, the output vector can be calculated.

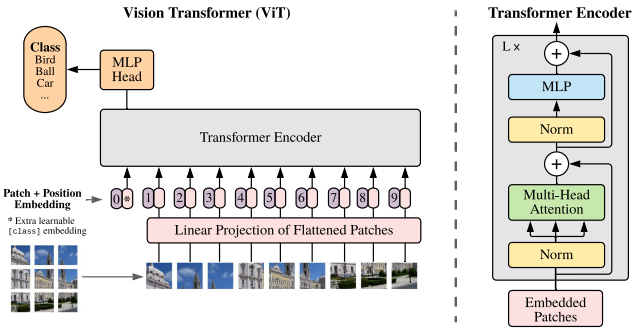


Fig. 4. Architecture of ViT.

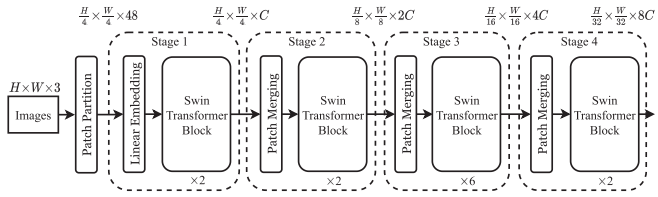


Fig. 5. Architecture of Swin transformer.

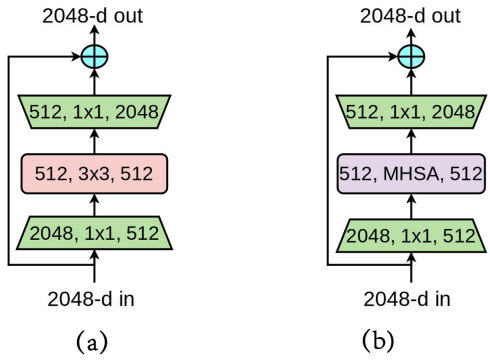


Fig. 6. Compare the architecture of ResNet bottleneck and bottleneck transformer. (a) ResNet bottleneck. (b) Bottleneck transformer.

The output vectors of all heads are merged into a longer vector through the concatenation operation. It contains information on each position in the input sequence, and the information at each position is considered equally. This is why transformer is more suitable for handling long text sequences. Multihead attention and feedforward networks use residual connections and layers among the sublayers. In the feedforward network, the two neural networks of the RELU activation function use the two layers of neural networks for nonlinear transformation.

Compared with the encoder, each decoder block is added with a multihead crossing-attention layer to embed the decoder into the encoder output. In addition, all sublayers in the encoder and decoder use the remaining connections and layers to improve the scalability of the transformer. To record the sequential information, each input embedding is attached to the start of the encoder and decoder stack with a position encoding. Finally, a linear layer and a softmax operation are adapted to predict the next word.

Compared with the traditional RNN, the transformer model can directly obtain global information. It is one of the advantages of high computing efficiency, parallel computing, and processing long text sequences. Therefore, it is widely used in NLP. In ChatGPT [55], [56], [57], transformer technology is used to generate text and text classification tasks. Its efficient computing power and accurate prediction results have been verified in practical applications.

2) Visual Transformer: Transformer has achieved great success in NLP. Subsequently, it was extended to computer vision and showed good performance in computer visual tasks, including image recognition, classification, segmentation, and so on. It has proven to be a simple and scalable framework. Compared with the traditional methods, it has obvious training efficiency advantages. It can use a pure transformer architecture or combined with CNN to achieve better results.

ViT: The overall framework of ViT [58] is shown in Fig. 4. First of all, the image is divided into 16×16 patches and then the flatted patches of the flatted linear mapping. The obtained patch and position encoding are sent to the transformer encoder for encoding. Finally, send the encoded features into the MLP head for classification. Among them, the transformer encoder is mainly the position encoder structure proposed in the transformer. The appearance of ViT is a preliminary attempt by transformer in computer visual tasks [59]. It is convolution-free and highly recognized by researchers for its excellent long-distance modeling capabilities.

Swin Transformer: Swin transformer [60] adapts the hierarchical construction method, similar to the CNNs. The sizes of the feature maps decreased with the feature layer deepened. ViTs feature maps' sizes are unchanged, with 16 times downsampling. Unlike ViT, its feature maps change 4 times, 8 times, and 16 times downsampling. In detail, its overall framework is figured, as shown in Fig. 5(a). It comprises patch partition, linear embedding, patch merging, and Swin transformer blocks. The overall architecture of two successive Swin-transformer blocks is shown in Fig. 5(b). The hierarchical characteristics of the Swin transformer have an essential role in visual recognition.

BotNet: BotNet [61] is a simple but effective backbone for visual representation. It introduces self-attention to many visual tasks, including image classification, object detection, and instance segmentation. BotNet consists of Bottleneck transformer blocks. In detail, its framework is shown in Fig. 6. For 2048-dimensional (2048-D) input, the ResNet bottleneck contains the convolutional operations, including $1 \times 1 \times 512$, $3 \times 3 \times 512$, and $1 \times 1 \times 2048$. The skipping connection of ResNet is still maintained. Compared with the ResNet bottleneck, the bottleneck transformer only replaces the original second convolutional operation with multihead attention (MHSA), as shown in Fig. 6. In the bottleneck transformer, MHSA is the central core novelty. It enables the model to capture different characteristics and modes in the input data. In addition, BotNet replaces the spatial convolutions in the last three bottleneck blocks of ResNet with global self-attention. It has significantly improved the baseline regarding instance segmentation and target detection and reduces the parameters to minimize delay.

3) Advantages and Disadvantages: NLP was a latecomer over the past ten years of a deep learning revolution. Anna Rumshisky, a computer scientist at the University of Massachusetts, said that NLP [62], [63] was behind the computer in a sense. Vision transformer (ViT) breaks the restrictions of incompetence computing in the RNN model. Note that the mechanism provides context information for any location in the input sequence. It is one of the advantages of parallelism, unlimited positioning operations, strong global characteristics, strong versatility, and strong scalability so that the generative pretrained (GPT) model [64], [65] has excellent performance. Specifically, the advantages are listed as follows.

1) Design innovation: It abandoned the most fundamental RNN or CNN in NLP and achieved excellent results. The design of it is very inspiring and worthy of in-depth research.
2) The key to transformer's design is that the distance between any two words is 1, which is effective for solving the difficult long-term dependencies in NLP.
3) Transformer cannot only be applied to machine translation in NLP but also not even limited to the NLP field. It is a direction of very scientific research potential.
4) The parallelism of the algorithm is good, which is in line with the current hardware environment.

Of course, its model still has some limitations such as follows.

1) Although the rough abandonment of RNN and CNN is very dazzling, it also causes the model to lose the ability to capture local characteristics. The combination of RNN, CNN, and transformer may bring better results [66].
2) The lost location information that transformer is important in NLP and adding position embedding to the feature vector is just a suitable measure, and it does not change the inherent defects in the transformer structure.
3) Although transformer helps integrate and improve AI tools [67], as with other emerging technologies, transformer also has expensive costs. A transformer model requires a lot of computing power during the pretraining stage to defeat the previous competitors [47], [68].
4) From the perspective of the transformer, there are problems of large memory occupation and high delay in architecture based on the transformer, which hinders their efficient deployment and reasoning. Recently, many studies have improved computing and memory efficiency around the original transformer architecture, but most of them are concentrated in the semisupervised field [25], [69].

B. Self-Supervised Learning

SSL plays a crucial role in training FMs. Many state-of-the-art FMs utilize SSL in the pretraining phase. This pretraining phase allows the FM to acquire rich features and representations, and then use the labeled data to fine-tune for specific downstream tasks. SSL is a form of unsupervised learning that aims to extract useful and generalizable feature representations from a large amount of unlabeled data for downstream tasks [70], [71], [72]. Referred to as the "dark matter" of intelligence, SSL differs from supervised learning, which is constrained by the availability of labeled data. Instead, self-supervised methods leverage a "semiautomatic" process to obtain "labels" directly from the data itself, saving significant manpower and time costs [71]. In recent years, SSL has achieved remarkable success in the field of deep learning, particularly in natural language processing, with the emergence of influential language models, such as BERT [4] and GPT-3 [73]. In computer vision, models, such as MAE [1] and DINOv2 [74], have been able to match or even surpass supervised models in certain scenarios. The general workflow of SSL in computer vision is illustrated in Fig. 7. SSL defines a pretext task based on unlabeled inputs to generate descriptive and interpretable representations [70], [72], [75], [76]. The pretext task is a predesigned task in the pretraining phase, where the objective function is learned by inputting unlabeled data. Typically, pretext tasks can be prediction-based, context-based, or generation-based, and the supervision signal is generated from the data itself [75]. After training on the pretext task, the learned representations are transferred as initial weights to downstream tasks to achieve their intended objectives.

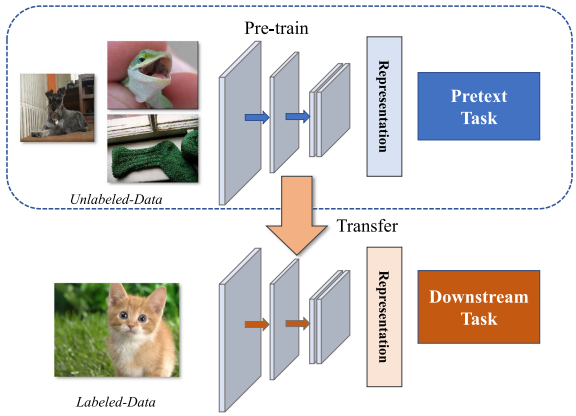


Fig. 7. General approach of SSL. First, an auxiliary task is trained using an unlabeled dataset to apply the SSL scheme. Subsequently, the learned network weights are transferred from the pretext task to the downstream task, enabling training on a small amount of data with labels.

1) SSL for Natural Images: Based on the different pretext task approaches, three different SSL methods can be identified: generative, contrastive, and predictive [70], [77], as shown in Fig. 8.

Generative Methods: Generative methods aim to learn representations by reconstructing or generating input data. The basic idea is to model the underlying data distribution to capture the statistical properties and dependencies of the input data. Generative methods can implicitly capture meaningful features and structures in the data without relying on explicit labels. These methods often utilize generative models, such as autoencoders and generative adversarial networks (GANs) [78], to perform reconstruction or generation tasks.

An autoencoder consists of an encoder network that maps input data to a latent space representation and a decoder network that reconstructs the data from the latent space. Based on the autoencoder, several variant methods have emerged. For example, variational autoencoder [79] combines the

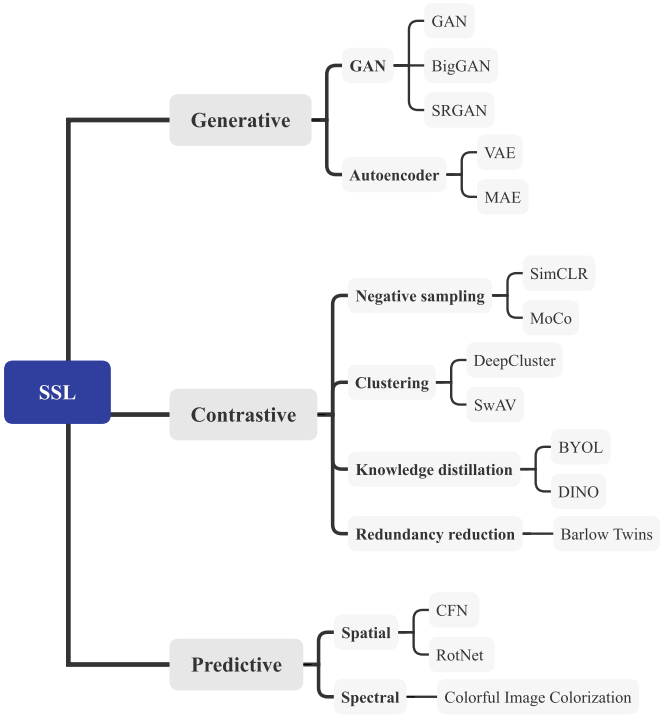


Fig. 8. Summary of popular self-supervision methods and typical models.

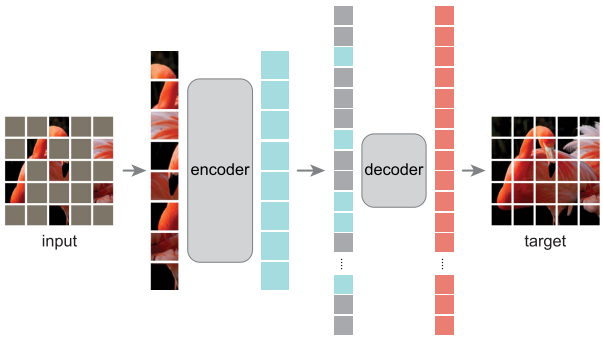


Fig. 9. Training procedure of the MAE (Image from [1].).

encoder–decoder structure of an autoencoder with probabilistic modeling. It assumes the existence of a prior distribution $P(Z)$ over the latent space and models the conditional distribution $P(X|Z)$. The encoder approximates the posterior distribution $P(Z|X)$ by inferring the model $Q(Z|X)$. Recently, generative methods have been commonly used for information recovery tasks, such as inpainting, where a portion of an image is removed, and the network's context encoder is trained to restore the missing pixel values based on the surrounding context [80]. This idea has evolved into the masked autoencoders (MAEs) task [1], where random masks are applied to input image patches, and the missing pixels are subsequently reconstructed, as shown in Fig. 9. In order to correctly reconstruct each pixel, the model needs to understand the different objects and relevant components present in the image. Therefore, the learned feature representations are useful for other downstream tasks.

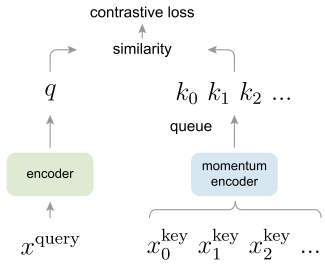


Fig. 10. Training procedure of the MoCo (Image from [84].).

GAN [78] consists of two networks: a generator $G : ZX$ and a discriminator $D : X[0, 1]$. The generator synthesizes fake samples from random noise in the latent space, while the discriminator attempts to distinguish between real and fake samples. These methods are primarily used for image generation and image super-resolution tasks. For example, BigGAN [81] introduces innovations, such as a large-scale GAN architecture, conditional batch normalization, category conditioning, and orthogonal regularization. These advancements enable BigGAN to generate high-quality, diverse, and category-specific images. SRGAN [82] is capable of recovering high-resolution textures in super-resolution tasks by learning from a large set of downsampled images.

Contrastive Learning Methods: Contrastive learning methods aim to learn representations by maximizing the similarity between semantically related samples and minimizing the similarity between unrelated samples. The key idea of contrastive learning is to create a contrastive objective function that encourages the model to bring similar samples closer together in the feature space and separate dissimilar samples. Classical contrastive learning methods can be categorized into four types: negative sampling, clustering, knowledge distillation, and redundancy reduction.

The negative sampling method involves creating a set of negative samples dissimilar to the anchor samples. The objective is to encourage the model to differentiate and separate positive and negative examples in the learned representation space, thereby learning valuable feature representations and avoiding model collapse. In SSL with negative sampling, models, such as SimCLR [83] and the MoCo series, are typical approaches. SimCLR uses a CNN as a feature extractor and employs various data augmentation strategies to increase data diversity. As shown in Fig. 10, the MoCo model [84] uses a momentum-based update strategy to update model parameters and performs contrastive learning by comparing a set of samples with a larger queue of negative samples. MoCo and its variants, such as MoCov2 [85] and MoCov3 [9], have demonstrated strong performance in SSL tasks, proving the effectiveness of negative sampling in training robust and discriminative representations.

The clustering approach utilizes clustering algorithms to group similar samples in an unsupervised manner. One representative method is DeepCluster, which first clusters images into different clusters and then trains a CNN to recognize assignments [86]. Another example is the SwAV [87] model, which introduces the concept of grouping similar representations into

a cluster. It generates multiple views of an image using data augmentation and then groups these views into clusters based on their similarity. SwAV encourages similar representations within each cluster and makes representations from different clusters dissimilar.

The knowledge distillation method refers to the process of transferring knowledge from a teacher model to a student model. The teacher model is usually a pretrained model with high performance, while the student model is trained using SSL. In SSL, based on knowledge distillation, several notable models have been proposed. One classic method is BYOL [23] and DINO (DINO [25] and DINOV2 [74]) series. BYOL focuses on predicting the representation of one augmented view from another view. It uses an online network as the student network and a target network as the teacher network, updating the target network using a momentum-based update strategy. The DINO model combines the transformer architecture and learns representations by maximizing consistency between two views of the same image, where one view is used as a query and the other view is used as a key to predict the output of the teacher network by the student network. Building upon DINO, DINOv2 further incorporates a clustering objective to encourage samples that are semantically similar to cluster together in the learned representation. This helps the model capture finer grained and meaningful structures in the learned representation.

Redundancy reduction refers to the process of reducing redundant information in learned representations. The goal is to ensure that the representations capture the essential and discriminative features of the data. For example, the Barlow twins model [88] maximizes the cross-correlation matrix of the representations while minimizing its diagonal elements. This encourages the representations to capture statistical dependencies between different parts of the input data while reducing redundancy. By maximizing cross correlation, the model learns to encode useful information across different views.

Predictive Methods: Predictive methods aim to learn useful representations by training models to perform prediction tasks. Popular image prediction tasks involve methods based on both spatial and spectral aspects of the image. In the spatial-based approach, the image prediction task involves predicting the relative positions of two patches from the same image or identifying the random order of a sequence of patches from the same image. The former trains CNNs to predict the relative positions of two randomly sampled patches in an image. The latter constructs image puzzles by decomposing the image into a series of nonoverlapping patches and predicts the relative positions of each patch to reconstruct the image. Image puzzle tasks require learning how parts are assembled within an object, the relative positions of different parts, and the shape of the object. Therefore, these representations are useful for downstream classification and detection tasks. For example, Noroozi and Favaro [89] proposed CFN, where a neural network is trained to solve the Jigsaw puzzle, learning both feature mappings of object parts and their correct spatial arrangements. Geometric transformation recognition tasks are used to identify the rotation angle of the entire image. This task requires the network to learn to locate salient objects in the image, recognize their orientations and object types, and then associate object orientations with the original image. RotNet, proposed by Gidaris et al. [90], performs unsupervised representation learning by predicting image rotations. Counting tasks aim to train models to count visual primitives in an image and learn the representation of the image by outputting the number of objects in the image, effectively learning spatial and object information in the image. However, spectral-based methods aim to automatically add realistic colors to grayscale images, which are referred to as the image coloring (IC) task. For example, CNN is used for IC followed by classification, detection, and segmentation downstream task validation in [91].

2) SSL for RS Images: SSL in RS is a method that utilizes unlabeled information in RS data to learn useful representations. In the field of RS, SSL has been widely applied to multispectral imagery, hyperspectral imagery, and synthetic aperture radar (SAR) imagery. Similar to the domain of natural images, RS SSL can also be categorized into three different approaches: generative, contrastive, and predictive.

Generative Methods: Generative methods in RS imagery often rely on techniques, such as autoencoders and GANs. However, generative methods in RS involve additional tasks. For instance, tasks, such as urban flood mapping and hyperspectral unmixing, are included. Specifically, Peng et al. [92] proposed the SSL framework for patch-based urban flood mapping using multitemporal multispectral satellite imagery. They utilized patch-level change vector analysis with features learned by a self-supervised autoencoder to generate patch-level change maps highlighting potential flood-affected areas. Jin et al. [93] introduced AAENet, a novel technique network for unsupervised hyperspectral unmixing. The proposed approach improved model performance and robustness by incorporating untied-weighted autoencoder, discrimination network, adversarial processes, and adding abundance priors to the framework.

Contrastive Methods: In contrastive methods, negative sampling remains a popular approach. For example, Hou et al. [94] proposed a contrastive learning-based algorithm for hyperspectral image classification, which consists of a pretraining phase and a fine-tuning phase. In the first phase, the model is pretrained by constructing positive and negative sample pairs to learn to discriminate between them. In the second phase, based on the pretrained model, features are extracted from hyperspectral images for classification, and a small number of labeled samples are used for fine-tuning the features. Similarly, Scheibenreif et al. [95] adopt a two-stage approach, where the model is trained to predict whether two image patches come from the same image. Swin transformer is combined with SSL for land-cover classification and segmentation. In clustering-based methods, contrastive learning-based dual dynamic graph convolutional network (GCN) for SAR image scene classification proposes a clustering-based contrastive learning approach using dual dynamic GCN for SAR image scene classification. The proposed clustering-based contrastive SSL model is used to transform SAR images into a higher level embedding space as richer representations without any labels, aiding subsequent node representations and information propagation in GCN.

Knowledge distillation methods are often inspired by models in the natural image domain. For instance, Muhtar et al. [96] proposed a method called IndexNet for semantic segmentation. The proposed IndexNet consists of two branches: the index contrastive branch and the instance contrastive branch. The index contrastive branch learns pixel-level representations by tracking object positions and maintaining sensitivity to changes in object positions to ensure consistency. The instance contrastive branch follows the standard BYOL learning process, learning image-level representations by combining image-level and pixel-level contrastive learning to capture spatiotemporal-invariant features.

Predictive Methods: RS prediction methods can also utilize other downstream tasks by constructing a pretext task based on rotation prediction. For example, Ji et al. [97] used rotation prediction to identify the input's 2-D rotation to guide the learning of transferable knowledge across categories. They combined contrastive learning to bring positive sample pairs closer and push away negative sample pairs, promoting intraclass consistency and interclass inconsistency. These pretraining tasks are jointly optimized in an end-to-end manner with semantic category prediction tasks, ultimately achieving RS image scene classification. In addition, in [98], IC is used as a pretexting task to learn feature representations, which are then transferred to a U-Net model for predicting the semantic segmentation of remote sense urban scenes.

C. Parameter-Efficient Tuning for FMs

Fine-tuning is a crucial method for applying pretrained models to downstream tasks. However, it involves updating parameters for both the entire model and each task model. Fine-tuning a large FM poses significant challenges in terms of computing resources and storage. To address this, the technology of parameter-efficient fine-tuning (PEFT) has been explored and implemented. The primary objective of PEFT is to enhance the performance of pretrained models on new tasks by minimizing the number of fine-tuning parameters and reducing computational complexity. This, in turn, mitigates the training cost associated with large pretrained models. In most cases, PEFT only requires the addition or updating of a small number of parameters in the model to facilitate its application on downstream tasks. Remarkably, these techniques achieve comparable accuracy compared with full fine-tuning. In this section, we introduce various PEFT strategies, such as prompt tuning, adapter tuning, and low-rank adapters (LoRA). For more parameter-efficient tuning strategies, interested readers can refer to the literature [99].

1) Prompt Tuning [100]: Prompting refers to constructing a language instruction to the input text of LLM so that the LLM can solve the downstream tasks without fine-tuning the whole model [73]. To construct better prompting texts, prompt tuning treats the prompts as task-specific continuous vectors and directly optimizes them via gradients during fine-tuning [100]. Prompt tuning only adds trainable vectors to the input embedding layer, initializing them with text. This approach allows fine-tuning with smaller learning parameters and offers higher computational efficiency. Despite fine-tuning fewer model parameters, Prompt tuning achieves accuracy comparable to full fine-tuning.

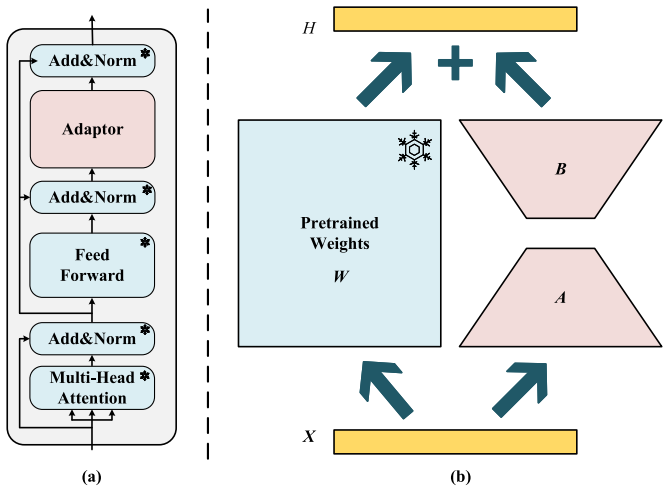


Fig. 11. Illustration of the parameter-efficient tuning algorithm, including adapter and LoRA. (a) Simple example of the adapter with transformer block. (b) Illustration of the LoRA algorithm.

Prefix tuning [101] is a kind of prompt tuning method. It uses a series of "virtual tokens" to create a prefix of tokens, providing "implicit" hints to the model. As a result, the parameters of the large language model remain frozen, and for each layer's input, a set of continuous task-related prefix tokens is learned to prompt the model. To ensure stable training, an MLP layer is added to reparameterize these prefix tokens. Once training is complete, only the parameters of the prefix layer need to be saved to obtain a fine-tuned model for a specific task.

Inspired by the prompt tuning, some research articles try to apply these efficient tuning methods to the visual domain. Visual prompt tuning (VPT) [102] prepends a set of learnable parameters to the pretrained ViT and conducts experiments on a wide variety of downstream recognition tasks. It shows that VPT achieves significant performance gains compared with other parameter-efficient tuning protocols. Multimodal prompt learning [103] learns prompts on both text and vision branches to ensure mutual synergy. In addition, branch-aware hierarchical prompts are also designed to progressively model the stagewise feature relationships. Oh et al. [104] proposed the black-box visual prompting, which efficiently adapts the large-scale pretrained models without knowledge about model architectures and parameters.

2) Adapter Tuning [105], [106]: Adapter tuning is a more general fine-tuning strategy designed for large models. It involves adding an adapter module within each layer or between layers while keeping the main parameters of the pretrained model fixed, as shown in Fig. 11(a). During the fine-tuning process, only the parameters in the adapter are trained to adapt to downstream tasks, reducing the computational overhead of training. The advantage of adapter tuning lies in its ability to retain the model's general knowledge while learning specific

knowledge for downstream tasks, avoiding catastrophic forgetting and task interference. This approach is now widely used in various applications.

3) LoRA [107]: LoRA is a well-known technology for fine-tuning the model. Fine-tuning typically involves modifying all parameters in the model, but today's large language models with billion-scale parameters make conventional fine-tuning strategies computationally expensive. Previous research has shown that neural networks are overparameterized, and their parameters can be represented separately using low-rank representation. Unlike fine-tuning all parameters, the LoRA algorithm suggests that we can learn a low-rank weight residual parameter, denoted as ΔW, to achieve the fine-tuning effect.

As shown in Fig. 11(b), for a weight W, rather than adjusting all parameters fully, we only need to learn the residual ΔW of this parameter. Moreover, this residual can be decomposed into two low-rank matrices, $\Delta W = AB$. As a result, we only need to fine-tune matrices A and B to achieve the fine-tuning effect, and the performance after fine-tuning is equivalent to that of full fine-tuning.

Due to its ease of use and effectiveness, LoRA has prompted the development of various methods to improve it, enabling fine-tuning models with more parameters, even on smaller computers. For example, AdaLoRA [108] considers the importance of parameters, adaptively allocates the budget for parameter optimization, and parameterizes incremental updates in the form of singular value decomposition. QLoRA [109] enhances Vallian LoRA from a quantitative perspective. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into LoRA. This enables fine-tuning of 65B parameter models on a single 48G GPU while maintaining comparable performance with 16-bit fine-tuned models.

III. Development of FMs

A. Language FM

Language FMs, also known as large language models, have gained significant attention in recent years. These models utilize a vast amount of text for unsupervised training and excel in text representation and understanding. Some of the notable language FMs are summarized in Table II.

One of the pioneering models is deep bidirectional transformers (BERT) [4], which achieve deep representation pretraining by joint conditioning on unlabeled samples in a bidirectional manner. BERT achieves state-of-the-art results on seven NLP tasks. Raffel et al. [110] proposed a method to convert all text-based language problems into a text-to-text format and trained a T5 model with 11B parameters. Subsequently, the well-known GPT transformer series models were introduced. GPT-3 [73], in particular, demonstrated the potential of large-scale parametric language models and inspired a wide range of applications. For instance, WebGPT [111] implemented question answering in a web browser environment, while Codex [112] performed fine-tuning on the basis of GPT to enable the model to master Python code-writing capabilities. Following these developments, several language FMs with large-scale parameters were proposed [113], [114], [115], [116], [117], showcasing impressive performance in language translation, summarization, question answering, and text completion. These models have displayed unprecedented capabilities [117].

Moreover, advancements, such as InstructGPT [118], have further improved the control and flexibility of large language models, ensuring the logic and values of the models align with human understanding. This has opened up new possibilities for utilizing language FMs in a more human-like manner.

B. Vision FM

Drawing inspiration from the construction of language FMs, the field of computer vision has also delved into large-parameter FMs. Table II summarizes some representative vision FMs.

The exploration of vision FMs can be categorized into three main aspects: training methods, parameter number, and tasks. BYOL [23] enables SSL by interacting between two networks. SimCLR [24] proposes a semisupervised learning method that combines unsupervised, distillation, and few-shot supervision to enhance the model's capabilities. SimMIM [26] simplifies the training process of MAE while maintaining accuracy.

Researchers have also explored breakthroughs in the number of parameters in visual models. Leveraging the scalability of the ViT model, Zhai et al. [2] scaled the ViT model to 1.8 billion parameters. Taking it a step further, Google proposed a model [3] with 22 billion parameters, demonstrating the visual scaling potential akin to large language models. In addition, InternImage [119] implements a large-scale CNN FM, achieving performance improvements similar to ViT.

While most of these models are focused on image-based tasks, the extension of natural image FMs to video domains has also been explored [120], [121], [122]. These advancements have opened up new possibilities for utilizing large-parameter FMs in various visual tasks, both for images and videos.

C. Vision–Language FM

In the current landscape, FMs have evolved to encompass more than just deep models; they now focus on utilizing vast amounts of data and computational power to tackle diverse problems. The goal is to use a unified model capable of addressing multiple modalities and tasks. Consequently, there is a growing emphasis on training vision–language models [123]. Table II summarizes some of the representative vision–language FMs.

CLIP [10] is a prominent example that leverages a large-scale collection of image–text pairs from the Internet for contrastive learning, enabling the creation of a unified representation of multimodal data. Inspired by CLIP, various multimodal image FMs have been proposed, differing in their model structures [30], [124], feature representations [125], [126], multimodal feature fusion approaches [27], feature alignment loss functions [28], pretraining methods [29], and more. These advancements have significantly improved the performance of multimodal FMs.

In addition, DALL-E [127] combines the diffusion model with multimodal FMs to generate images from text. SAM [31] introduces a promptable model with training strategies that enable the segmentation of objects using text, points, and lines as prompts. GPT-4 [128] exhibits superhuman capabilities on

TABLE II
SUMMERY OF LANGUAGE, VISION, AND VISION–LANGUAGE FMS

Model name	Publish time	Parameter number	Contribution
Language foundation models			
BERT [4]	2018	336M	Propose a transformer-based language representation model using a bidirectional approach to train on large amounts of text data.
T5 [111]	2019	11B	Convert all text-based language problems into a text-to-text format, allowing us to use the same model, loss function, and hyperparameters on any NLP task.
GPT-3 [73]	2020	175B	Demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance.
WebGPT [112]	2021	175B	Fine-tune GPT-3 to answer long-form questions using a text-based web-browsing environment.
Codex [113]	2021	175B	Propose a GPT language model fine-tuned on publicly available code from GitHub, and study its Python code-writing capabilities.
BLOOM [116]	2022	176B	Propose a language model capable of generating text in 46 natural languages and 13 programming languages.
GLM [115]	2022	130B	Propose a bilingual (English and Chinese) pretrained language model with 130 billion parameters.
Flan-T5 [117]	2022	11B	Explore instruction fine-tuning with a particular focus on scaling the number of tasks, scaling the model size, and finetuning on chain-of-thought data.
InstructGPT [119]	2022	1.3B	Propose an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback.
PaLM [118]	2022	540B	Further explore the impact of scale and train a 540-billion parameter, densely activated, transformer language model.
LLaMA [114]	2023	65B	Release a collection of foundation language models ranging from 7B to 65B parameters publicly.
Vision foundation models			
BYOL [23]	2020	375M	Propose an approach to self-supervised image representation learning through the interaction and learn between online and target network.
SimCLR v2 [24]	2020	795M	Propose a semisupervised learning algorithm contraining unsupervised pretraining of a big model, supervised fine-tuning on a few labeled examples, and distillation with unlabeled examples for refining and transferring the task-specific knowledge.
DINO [25]	2021	85M	Show the potential of self-supervised pretraining a standard ViT model, achieving performance that are comparable with the best convnets specifically designed.
MAE [1]	2022	632M	Propose a MAEs, which are scalable self-supervised learners for computer vision.
SimMIM [26]	2022	197M	Propose a simplified framework for masked image modeling without the need for special designs, such as block-wise masking and tokenization via discrete VAE or clustering.
Scaling ViT [2]	2022	1.8B	Demonstrate that the performance-compute frontier for ViT models with enough training data roughly follows a (saturating) power law.
InternImage [120]	2023	1.08B	Propose a new large-scale CNN-based FM, which can obtain the gain from increasing parameters and training data like ViTs
ViT 22B [3]	2023	22B	Present a recipe for highly efficient and stable training of a 22B-parameter ViT.
VideoSwin [121]	2022	200M	Instead an inductive bias of locality in video transformers based on the Swin transfomrmer.
BEVT [122]	2022	88M	Conduct masked image modeling on image data and masked video modeling on video data jointly.
VideoMAE [123]	2023	2B	Propose a scalable and general self-supervised pretrainer for building video FM.
Vision- language foundation models			
CLIP [10]	2021	370M	Demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet.
ALBEF [27]	2021	209.5M	Propose a contrastive loss to align the image and text representations before fusing them through cross-modal attention.
Florence [126]	2021	893M	Propose a new computer vision FM to expand the representations from coarse (scene) to fine (object), from static (images) to dynamic (videos), and from RGB to multiple modalities.
CoCa [28]	2022	2.1B	Propose a minimalist design to pretrain an image–text encoder–decoder FM jointly with contrastive loss and captioning loss.
DALL-E 2 [128]	2022	6.5B	A diffusion model can generate images from text descriptions based on the embedding space of CLIP.
Flamingo [125]	2022	80B	Propose an architecture to bridge powerful pretrained vision-only and language-only models, handle sequences of arbitrarily interleaved visual and textual data, and seamlessly ingest images or videos as inputs.
InternVideo [127]	2022	-	Explores masked video modeling and video-language contrastive learning as the pretraining objectives, and selectively coordinates video representations of these two complementary frameworks in a learnable manner.
BLIP-2 [29]	2023	12.8B	Propose a generic and efficient pretraining strategy that bootstraps vision–language pretraining from off-the-shelf frozen pretrained image encoders and frozen large language models.
BEiT-v3 [30]	2023	1.9B	Propose a multiway transformers for general-purpose modeling, where the modular architecture enables both deep fusion and modality-specific encoding.
PaLI [130]	2023	17B	Scale up a model to the joint modeling of language and vision with flexible task interface.
SAM [31]	2023	636M	Propose a promptable model, which can transfer zero shot to new image distributions and tasks.
GPT-4 [129]	2023	-	Propose a large-scale, multimodal model that exhibits human-level performance on various professional and academic benchmarks.
mPLUG-2 [131]	2023	-	Introduce a multimodule composition network by sharing common universal modules for modality collaboration and disentangling different modality modules to deal with modality entanglement.
Metatransformer [132]	2023	-	Propose a unified framework that can handle a wide range of tasks, including fundamental perception (text, image, point cloud, audio, video), practical application (X-Ray, infrared, hyperspectral, and IMU), and data mining (graph, tabular, and time-series).

TABLE III
SUMMERY OF RSFMS

Model name	Publish time	Parameter number	Contribution
RSP [32]	2022	27M	Conduct an empirical study of RS pretraining on aerial images.
RVSA [17]	2022	89M	Train a plain ViTs with about 100 million parameters and propose a new rotated varied-size window attention.
RingMo [18]	2022	88M	Construct a large-scale dataset by collecting two million RS images and propose a training method designed for dense and small objects.
Geograph [33]	2023	24M	Propose a contrastive learning methods exploiting the spatiotemporal structure of RS data.
SatMAE [20]	2023	307M	Propose a pretraining framework for temporal or multispectral satellite imagery based on MAE.
ScaleMAE [22]	2023	307M	Propose a masked image modeling method using the area of the Earth covered by the image to determine the scale of the positional encoding.
Billion [19]	2023	2.4B	Conduct an empirical study of the effect of increasing the number of model parameters in RS.
GFM [21]	2023	80M	Investigate the potential of continual pretraining from large-scale ImageNet-22k models and propose a multi-objective continual pretraining paradigm.
CSP [133]	2023	-	Propose a contrastive spatial pretraining framework leveraging the abundant geospatial information associated with images.

various professional and academic datasets. mPLUG-2 [130] introduces a multimodule composition network, including text, image, and video. Metatransformer [131] proposed a unified framework performing learning across 12 modalities with unpaired data (e.g., natural language, 2-D images, 3-D point clouds, audio, video, time series, and tabular data).

These models are all built upon large-scale training data and self-supervised methods, harnessing the potential of unlabeled multimodal data encompassing both vision and language to train the FMs. As a result, these models can effectively perform a wide range of tasks that involve both vision and language processing.

D. RS Foundation Model

The research on natural image FMs has seen significant progress, and the field of RS has also garnered substantial attention in this regard. However, due to the inherent domain gap between natural images and RS images, directly applying pretrained models from natural images to RS images often leads to suboptimal results. To address this challenge, the construction of RSFMs can be divided into two approaches: training from scratch and continuous training using pretrained natural image models. The RSFMs are summarized in Table III.

1) Training From Scratch: The training from scratch approach involves collecting a large number of RS images and using the training methods employed in natural image FMs. Wang et al. [32] conducted experiments on RS pretraining models, showing that pretraining methods can effectively alleviate data differences but may still be influenced by task differences as downstream tasks require distinct representations from scene recognition tasks. Sun et al. [18] collected 2 million RS images to build a large-scale dataset covering diverse scenes and objects worldwide for pretraining. They proposed the RingMo masked image modeling method, addressing the problem of dense small targets often overlooked in complex RS scenes. Wang et al. [17] trained a visual RS transformer model with approximately 100 million parameters and introduced a new rotating variable-size window attention method to accommodate the characteristics of dense RS targets. Addressing the issue of large differences in the scale of RS targets, Reed et al. [22] proposed the scale-MAE method. This method explicitly learns the relationship between data at different known scales, enabling robust multiscale representations. Furthermore, SatMAE [20] established an MAE-based pretraining framework for temporal or multispectral satellite imagery, extending the paradigm to multispectral imagery as well as temporal dimensions. Geograph [33] introduced a method for comparative learning of spatiotemporal structure for RS data. In addition, CSP [132] utilized geospatial information in the image to construct a pretraining framework for contrastive learning.

2) Continuous Training: While training from scratch has propelled the development of RSFMs, it can be resource-intensive and challenging for large-scale models. As a result, some researchers have turned their focus to the method of continuous training, which utilizes the existing pretrained natural image FMs. Cha et al. [19] implemented a billion-level RS image FM based on Wang et al.'s [17] work. Mendieta et al. [21] constructed a compact yet diverse dataset called GeoPile to increase the amount of information in the pretraining data. They introduced the GFM model and carried out continuous pretraining based on the large-scale ImageNet-22k pretrained model to achieve an efficient geospatial FM with minimal resource cost and carbon impact. These continuous training approaches offer a cost-effective and efficient way to leverage the existing pretrained models in the RS domain.

IV. APPLICATIONS OF RSFM

In this section, we introduce the important applications of RSFMs. The applications are shown in Fig. 12. The applications are divided into three types: classification task, location task, and understand task. The classification tasks classify the image into a certain category at the image level or pixel level. Location tasks locate the target with boxes or masks. The understand tasks involve in the process of language. The representative algorithms in recent years are summarized in Tables IV and V.

A. Classification Task

1) Scene Classification: Scene classification is an image classification task similar to natural images. Given an RS image, it needs to be classified into a specific category according to the category settings [133]. In the whole RS image, the scene information contained is complex, so the picture used for scene

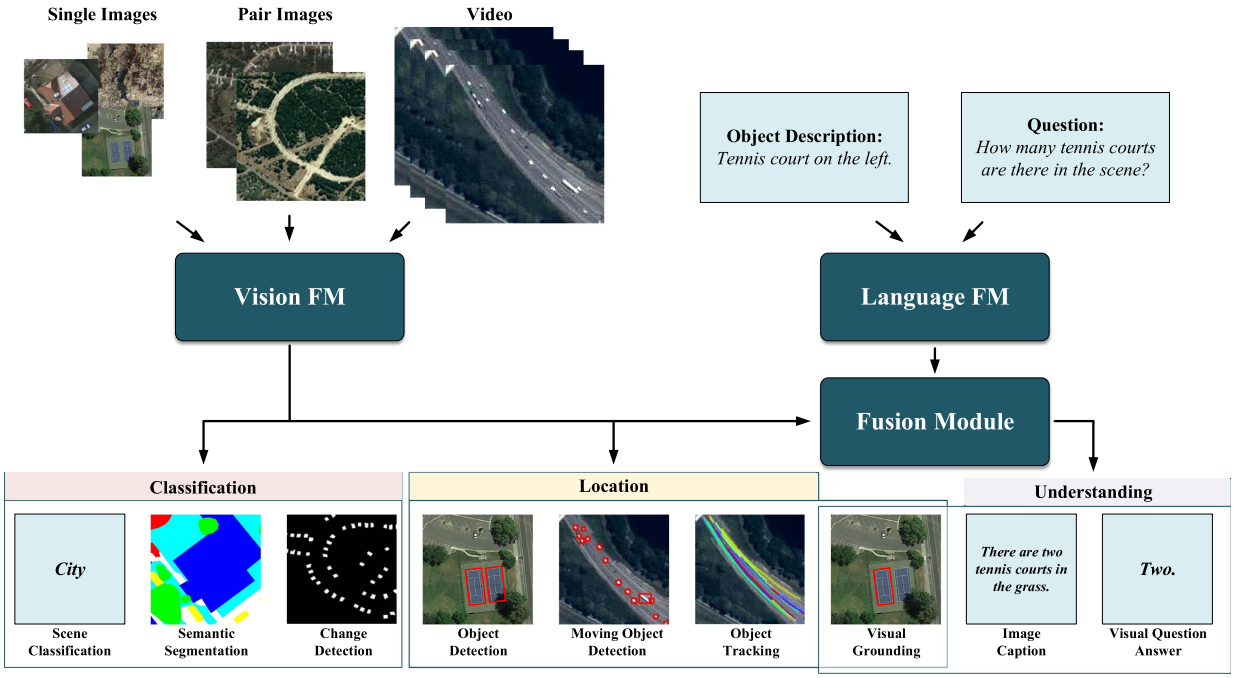


Fig. 12. Application of RSFMs. The applications are divided into three types: classification task, location task, and understand task. The classification task and location task are pure-visual tasks. The understand tasks involve in the process of language.

classification is usually cut out from the whole image according to a specific target to obtain a single-category image.

Scene classification tasks mainly require model representation capabilities, so many new algorithms are also applied to this field. Yang et al. [134] proposed an explainable multiscale spatial–spectral transformer. Spatial attention is a popular algorithm to achieve multiscale fusion [135], [136], [137]. Contrastive learning is widely applied to learn robust representation in RS [138], [139], [140], [141], [142]. In addition, GAN [143], neural architecture search [144], autoencoder [145], knowledge distillation [146], and collaborative framework [147] have also been applied to scene classification.

2) Semantic Segmentation: Semantic segmentation, also known as land-cover and land-use classification in RS, is one of the most important and widely used tasks in RS interpretation. This task involves classifying each pixel of an image into specific categories representing different ground objects. The classification process must take into account various factors, such as multiscale characteristics of ground objects, texture features, and spectral characteristics. However, the ground resolution of RS images can be limiting, as often a single pixel contains spectral information from multiple ground objects, making it challenging to accurately distinguish the boundaries of individual ground objects. Nevertheless, with advancements in high-resolution RS imaging technology, the accuracy of ground object segmentation has significantly improved and become widely used in various applications.

Numerous studies have been conducted in the field of semantic segmentation in RS. To leverage the potential of large amounts of unlabeled data, various algorithms, such as semisupervised [148], self-supervised [149], [150], [151], [152], [153], and self-training [154], [155] methods, are widely employed in RS semantic segmentation. These methods make effective use of unlabeled data to improve the accuracy of segmentation. In addition, factors, such as multiscale [156], [157], spatial–temporal [158], [159], [160], and boundary [161] information, which are crucial for semantic segmentation, are often carefully considered and incorporated into the design of algorithms. Furthermore, there are investigations into the interaction-based segmentation of RS images, which explore methods that involve interactions between pixels for more accurate segmentation results [162].

3) Change Detection: Remote sensing change detection (RSCD) is a process that involves identifying and extracting differences between multitemporal RS images captured in the same geographic area. The typical workflow of RSCD methods includes several steps, such as RS image preprocessing (alignment, correction, noise reduction, etc.), selecting appropriate change detection methods, and analyzing the results.

Due to the limited availability of labeled data for training, clustering algorithms remain the mainstream approach for change detection [167], [172], [173], [174]. Attention modules have also emerged as a noteworthy technique in this field [166], [177]. Notable attention modules include hybrid attention [168], convolutional block attention module [169], and spatialwise attention [175], which have been proposed to enhance the performance of change detection algorithms. In addition, the characteristics of GANs [165] and the Swin-transformer model [170] have been integrated into change detection algorithms, further improving their effectiveness.

Researchers have also proposed innovative methods to reason about both single-temporal and cross-temporal semantic correlations for change detection [171], and spatial–spectral cross-fusion approaches, such as SSCFNet [176], have been introduced to improve change detection performance.

TABLE IV
SUMMARY OF THE REPRESENTATIVE ALGORITHMS OF CLASSIFICATION APPLICATION

	Method	Publication	Modality	Key characteristics
Scene classification	Tang et al. [136]	JSTARS2021	Optical	Local features; Attention consistent network
	Peng et al. [145]	TGRS2021	Optical	Neural architecture search; Architecture regularization
	Liu et al. [139]	TNNLS2022	SAR	Contrastive learning; Dual dynamic GCN
	Zeng and Genget [140]	ISPRS2022	Optical	Few-shot learning; Task-specific contrastive loss; Self-attention and mutual-attention
	EMTCAL [163]	TGRS2022	Optical	Multiscale; Cross-level attention; Transformer.
	Qian et al. [146]	TGRS2022	SAR	G^0-based convolutional variational autoencoder; Hybrid network; Structural constraint
	Miao et al. [144]	TGRS2022	Optical	Involution GAN ; Siamese network; Semi-supervised
	Xu et al. [147]	TGRS2022	Optical	Knowledge distillation; ViT
	GCSANet [137]	JSTARS2022	Optical	Global context spatial attention; Multiscale
	Huang et al. [141]	GRSL2022	Optical	Spatial–temporal-invariant contrastive learning; Optimal transport
	Li et al. [138]	RS2022	Optical	Multiscale; Self-adaptive attention
	Li et al. [142]	TGRS2023	Optical	Contrastive learning; Hard feature transformation
	Yang et al. [135]	TGRS2023	Optical	Explainable; Multiscale representation; Spatial frequency; Texture-enhanced encoder; Transformer
	Zhao et al. [148]	TGRS2023	Optical	Local and long-range collaborative framework; Cross-feature calibration; Semi-supervised; ViT
	Xu et al. [143]	JSTARS2023	Optical	Contrastive learning; Occlusion images; Out-of-distribution
Semantic Segmentation	Shang et al. [156]	RS2020	MSI	Multiscale context; Adaptive fusion; Channel attention
	Li et al. [149]	JSTARS2021	Optical	SSL ; Multitask loss; Triplet Siamese network; High-level and low-level image features
	Luo et al. [158]	ISPRS2022	Optical	Domain adaptation; Cross spatiotemporal
	Li et al. [161]	TGRS2022	Optical	Boundary attention; Multilevel aggregation; Multitask learning
	MANet [157]	TGRS2022	Optical	Aware relation; Collaborative learning; Feature refinement; Multiscale
	Lu et al. [154]	TGRS2022	Optical	Data augmentation; Pseudolabeling-based self-training; Semisupervised learning; Linear sampling
	Bai et al. [159]	TGRS2022	HSI	Multibranch prediction; Self-attention; Spatial attention; Transformer
	Xue et al. [152]	TGRS2022	Optical, HSI	Generative self-supervised; Multimodal data
	Xue et al. [153]	TGRS2022	Optical, HSI	Multiview learning; Contrastive learning; Multitask learning; Multimodal images
	Yang et al. [150]	TGRS2022	Optical	Contrastive learning; SSL
	JAGAN [160]	JSTARS2022	HSI	Channel-space attention; GAN
	DIAL [162]	JSTARS2022	Optical	Active learning; Interactive segmentation
	Li et al. [155]	RS2022	Optical	Unsupervised domain adaptation; Self-training; Gradual class weights; Local Dynamic Quality
	Lu et al. [148]	TGRS2023	Optical	Semisupervised; Weak-to-strong consistency learning; Sparse dual-view cross-sample image generation
	Ghanbari et al. [164]	JSTARS2023	SAR	Local and global spatial dependency; Superpixels; Graph-based learning
	Marsocci et al. [151]	JSTARS2023	Optical	Continual learning; SSL ; Barlow twins
Change Detection	Dong et al. [165]	RS2020	Optical, SAR	GANs ; Deep belief networks; SSL
	Chen et al. [166]	RS2021	Optical	Siamese network; Attention mechanism
	Zhang et al. [167]	TIP2022	SAR	Contourlet fusion; Nonlocal clustering; Fuzzy clustering
	Li et al. [168]	TGRS2022	Optical	Densely attentive refinement network; Hybrid attention
	Lv et al. [169]	TGRS2022	MSI	Convolutional block attention module; Multiscale dilation convolution module
	SwinSUNet [170]	TGRS2022	Optical	Siamese U-shaped structure; Swin transformer
	Ding et al. [171]	TGRS2022	Optical	Bitemporal semantic reasoning network; CNN
	Zhang et al. [172]	TGRS2022	SAR	Multiobjective sparse feature learning; Cross-entropy clustering loss
	Dong et al. [173]	TGRS2022	SAR	Deep clustering; Noise-robustness loss; Shearlet transform
	Dong et al. [174]	TGRS2022	SAR	Deep clustering; Multiscale fusion; Octave convolution; Self-attention
	HFA-Net [175]	PR2022	Optical	High frequency; Spatial-wise attention
	SSCFNet [176]	JSTARS2023	Optical	Combined enhancement; Spatial-Spectral Cross Fusion
	Zhang et al. [177]	RS2023	Optical	Siamese network; Attention module; Transformer module; Multi-scale feature fusion

The classification application includes scene classification, semantic segmentation and change detection.

Furthermore, multiscale geometric techniques, such as Shearlet [173] and contourlet [167], have been applied to change detection to provide multiscale and multidirectional features, leading to better performance in detecting changes in RS images.

B. Location Task

1) Object Detection: Object detection is a crucial task in RS interpretation that involves identifying and localizing objects of interest within a scene. However, object detection in RS presents unique challenges compared with natural scenes. RS targets are often densely distributed, making it difficult for conventional horizontal bounding boxes to effectively capture and surround the targets. Moreover, the significant variation in scale among RS objects adds complexity to the detection task.

To address these challenges, researchers have proposed various approaches to improve object detection in RS imagery. For example, Zhang et al. [179] introduced the Laplacian feature pyramid to capture multiscale features, enhancing the detection performance. Ye et al. [184] developed an adaptive attention fusion method in conjunction with EfficientDet [236] to better handle multiscale objects. Bai et al. [181] leveraged time–frequency analysis and deep reinforcement learning to reduce computational complexity while ensuring detection accuracy.

In addition to the traditional target detection methods, recent efforts have focused on advancing learning algorithms for object detection. Weakly supervised learning [178], [183], [186], SSL [188], and distillation strategies [182], [187] have gained attention as effective approaches for improving object detection performance in RS imagery. These learning

TABLE V
SUMMARY OF THE REPRESENTATIVE ALGORITHMS OF LOCATION AND UNDERSTANDING APPLICATIONS

	Method	Publication	Modality	Key characteristics
Object Detection	Li et al. [178]	JSTARS2021	Optical	Point-based; Progressive candidate bounding box mining; Weakly supervised
	Zhang et al. [179]	TGRS2022	Optical	Foreground modeling; Foreground anchor reweighting loss
	Zhang et al. [180]	TGRS2022	Optical	Laplacian feature pyramid; Trainable Laplacian operator
	Bai et al. [181]	TGRS2022	Optical	Wavelet decomposition; Discrete wavelet multiscale attention mechanism; Reinforcement learning; Time-frequency analysis
	Liu et al. [182]	TGRS2022	Optical	Feature distillation structure; Global perception; Axial attention; Salient object detection
	Cheng et al. [183]	TGRS2022	Optical	Weakly supervised; Self-guided proposal generation
	Ye et al. [184]	RS2022	Optical	Adaptive attention fusion; EfficientDet; Spatial attention; Complete intersection over Union
	Zhang et al. [185]	ISPRS2023	Optical	Generalized few shot; Transfer learning; Metric learning; Representation compensation
	MOL [186]	ISPRS2023	Optical	Weakly supervised; Noisy learning; Multiview learning; Temporal consistency
	Li et al. [187]	TGRS2023	Optical	Instance-aware distillation; Relation based; Parameter-free masking module
	Zhang et al. [188]	JSTARS2023	Optical	Object centric; Attention-guided mask generator; Masked image modeling; Self-supervised learning; Vision transformer
	Han et al. [189]	JSTARS2023	Optical	Capsule reasoning; Transformer; DETR based; Multilevel feature fusion
Moving Object Detection	Zhang et al. [190]	TGRS2020	Optical	Background subtraction; Low-rank matrix decomposition; Structured sparsity regularization
	Zhang et al. [191]	TGRS2020	Optical	Background subtraction; Online robust principal component analysis; Structured sparsityinducing norm
	Feng et al. [192]	ISPRS2021	Optical	Keypoint-based detection; Spatial motion information guided; Relative spatial relationship; MOT
	Zhang et al. [193]	TPAMI2022	Optical	Moving-confidence-assisted matrix decomposition; Moving- confidence- assisted Matrix Decomposition
	Pi et al. [194]	TGRS2022	Optical	Motion information; Low resolution; Transformer
	DSFNet [195]	GRSL2022	Optical	Two-stream detection network; Dynamic and static information fusion
	SDANet [196]	TIP2023	Optical	Anchor-free detector; Semantic embedded; Weakly supervised learning; Density matching; Road information; Bi-directional conv-RNN
Object Tracking	HRSiam [197]	TIP2021	Optical	SOT; Siamese network; High-spatial-resolution representation; Gaussian mixture model
	DFAT [198]	TGRS2022	Optical	SOT; Dynamic feature adaptive; Candidate experts
	Cui et al. [199]	TGRS2022	Optical	SOT; Deep reinforcement learning; Occlusion; Temporal and spatial context
	MBLT [200]	TGRS2022	Optical	SOT; Motion and background learning; Location probability; Segmentation
	Song et al. [201]	TGRS2022	Optical	SOT; Channel and spatial attention; Cross attention; Composite feature combine
	Li et al. [202]	TGRS2022	Optical	SOT; Correlation particle filter; Kalman filter
	Chen et al. [203]	JSTARS2022	Optical	SOT; Historical model; Correlation filter; Antidrift tracker correction
	Li et al. [204]	GRSL2022	Optical	SOT; Correlation filter; Interacting multiple models
	Nie et al. [205]	GRSL2022	Optical	SOT; Temporal motion compensation; Multidimensional information aware; Siamese
	Li et al. [206]	TCYB2023	Optical	SOT; Dual-branch spatial-channel coattention; Collaborative learning; Geometric constraint
	Zhang et al. [207]	TEVC2023	Optical	SOT; Quantum evolution; Rotation operator; Trajectory inference; Balanced Intersection over Union
	SiamMDM [208]	TGRS2023	Optical	SOT; Dynamic template update; Multiple response map fusion; Score-guided target motion trajectory prediction
	Ao et al. [209]	TIP2020	Optical	MOT; Probabilistic noise modeling; Multi morphological; Evaluation protocols
	He et al. [210]	TGRS2022	Optical	MOT; Graph reasoning; Multitask learning; Spatiotemporal; AIR-MOT dataset
	Zhang et al. [211]	TGRS2023	Optical	MOT; Bidirectional; Invalid fragment trajectory backtracking; Trajectory criteria; Integrate with SOT
	CFTracker [212]	TGRS2023	Optical	MOT; Cross-frame feature update; Cross-frame training flow; Joint detection and tracking
VG	Sun et al. [213]	ACM MM2022	Optical	RSVG dataset; Numerical geospatial relations; Geospatial relation graph
	Zhan et al. [214]	TGRS2023	Optical	DIOR-RSVG dataset; Transformer; Multigranularity visual language fusion
	Yuan et al. [215]	arXiv2023	Optical	RefSegRS dataset; Language-guided cross-scale enhancement; Segmentation
Image Caption	Wu et al. [216]	IJCNN2020	Optical	Long short-term memory network; Scene attention
	Ye et al. [217]	TGRS2022	Optical	Joint training; Multilabel attributes; Multilabel classification; Differentiable sampling operator; Dynamic contrast loss
	Li et al. [218]	TGRS2022	Optical	Recurrent attention; High-level attentive maps; Long short-term memory; Semantic gate
	Wang et al. [219]	ISPRS2022	Optical	Multi-label; Semantic attribute extractor; Cross-modal semantic feature fusion operators
	Li et al. [220]	RS2022	Optical	Multi-level attention mechanism; Encoder-decoder
	TypeFormer [221]	GRSL2022	Optical	Captiontype controller; Multiscale vision transformer
	Wang et al. [222]	IGARSS2022	Optical	Pure transformer
	Yang et al. [223]	ISPRS2022	Optical	Meta learning; Support tasks
	Wang et al. [224]	JSTARS2022	Optical	Multiscale; Multiinteraction
	Zia et al. [225]	IJAEOG2022	Optical	Multi-scale; Adaptive attention; Topic sensitive word embedding
	Zhang et al. [226]	RS2023	Optical	Multi-source interactive; Stair attention
	Chg2Ca [227]	arXiv2023	Optical	Change caption; Siamese CNN-based; Hierarchical self-attention
VQA	Zhang et al. [228]	TGRS2023	Optical	Hash-based spatial multiscale visual representation; Spatial hierarchical reasoning
	Prompt-RSVQA [229]	CVPRW2022	Optical	Prompt; DistilBERT
	Yuan et al. [230]	TGRS2022	Optical	Self-paced curriculum learning; Spatial transformer; Language-guided; RNN
	Bazi et al. [231]	TGRS2022	Optical	CLIP; Co-attention; Transformer
	Al et al. [232]	IJRS2022	Optical	Open-set dataset; VQA-TextRS dataset; Transformer
	Chappuis et al. [233]	IGARSS2022	Optical	Recurrent neural network; BERT
	Yuan et al. [234]	TGRS2022	Optical	Multitemporal; Change detection; CDVQA dataset
	Bashmal et al. [235]	JSTARS2023	Optical	Visual question generation; Language transformers; GPT-2; TextRS-VQA dataset

The location application includes object detection, moving object detection, target tracking and visual grounding (VG). The understanding application includes visual grounding, image caption and visual question answering (VQA).

methods help enhance the model's ability to learn from limited labeled data, leading to better object detection results in RS applications.

2) Moving Object Detection: The advancement of RS photography technology has allowed satellites to capture continuous videos by staring at specific areas [237]. Moving object detection in satellite videos involves extracting objects in motion, such as airplanes, ships, and cars. However, due to factors, such as changes in lighting, weather, and viewing angles, the moving targets often occupy only a small portion of the scene. As a result, moving target detection faces challenges of false detections and missed detections caused by image blur and shaking.

One popular approach for moving object detection in RS is background subtraction [190], [191]. Zhang et al. [193] explored background modeling and incorporated sparse constraints to achieve the accurate extraction of moving objects. However, in low-quality videos, background subtraction methods tend to produce a large number of false alarms and may miss many positive targets. To address these limitations, several methods based on temporal and appearance features have been proposed [192], [194]. For example, DSFNet [195] is a two-stream detection network that considers both dynamic and static information. SDANet [196] is an anchor-free detector that utilizes road information to suppress false alarms. These approaches aim to improve the accuracy and robustness of moving object detection in RS videos, making it an active area of research with promising applications in various domains.

3) Object Tracking: Object tracking in satellite videos involves continuously tracking single or multiple objects [238]. However, due to the limited ground resolution of satellite imaging, the targets in satellite images are usually very small, providing limited detail information, which can lead to tracking deviations or difficulty in distinguishing targets from the background. In addition, the presence of clouds and buildings can further hinder object tracking.

In single-object tracking (SOT), correlation filter-based algorithms remain comparable [202], [203], [204]. To address challenges in high-spatial-resolution representation, HRSiam [197] is proposed. Cui et al. [199] explore reinforcement learning to tackle occlusion problems during tracking. MBLT [200] leverages motion and background information to improve object tracking in satellite videos. Spatial-channel attention is also utilized in SOT for RS [201], [206], and dynamic information in videos is explored in [198], [205], and [208].

In multiple-object tracking (MOT), Ao et al. [209] propose probabilistic noise modeling algorithms and evaluation protocols for MOT. He et al. [210] develop graph reasoning algorithms that leverage the relations between objects and introduce the AIR-MOT dataset. Zhang et al. [211] utilize bidirectional tracking for trajectory verification to mitigate the influence of similar objects. CFTracker [212] introduces a cross-frame feature update and training flow to enhance tracking performance. These advancements in object tracking techniques for satellite videos hold significant promise for improving the accuracy and robustness of RS applications.

C. Understanding Task

Understanding tasks in RS interpretation encompass tasks involving linguistic descriptions, including visual grounding (VG), image captioning, and visual question answering (VQA).

1) VG: VG, also known as referring location, is a derivative task of target detection. In contrast to target detection, where the category of the target is predefined, VG requires locating the target in the image based on a given linguistic expression. This task demands not only language understanding but also comprehension of the categories and relationships of targets in RS images to achieve the accurate localization. RS VG is still in its early stages of development, and datasets, such as RSVG [213], DIOR-RSVG [214], and RefSegRS [215], have been introduced.

Zhan et al. [214] proposed the MGVLF module, which combines image features extracted from CNNs and text features obtained using BERT to achieve target localization. Similarly, Sun et al. [213] proposed the GeoVG model, which also utilizes BERT to encode text. Moreover, the geospatial relations of the target are taken into consideration to improve accuracy.

2) Image Caption: Image captioning involves summarizing the text describing an image based on the information in the image. In the case of RS images, which are taken from a high altitude, the targets are small and numerous. As a result, RS image caption algorithms tend to focus on describing the dominant content in the scene while overlooking smaller objects of interest. In addition, current RS image description datasets suffer from issues, such as small image sizes and limited richness of content. Existing algorithms face difficulties when applied to large-scale images and struggle to fully describe the content using rich, hierarchical, and coherent language.

Numerous studies have been conducted on RS image captioning, with attention mechanisms and semantic information often employed [216], [218], [219], [220], [225], [226]. Wang et al. [222] proposed a pure transformer for image captioning in RS. TypeFormer [221] was introduced to control the type of generated captions, while Chg2Ca [227] extended captioning to change descriptions in RS. In addition, multilabel [217] and metalearning [223] have also been considered.

3) VQA: VQA involves answering questions based on image information. In RS, VQA is mainly divided into three types:

1) determining whether a specific target is present in the image;
2) identifying the target within the question description area;
3) counting the number of targets.

Current RS VQA models often overlook information in the image space, leading to lower accuracy in information answering. Moreover, the design of current VQA tasks is relatively simple and cannot handle more complex questions. Furthermore, these tasks do not consider the role of landmark buildings in question answering, necessitating the integration of real geographic information to enhance the practicality of visual answering tasks.

Zhang et al. [228] propose a spatial hierarchical reasoning network to model and reason the relationships between entities. Yuan et al. [230] introduce a self-paced curriculum learning

approach for VQA in RS. Models, such as BERT [229], [233], CLIP [231], and GPT [235], have been widely applied, and addressing the open-set problem of VQA in RS is also a focus of research [232]. The VQA task related to change detection is an emerging research direction [234]. Furthermore, visual question generation is also a valuable area for the development of VQA [235].

V. Exploration of the Effectiveness of Existing FMs on Various RS Applications

RS interpretation is a technology utilized for analyzing and comprehending images. Numerous FMs have been developed for image interpretation. In this section, we conduct systematic experiments to compare the performance of RSFMs with natural FMs in RS applications. The experiments cover three key aspects: global representation (scene classification), local representation (semantic segmentation), and object localization (object detection). Furthermore, we provide a detailed discussion on the strengths and weaknesses of the currently existing FMs.

A. Scene Classification

We conducted experiments on representation capabilities using five commonly used scene classification datasets in RS.

1) Dataset: We conducted experiments on scene classification using a total of five datasets.

WHU-RS19 [239]: This dataset consists of 1005 images collected from Google Earth imagery, covering 19 categories of RS scenes. The images are fixed at a size of 600×600 pixels with a spatial resolution of 0.5 m. Each category contains approximately 50 images.

UCMerced [240]: UCMerced is obtained from the United States Geological Survey National Map and comprises 21 classes, with 100 images per category. Compared with WHU-RS19, UCMerced offers a higher spatial resolution of 0.3 m. The images in UCMerced are cropped into smaller regions of 256×256 pixels.

AID [241]: AID is an aerial imagery dataset that includes 10 000 images with 30 categories of RS scenes. The number of sample images varies from 220 to 420 for each class. The spatial resolutions of the images range from 8 to 0.5 m, presenting a challenge in scene classification. AID dataset was constructed to consider higher intraclass variations and smaller interclass dissimilarity for comprehensive comparisons.

RESISC [242]: Also known as NWPU-RESISC45, this dataset consists of 31 500 images extracted from RS images in Google Earth. It comprises 45 categories, with 700 images in each class. The spatial resolution of RESISC varies from 30 to 0.2 m. The image size is standardized to 256×256 pixels. RESISC covers over 100 countries and regions worldwide, providing rich image variations, high within-class diversity, and between-class similarity.

fMoW [41]: The functional Map of the World (fMoW) is a large-scale RS dataset used for training machine learning models. It enables the prediction of building functions and land use based on the time series of satellite imagery. The dataset contains over 1 million images from more than 200 countries, annotated with 63 categories. There are two versions available: fMoW full, consisting of four-band and eight-band multispectral images, and fMoW-RGB, which is in JPEG format and contains RGB images converted from multispectral data. Since most FMs only work with RGB images, we utilized the fMoW-RGB dataset for our experiments.

These datasets provide a diverse range of RS scenes and facilitate a thorough evaluation of the representation capabilities of FMs.

2) Experimental Analysis: Except for the fMoW dataset, the other datasets were not initially divided into training and test sets. Therefore, we randomly divided these datasets into training and test sets using different training ratios. Specifically, we selected 50% of the WHU-RS19 dataset, 20% and 50% of the AID dataset, 50% and 80% of the UCMerced dataset, and 10% and 20% of the RESISC dataset as the training sets, respectively. The remaining images were used as the test sets. To mitigate the impact of randomness, each separation with a specific training ratio was performed three times. For the fMoW dataset, we used the official training and test sets provided by the dataset creators. The experimental results are presented in Table VI. In this experiment, four RSFMs are selected for comparison, including RSP [32], RVSA [17], SatMAE [20], and ScaleMAE [22]. In addition, SwinV1 [60], SwinV2 [26], CLIP [10], ALBEF [27], BEiT-v3 [30], BLIP-2 [29], and SAM [31] are involved in experiments.

Among the FMs for RS, the RSP model has demonstrated promising results on multiple datasets. This can be attributed to the model's utilization of MillionAID's classification labels during pretraining, which enables better performance on scene classification datasets. On the other hand, RVSA, SatMAE, and ScaleMAE are FMs trained using label-free self-supervised algorithms. SatMAE exhibits superior performance on the WHU-RS19 and UCM datasets. The ScaleMAE model achieves the best results on AID, RESISC, and fMow datasets while also performing competitively with other RSFMs on WHU-RS19 and UCMerced datasets. The success of ScaleMAE can be attributed to its consideration of different ground sampling distances during the training process, leading to effective adaptation to datasets with multiresolution characteristics.

Among the natural FMs, the BLIP2 and CLIP series models have achieved remarkable results. The BLIP2 model utilizes the image encoder from CLIP, resulting in a similar performance to CLIP. The CLIP model demonstrates excellent representation capabilities for RS images, thanks to the inclusion of RS-related data in its dataset. Furthermore, compared with models trained with masked image modeling techniques, such as Swin transformer, FMs trained using image–language pairs training place greater emphasis on capturing high-level semantic information. As a result, they exhibit superior representation abilities. On the other hand, the underwhelming results of SAM can be attributed to its design for segmentation tasks. The large feature map causes background information to interfere with target information during average pooling, consequently impacting classification outcomes.

Moreover, when comparing all FMs, we observe that natural FMs consistently outperform current RSFMs. Even when Swin transformer is trained without RS images, it still demonstrates superiority on WHU-RS19, UCM, AID, and RESISC datasets.

TABLE VI
NUMERICAL RESULTS OF COMPARISONS WITH FMS ON FIVE SCENE CLASSIFICATION DATASETS WITH DIFFERENT TRAINING RATIOS

Method Name	Tag	WHU-RS19	AID		UCMerced		RESISC		fMoW
		0.5	0.2	0.5	0.5	0.8	0.1	0.2	
RSP	Swin-T	83.2±1.23	82.4±0.23	86.2±0.06	**93.8±0.55**	95.0±0.67	67.7±0.31	71.7±0.20	36.3
	ViTAEv2-S	86.1±1.13	86.6±0.40	89.6±0.22	**95.7±0.59**	**96.1±0.40**	74.5±0.33	77.7±0.32	37.8
RVSA	ViT	66.6±1.77	65.2±0.76	74.2±0.23	69.9±0.84	79.3±1.21	57.2±0.50	64.2±0.55	34.4
	ViTAE	67.9±0.74	67.7±0.40	76.2±0.33	72.6±1.70	81.4±1.27	59.5±0.41	66.1±0.24	34.5
SatMAE	default	82.7±1.66	77.6±0.53	82.9±0.48	79.3±0.84	82.9±0.19	64.2±0.41	69.3±0.17	32.7
ScaleMAE	default	84.7±0.37	86.5±0.39	90.6±0.31	75.6±1.83	83.4±1.46	75.8±0.16	80.6±0.12	43.8
SwinV1	Swin-B-IN22K-224	85.6±1.03	77.6±0.45	82.1±0.42	88.2±1.27	90.4±0.30	67.0±0.52	72.0±0.19	34.2
	Swin-L-IN22K-224	89.4±0.52	87.6±0.16	90.7±0.13	91.9±0.90	93.8±0.97	77.9±0.48	81.3±0.39	38.9
SwinV2	SwinV2-B-IN1K-256	90.9±0.58	89.7±0.37	92.4±0.38	93.5±1.18	95.4±0.30	80.1±0.19	83.8±0.32	40.6
	SwinV2-B-IN22K-192	90.1±0.32	87.3±0.07	90.4±0.42	92.7±1.13	95.1±0.11	76.4±0.57	79.9±0.23	38.8
	SwinV2-L-IN22K-192	90.6±0.67	88.5±0.06	91.6±0.30	92.9±1.09	95.5±0.85	79.0±0.11	82.2±0.19	39.8
CLIP	ViT-B/32	88.9±0.86	92.5±0.37	94.3±0.50	89.6±0.39	93.8±1.36	85.4±0.06	87.5±0.13	42.7
	ViT-B/16	87.1±0.37	89.7±0.40	92.2±0.28	88.9±0.92	93.0±1.19	81.4±0.32	83.8±0.14	38.0
	ViT-L/14	**91.2±0.49**	**93.9±0.24**	**95.8±0.08**	92.2±0.68	95.5±1.21	**89.2±0.04**	**90.8±0.10**	**50.0**
	ViT-L/14-336	**91.9±0.93**	**94.0±0.26**	**96.1±0.11**	92.5±0.47	**95.9±0.81**	**89.6±0.19**	**91.2±0.08**	**52.2**
ALBEF	default	84.9±0.49	83.8±0.31	86.8±0.38	87.5±1.28	92.1±0.85	72.3±0.47	76.1±0.16	34.2
	4M	89.3±0.28	89.0±0.38	90.7±0.49	88.6±1.13	92.8±1.11	76.6±0.16	79.6±0.09	35.4
BEiT-v3	BEiT3-L-p16-224	77.6±1.26	78.2±0.63	83.7±0.53	83.6±0.70	88.6±1.01	69.2±0.29	73.9±0.17	31.7
BLIP-2	default	**93.1±0.58**	**95.6±0.17**	**96.8±0.02**	**93.9±0.91**	**96.8±0.74**	**91.8±0.24**	**93.1±0.06**	**53.0**
SAM	ViT-B	74.1±0.43	58.7±0.36	64.4±0.36	66.6±0.88	71.6±1.81	39.9±0.22	44.3±0.28	-

The top three results are masked in red, green and blue.

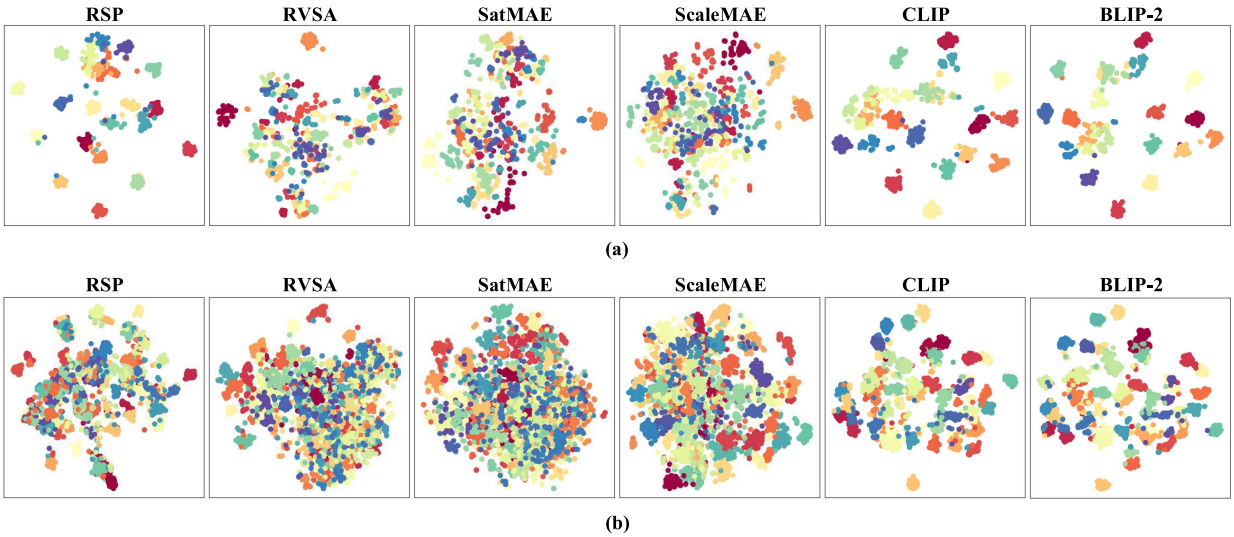


Fig. 13. Feature visualizations of six FMs. (a) UCMerced dataset. (b) RESISC dataset.

Furthermore, we use the T-SNE algorithm to visualize features, as shown in Fig. 13. The RSP shows good features in UCMerced dataset but performs worse in the complex dataset, RESISC. The CLIP and BLIP-2 perform well across these datasets. This further corroborates the notion that large-scale pretrained natural FMs remain highly competitive in the field of RS. In addition, this insight inspires us to leverage natural FMs to enhance the development of RSFMs in terms of efficiency and performance.

B. Semantic Segmentation

Semantic segmentation, also known as land-cover classification in RS, differs from scene classification as it involves classifying pixels of an image on a pixel level. This allows for the evaluation of the local representation capabilities of FMs. In our experiments, we combine the UperNet [243] with the FMs and fine-tune them using five semantic segmentation datasets to compare their performance.

1) Datasets: We conducted experiments on semantic segmentation using a total of five datasets.

DFC22 [244]: The DFC22 dataset is based on the MiniFrance dataset [245] and is designed for training semisupervised semantic segmentation models for land-use/land-cover mapping. It contains 766 labeled images with a resolution of approximately 2000 × 2000. In our experiments, we compared the fine-tuned performance of each FM. We uniformly resized all images to

TABLE VII
NUMERICAL RESULTS OF COMPARISONS WITH FMS ON FIVE SEMANTIC SEGMENTATION DATASETS

Model	Backbone	Pretrained data	DFC22		Vaihingen		MER		GID-15		Potsdam	
			mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
SatMAE	ViT-L	fMoW	46.3	58.7	69.3	77.0	64.1	73.0	88.1	92.0	75.7	83.8
ScaleMAE	ViT-L	fMoW	50.2	63.2	72.3	79.7	64.3	73.5	90.7	93.5	76.9	84.5
RVSA	ViT-B_kvdiff	MillionAID	49.0	61.1	71.6	79.7	64.3	74.0	92.6	95.6	77.6	84.7
	ViTAE-B_kvdiff		49.3	61.4	71.4	79.3	64.4	74.5	92.2	95.7	77.7	84.6
Deeplabv3+	R50	IN1K	48.1	59.6	73.1	80.2	63.9	74.6	94.6	96.7	78.3	84.9
ConvNeXt	ConvNeXt-T	IN1K	50.1	63.4	73.7	80.7	64.0	74.7	91.6	95.3	79.0	85.9
	ConvNeXt-S		49.0	61.3	73.5	81.1	64.5	74.8	92.1	95.4	79.3	86.0
	ConvNeXt-B		49.7	62.8	74.2	81.4	64.9	75.4	94.1	96.7	78.9	85.5
	ConvNeXt-B	IN22K	50.0	63.0	75.3	82.3	65.0	75.1	94.5	**96.9**	79.8	86.3
VAN	VAN-B	IN1K	49.4	61.6	73.9	81.1	65.3	75.7	94.2	96.5	**79.9**	**86.4**
MViTv2	MViTv2-T	IN1K	50.1	62.4	**75.4**	**83.7**	**68.7**	**77.0**	93.0	95.6	79.0	85.7
	MViTv2-S		50.5	62.6	**75.7**	**82.6**	**68.3**	**77.0**	93.5	96.3	79.3	**86.4**
	MViTv2-B		50.5	62.5	**75.6**	**82.9**	67.2	**76.7**	94.3	96.5	**80.0**	86.3
SegFormer	SegFormer-B2	IN1K	40.0	52.0	62.0	70.9	56.1	65.4	81.3	87.5	73.1	81.6
ViT	ViT-B	IN1K	49.3	60.8	71.4	78.8	66.5	75.0	92.5	95.2	77.9	84.9
	ViT-L		48.9	61.2	71.2	78.3	66.3	74.6	92.4	95.0	77.4	84.6
	ViT-L	IN22K	48.3	60.5	70.7	78.2	65.1	74.1	89.8	92.8	77.3	85.1
SwinV1	Swin-T	IN1K	50.2	62.7	73.8	80.9	64.1	74.5	92.0	95.3	79.1	86.3
	Swin-S		**51.1**	**64.1**	74.2	81.4	65.9	74.9	94.3	96.8	79.3	85.9
	Swin-B		49.9	62.6	73.2	80.4	64.5	75.4	**95.0**	**97.1**	79.3	86.2
	Swin-B	IN22K	**51.6**	**65.6**	75.1	81.9	66.7	75.9	**94.8**	**97.1**	**80.1**	**86.8**
SwinV2	Swinv2-T-p4w8	IN1K	50.6	63.2	73.5	80.8	64.7	74.9	94.0	96.5	79.0	85.6
	Swinv2-S-p4w8		50.3	62.1	73.4	80.7	65.7	76.3	94.4	96.8	79.1	85.7
	Swinv2-B-p4w8		50.1	62.4	73.6	80.4	65.4	75.7	94.7	96.8	79.4	85.7
	Swinv2-B-p4w12	IN22K	**50.7**	63.3	74.9	81.9	66.0	76.1	**95.0**	96.9	79.3	85.9
CLIP	CLIP-ViT-B	WebImageText	49.2	60.2	74.1	80.8	**67.9**	76.3	-	-	-	-
DenseCLIP	DenseCLIP-ViT-B	WebImageText	49.9	**63.9**	74.1	81.0	67.2	76.5	-	-	-	-

The top three results are masked in red, green and blue.

512×512 pixels and randomly split them into training and test sets with a 4:1 ratio. No unlabeled images were used.

Vaihingen [246]: The Vaihingen dataset, released by the International Society for Photogrammetry and Remote Sensing (ISPRS) Commission, consists of RS images of the Vaihingen village. It is divided into 33 patches, with 17 patches used as the test set and the remaining 16 patches as the training set. The patch sizes range from 1996×1995 pixels to 3816×2550 pixels. In our experiments, we resized all data to 512×512 pixels with an overlap size of 128 pixels.

MER [247]: The Mars-Seg (MER) dataset consists of 4155 RGB images and 1024 grayscale images, capturing the Martian landscape. The RGB images have a resolution of 560×500, while the grayscale images have a resolution of 1024×1024. To conduct our experiments, we uniformly resized the MER images to 512×512 pixels. We randomly split both the RGB and grayscale images into training and test sets with a 4:1 ratio.

Gaofen Image Dataset (GID-15) [248]: The GID-15 is a large dataset for land-use and land-cover classification. It contains ten high-quality Gaofen-2 images from different cities in China, with a resolution of 7200×6800 pixels. In our experiments, we applied the sliding window method with a window size of 512×512 and a stride size of 384 to extract patches from the images. The resulting patches were then randomly divided into training and test sets using a 4:1 ratio.

Potsdam [249]: The ISPRS Potsdam dataset comprises 38 high-resolution aerial images with a resolution of 0.5 m. It is annotated with six categories, including impervious surfaces, buildings, low vegetation, trees, cars, and clutter. The dataset is divided into 24 training images and 14 testing images, each with a size of 6000×6000 pixels. In our experiments, we used the sliding window method with a window size of 512×512 and a stride size of 384 to extract image patches.

2) Experimental Analysis: The experimental results are presented in Table VII. In this experiment, three RSFMs are selected for comparison, including SatMAE [20], ScaleMAE [22], and RVSA [17]. In addition, Deeplabv3+ [250], ConvNeXt [251], VAN [252], MViTv2 [253], SegFormer [254], ViT [58], SwinV1 [60], SwinV2 [255], CLIP [10], and DenseCLIP [256] are involved in experiments.

From the table, we can observe that different RSFMs yield varying effects on semantic segmentation. ScaleMAE performs well on DFC22 and Vaihingen datasets, achieving mIoU scores of 50.2 and 72.3, respectively. RVSA shows better performance on the larger datasets GID-15 and Potsdam. The three RSFMs perform similarly on the MER dataset.

Among the natural base models, Swin transformer achieves the best results on DFC22, GID-15, and Potsdam datasets. This can be attributed to two factors: its ability to represent RS images and the structure of the model. Our scene classification experiments have already demonstrated the Swin transformer's strong representation capability for RS images. In semantic segmentation, Swin transformer provides multiscale features, effectively improving accuracy. MViTv2 also performs admirably,

particularly excelling on the Vaihingen and MER datasets. It is worth noting that MViTv2 is only pretrained on ImageNet 1K, indicating that its performance gains are primarily derived from its excellent multiscale design and improved representation ability of local features. Furthermore, the tiny and small versions of the MViTv2 series models are also competitive. This highlights the fact that the number of parameters cannot solely determine the performance of a model when it is applied to a specific RS dataset.

By comparing all the FMs, it becomes evident that natural FMs outperform current RSFMs in terms of performance. This can be attributed to two main factors. First, the current RSFMs are pretrained on the fMoW and MillionAID datasets, which do not comprehensively cover all RS datasets. Consequently, these models do not exhibit significant advantages in the local representation segmentation. In addition, the current FMs predominantly utilize ViT as the underlying structure, which also affects their performance. Therefore, designing an RSFM requires not only a well-suited pretraining algorithm but also an excellent multiscale structure that enables the model to meet the requirements of diverse applications.

C. Object Detection

Object detection is a crucial task in RS interpretation, as it necessitates the model's ability to handle objects with significant size variations while also performing accurate classification. In our experiments, we integrate the oriental-RCNN [257] with the FMs and fine-tune them on two widely used RS object detection datasets to assess their performance.

1) Datasets: We utilized two object detection datasets for our experiments: DOTA v1.0 and DIOR-R.

DOTA v1.0 [258]: DOTA is a renowned dataset widely used for rotated object detection in the RS domain. We employed DOTA v1.0 for evaluation purposes. This dataset consists of 15 common categories, 2806 images, and 188 282 instances, gathered from various sensors and platforms. The image sizes in DOTA v1.0 range from 800×800 to 4000×4000 pixels. The labels for the training and validation sets are publicly available. The categories in DOTA v1.0 include plane (PL), baseball diamond (BD), bridge (BG), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HB), swimming pool (SP), and helicopter (HC). For our experiment, we trained the model on the train set and evaluated its performance on the validation set.

DIOR-R [259]: The DIOR-R dataset is an extension of the previous DIOR dataset [260]. It comprises 23 463 images and 192 518 instances, encompassing a wide range of scenes and 20 common object classes. The images in DIOR-R have a fixed size of 800 pixels. The object categories in DIOR-R include airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway service area (ESA), expressway toll station (ETS), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM).

2) Experimental Analysis. DOTA: The experimental results for DOTA are presented in Table VIII. In this experiment, ResNet [261], PVTv2 [262], Poolformer [263], VAN [252], MViTv2 [253], ConvNeXt [251], SwinV1 [60], SwinV2 [255], and RVSA [17] are involved in experiments.

Overall, there is no significant gap between the FMs designed for natural images and those tailored for RS. Swin transformer achieved an mAP of 76.5, while ConvNext and RVSA both achieved 76.3 mAP. However, Swin transformer and ConvNext yielded comparable results to the ViTAE-base model in RVSA when using a large-parameter model. Consequently, RVSA demonstrates superior performance in RS target detection. This can be attributed to the incorporation of a rotating variable-size window attention method in the RVSAs structural design, effectively enhancing the model's accuracy in object detection.

From a category standpoint, Swin transformer exhibits a significant advantage over RVSAs ViTAE-B in the baseball-diamond and roundabout categories, achieving 11.6 and 6.7 mAP scores higher, respectively. These two object classes are less sensitive to rotating boxes, rendering the design of RVSA less beneficial in terms of performance improvement. Conversely, RVSA surpasses Swin transformer by 8.5 mAP in the soccer-ball-field category. Therefore, incorporating the FM can enhance the accuracy of the downstream applications. Certain modules specifically designed for downstream tasks remain crucial even in FMs and can effectively improve model performance.

DIOR-R: Similarly, we can observe similar patterns in the DIOR-R dataset experiments as in the DOTA dataset. The experimental results on the DIOR-R dataset are presented in Table IX. Swin transformer, ConvNext, and RVSA all achieved great performance. RVSA demonstrated excellent performance in both the tennis court and vehicle categories. Swin transformer employs a multiscale transformer structure, enabling better detection of super large airports. ConvNext, with its convolutional structure, exhibited superior detection performance for ships.

In this section, we conduct experiments focusing on three essential aspects: global representation (scene classification), local representation (semantic segmentation), and object localization (object detection). Our findings from these experiments reveal that the foundational models trained with natural images exhibit comparable performance with the ones developed for RS. The CLIP model stands out for its remarkable performance in scene classification; however, it does not perform as well in semantic segmentation. At present, no single FM can excel across all these applications in our experiments. This underscores the necessity for further development of FMs to suit a wide array of applications in the RS field.

VI. Brain-Inspired RSFM

A. Overall Architecture of the Brain-Inspired RSFM

The FM aims to address multiple modalities and tasks with a unified approach. However, based on the experiments conducted in Section V, we have identified several shortcomings in the current FMs' performance. Particularly, there is a lack of RSFMs

TABLE VIII
NUMERICAL RESULTS OF COMPARISONS WITH FMS ON DOTA V1.0 DATASETS WITH DIFFERENT TRAINING RATIOS

	Model	Pretrained data	mAP	PL	BD	BG	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HB	SP	HC
R50	R50	IN1K	66.6	88.7	73.5	45.3	63.8	61.3	82.8	87.7	90.4	61.2	62.0	56.9	61.8	63.5	54.5	45.8
PVTv2	pvtv2-b0	IN1K	71.9	89.4	75.7	51.9	76.7	66.2	84.6	88.5	90.6	67.7	69.2	63.9	66.1	70.9	57.6	58.9
	pvtv2-b1		74.6	89.6	83.1	55.5	77.9	68.9	85.7	89.2	90.7	75.3	69.4	65.9	70.1	74.9	59.5	62.8
	pvtv2-b2		74.6	89.9	84.6	58.1	80.0	68.2	86.0	89.3	90.7	74.8	69.8	64.2	71.0	76.0	60.1	56.6
	pvtv2-b3		74.6	89.9	78.9	59.1	78.8	67.6	86.1	**89.4**	90.6	76.8	70.8	65.4	69.0	76.1	60.6	60.1
Poolformer	poolformer-s12	IN1K	70.7	88.8	77.4	48.8	77.5	59.3	83.4	88.5	90.5	67.2	62.4	62.2	67.2	66.6	58.0	61.9
	poolformer-s24		73.5	89.6	77.5	56.6	78.6	65.5	84.5	88.8	90.7	74.0	62.4	66.4	66.3	75.1	58.8	67.3
	poolformer-s36		72.8	89.6	77.8	50.6	74.7	60.1	84.2	88.9	90.7	72.2	62.6	63.4	71.4	74.2	59.7	71.7
VAN	van-b0	IN1K	72.7	89.2	77.6	51.4	77.6	65.5	83.6	88.2	90.5	73.6	69.4	64.5	65.8	73.6	58.2	61.7
	van-b1		73.7	89.6	76.6	53.0	81.0	63.4	84.8	89.0	90.7	74.5	69.5	67.0	71.7	75.2	58.8	60.6
	van-b2		74.6	89.8	84.0	50.7	79.4	67.7	85.7	88.9	90.7	74.6	62.7	71.2	67.8	**76.7**	59.2	70.6
	van-b3		74.9	89.8	77.6	55.7	81.6	62.0	85.0	89.2	90.8	75.0	70.4	68.9	66.7	76.5	62.4	72.5
	van-b4		74.0	89.8	77.1	52.8	**83.2**	62.3	79.2	89.1	90.7	**76.9**	62.7	70.6	67.5	76.4	60.0	71.1
MViTv2	MViTv2-T	IN1K	75.2	89.7	**86.4**	57.6	82.0	68.3	**86.5**	89.3	90.7	74.4	70.2	69.1	73.5	75.4	59.6	55.7
	MViTv2-S		75.1	90.0	84.9	59.0	77.9	**69.2**	**86.6**	89.3	90.6	74.2	70.1	67.4	**74.4**	76.2	55.2	61.7
ConvNext	ConvNext-T	IN1K	74.1	89.4	81.6	54.0	81.2	66.2	84.9	89.2	90.8	74.9	69.8	64.7	71.0	74.2	57.5	62.7
	ConvNext-S		73.9	89.5	84.1	57.7	79.4	61.1	85.0	89.2	90.8	75.2	62.4	65.7	73.0	76.0	59.3	60.5
	ConvNext-B		75.0	89.7	77.6	55.3	81.1	63.2	85.7	89.2	90.8	76.8	69.5	69.4	72.8	76.5	58.7	68.6
	ConvNext-L		75.6	89.8	77.8	59.1	82.2	63.4	85.4	89.2	90.8	**77.0**	70.1	69.6	73.3	76.5	59.2	70.6
	ConvNext-B	IN22K	75.2	89.8	78.1	58.6	**83.5**	**70.3**	85.9	**89.4**	90.8	76.8	70.4	**71.8**	73.6	76.6	59.1	52.8
	ConvNext-L		**76.3**	89.8	78.0	**60.0**	82.0	**70.6**	86.0	**89.4**	90.8	76.0	70.7	69.2	74.3	**76.9**	60.0	70.7
	ConvNext-XL		75.6	90.1	78.7	**60.5**	81.0	63.0	85.4	89.3	90.8	76.8	70.8	70.3	73.3	76.4	59.5	68.8
SwinV1	Swin-T	IN1K	74.2	89.6	77.1	57.3	81.8	66.2	85.3	89.1	90.7	73.2	62.3	68.3	73.7	74.3	57.0	67.2
	Swin-S		75.1	89.5	85.4	58.4	76.1	61.9	85.2	88.8	90.6	74.7	69.8	67.1	**75.4**	75.4	57.7	70.9
	Swin-B		74.9	89.7	78.3	57.3	80.4	61.2	85.5	88.8	90.7	73.9	69.8	68.8	68.0	76.0	61.2	73.5
	Swin-B	IN22K	75.5	89.7	**86.5**	59.1	80.7	67.2	86.0	88.9	90.7	74.8	70.2	69.7	67.7	76.1	58.5	66.9
	Swin-B_w12		76.0	89.7	84.8	57.0	78.1	66.9	86.0	89.0	90.6	73.3	70.1	69.2	73.5	75.4	60.0	**76.3**
	Swin-L		75.8	89.9	83.2	**59.3**	80.1	67.6	86.1	89.2	90.7	76.1	70.3	68.8	67.7	76.2	58.9	72.1
	Swin-L_w12		**76.5**	89.7	**86.2**	58.1	78.8	66.8	86.0	89.1	90.7	**77.1**	70.2	69.3	**75.3**	76.2	58.9	**75.4**
SwinV2	Swinv2-T	IN1K	73.7	89.8	77.6	56.7	81.3	60.2	85.1	89.3	90.7	72.6	62.7	67.9	71.6	75.2	58.0	67.3
	Swinv2-S		73.0	89.7	78.7	57.4	75.9	61.7	79.2	89.0	90.7	73.3	62.7	66.5	68.3	76.1	59.4	66.3
	Swinv2-B		73.9	89.9	77.8	58.5	76.4	60.2	79.3	88.9	90.7	75.9	62.9	71.3	67.2	76.2	59.3	73.7
	Swinv2-B_w12	IN22K	74.9	89.9	77.4	56.2	80.6	61.5	**86.4**	89.3	90.8	75.9	70.3	69.9	67.5	76.3	60.4	71.0
	Swinv2-L_w16		75.0	90.0	77.0	53.2	**82.6**	64.3	79.4	89.3	90.8	76.0	70.3	**71.9**	68.7	76.6	60.2	**74.3**
RVSA	ViT-B	MillionAID	74.8	90.1	71.9	56.9	73.3	65.7	84.4	88.6	**90.9**	69.1	**89.0**	69.5	65.2	**76.6**	**66.0**	64.2
	ViT-B_kvdiff		75.2	**90.1**	72.1	57.1	74.5	65.9	84.6	88.3	**90.8**	71.8	**88.7**	71.3	68.1	76.0	65.1	63.6
	ViTAE-B		**76.3**	**90.2**	74.6	56.5	73.6	65.5	84.9	88.4	**90.9**	71.6	**88.8**	**77.8**	68.6	76.0	**67.0**	70.0
	ViTAE-B_kvdiff		74.0	**90.1**	71.4	54.7	75.3	65.0	85.1	89.4	90.8	69.9	88.2	70.9	62.7	75.5	**67.0**	53.4

The top three results are masked in red, green and blue.

that can effectively handle multimodal data. These existing data-driven FMs still have limitations in terms of data size, model structure, and learning strategies.

To address these challenges, we propose a brain-inspired framework for an RSFM, as illustrated in Fig. 14. This framework aims to integrate multimodal data in RS, such as image, video, point cloud, and text, and represent them uniformly for data-driven learning. Moreover, it incorporates prior knowledge, such as object spectral signature, road network information, and terrain and geographical location, into the model for knowledge-driven learning. By combining both data-driven and knowledge-driven approaches, we expect to enhance the model's performance and adaptability. More importantly, the brain-inspired properties can guide us to construct the model, represent the data, build learning algorithms, and process reasoning. Thus, in the following sections, we will delve into the key brain-inspired properties, focusing on four aspects: structure, perception, learning, and cognition.

B. *Basic Properties of Brain-Inspired RSFM*

1) Structure: The foundation of a functional model lies in its structure. Just as the human brain possesses a complex architecture to enable its comprehensive functions, we seek to

TABLE IX
NUMERICAL RESULTS OF COMPARISONS WITH FMS ON DIOR-R DATASETS WITH DIFFERENT TRAINING RATIOS

Method	Model	Pretrained data	mAP	APL	APO	BF	BC	BR	CH	DAM	ESA	ETS	GF
R50	R50	IN1K	52.7	60.9	14.7	71.0	80.5	27.2	72.1	18.8	56.5	48.0	56.3
PVTv2	pvtv2-b1	IN1K	63.5	70.3	35.8	80.1	81.4	40.5	72.5	29.1	77.9	67.3	71.6
	pvtv2-b2		66.7	71.8	46.9	79.9	81.3	44.2	80.0	36.5	79.8	70.0	77.1
VAN	van-b2	IN1K	66.2	79.6	42.3	80.5	81.4	44.6	80.1	36.1	79.2	70.5	69.4
MViTv2	MViTv2-T	IN1K	67.1	79.6	49.8	80.5	81.3	44.3	80.5	37.4	79.1	68.8	77.6
	MViTv2-S		68.1	80.7	50.3	80.6	81.1	44.3	81.1	37.3	79.9	70.4	**79.2**
ConvNext	ConvNext-T	IN1K	63.6	71.5	29.8	79.7	81.4	41.1	72.5	27.0	75.8	67.7	67.9
	ConvNext-S		66.2	71.1	40.6	79.5	81.3	43.6	80.4	36.5	79.3	69.3	75.9
	ConvNext-B		67.0	71.9	42.1	80.2	88.8	44.5	72.6	36.5	84.1	70.8	78.6
	ConvNext-L		68.1	**80.9**	42.6	80.3	**89.4**	44.8	80.4	39.2	79.4	71.2	78.4
	ConvNext-B	IN22K	68.5	80.7	50.0	80.3	81.3	45.6	81.3	37.9	85.2	71.5	77.6
	ConvNext-L		**70.1**	**81.4**	50.9	80.6	**89.6**	**47.4**	81.3	38.6	86.8	**77.7**	78.5
SwinV1	Swin-T	IN1K	65.2	79.9	37.7	80.1	88.0	38.9	72.5	35.1	77.7	68.1	75.3
	Swin-S		66.8	71.5	44.3	80.5	**89.2**	44.3	72.7	36.5	80.0	70.5	78.4
	Swin-B		67.1	71.9	43.3	80.0	89.0	45.1	80.3	38.3	79.7	70.8	78.2
	Swin-B	IN22K	69.2	80.5	50.9	80.8	88.7	45.4	81.3	**41.2**	87.5	71.1	77.8
	Swin-B_w12		69.0	80.8	**51.0**	80.8	81.2	47.1	**81.4**	**41.2**	87.4	**76.0**	78.6
	Swin-L		**70.1**	80.7	**52.8**	**81.1**	89.0	46.4	**81.6**	**40.8**	**88.6**	**76.2**	**79.1**
	Swin-L_w12		69.8	79.7	**58.2**	**80.9**	81.2	46.4	**81.4**	40.6	**88.4**	75.7	**78.8**
SwinV2	Swinv2-T	IN1K	64.1	71.1	33.7	79.0	81.4	41.8	72.6	34.3	78.3	67.1	76.4
	Swinv2-S		66.9	80.2	44.7	79.7	81.3	44.2	79.3	38.4	80.0	70.4	78.5
	Swinv2-B		66.6	72.1	41.9	80.6	88.4	45.0	72.7	39.0	79.8	70.9	77.4
	Swinv2-B_w12	IN22K	67.2	72.3	43.9	80.4	**89.2**	45.5	80.9	37.2	85.6	71.6	76.3
	Swinv2-L_w16		67.8	80.5	43.9	80.8	81.3	45.3	81.0	34.4	86.2	71.5	77.2
RVSA	ViT-B	MillionAID	67.3	80.4	41.3	80.6	81.3	46.1	72.6	32.7	86.7	70.0	72.6
	ViT-B_kvdiff		67.0	80.4	40.7	80.5	81.4	46.4	72.6	29.1	84.4	69.2	74.1
	ViTAE-B		**69.9**	**81.2**	50.3	**80.9**	86.9	**50.3**	79.1	36.5	**88.2**	73.8	76.6
	ViTAE-B_kvdiff		68.5	80.3	41.0	80.1	81.3	**47.3**	77.4	36.7	87.3	70.8	75.6
				GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM
R50	R50	IN1K		72.3	16.9	45.9	80.3	49.8	61.3	81.0	38.8	41.5	60.8
PVTv2	pvtv2-b1	IN1K		83.1	33.6	56.8	81.0	78.5	61.5	81.5	54.8	46.8	65.2
	pvtv2-b2			82.7	41.6	58.5	81.1	79.8	68.8	81.3	59.0	48.1	65.3
VAN	van-b2	IN1K		83.5	40.7	58.2	81.1	76.3	62.1	81.4	62.9	49.3	65.6
MViTv2	MViTv2-T	IN1K		83.2	38.8	58.5	81.1	80.8	60.9	81.4	64.0	48.3	65.2
	MViTv2-S			84.0	41.9	58.6	81.1	81.4	68.7	81.4	64.2	49.3	65.4
ConvNext	ConvNext-T	IN1K		84.1	36.7	55.6	81.0	80.3	70.7	81.3	56.1	47.2	64.9
	ConvNext-S			**84.2**	41.1	58.4	81.1	76.9	70.5	81.3	63.2	44.0	66.1
	ConvNext-B			**84.4**	41.3	58.6	81.1	77.4	70.0	81.4	64.1	44.1	**66.5**
	ConvNext-L			**84.2**	43.2	58.8	**81.2**	78.2	70.5	81.4	62.3	49.9	**66.2**
	ConvNext-B	IN22K		83.7	43.9	59.7	81.1	77.0	70.6	81.4	65.9	49.2	65.9
	ConvNext-L			83.2	**45.3**	59.7	**89.3**	76.9	70.7	81.5	**67.1**	**50.0**	**66.2**
Swin	Swin-T	IN1K		83.2	35.0	56.1	81.0	79.8	70.7	81.5	55.6	43.3	63.8
	Swin-S			83.8	42.3	57.9	81.1	77.9	70.6	81.4	65.8	43.6	64.2
	Swin-B			**84.7**	43.1	58.5	81.0	**82.9**	62.6	81.4	62.0	43.7	65.1
	Swin-B	IN22K		83.3	44.5	59.1	81.0	77.3	71.0	81.4	**66.9**	48.7	65.0
	Swin-B_w12			82.8	43.7	59.2	81.0	76.7	70.8	81.4	66.0	47.8	65.2
	Swin-L			83.6	44.0	60.0	81.0	**82.9**	**71.1**	81.4	65.7	49.5	65.6
	Swin-L_w12			83.4	**45.4**	59.6	81.0	82.6	70.9	81.3	**66.8**	49.1	65.2
SwinV2	Swinv2-T	IN1K		81.7	37.5	56.1	80.9	77.2	70.4	81.4	54.6	43.4	63.9
	Swinv2-S			82.6	40.0	58.5	81.0	77.3	70.0	81.5	62.8	43.7	64.5
	Swinv2-B			83.1	40.4	58.7	81.0	77.4	70.4	81.4	63.3	43.8	65.0
	Swinv2-B_w12	IN22K		76.0	42.9	59.1	81.0	77.7	70.6	81.5	62.7	43.9	65.3
	Swinv2-L_w16			82.7	44.0	59.1	**81.2**	82.8	70.3	81.5	63.5	44.0	65.2
RVSA	ViT-B	MillionAID		82.1	41.3	**60.3**	81.2	80.3	**71.1**	**89.9**	62.2	49.7	64.1
	ViT-B_kvdiff			81.9	41.3	59.7	81.2	81.6	70.5	**89.9**	61.0	49.8	64.8
	ViTAE-B			83.3	44.5	**60.8**	**81.2**	**83.6**	**71.1**	**89.9**	63.2	**50.3**	65.4
	ViTAE-B_kvdiff			82.3	**44.9**	**60.4**	**81.2**	82.8	70.8	89.4	66.7	**50.2**	62.8

The top three results are masked in red, green and blue.

design a model with similar characteristics. In this section, we explore the brain's spiking structure, diversity, and geometry.

Spiking: The human brain, consisting of 86 billion neurons, communicates through highly structured connections called synapses [265]. Neurons exchange information in a sparse and asynchronous manner through discrete action potentials or "spikes" [266]. To emulate this essential characteristic of the human brain, the spiking neural network (SNN) was introduced, as shown in Fig. 15. Unlike the traditional neural networks, SNN processes sparse spatiotemporal signals by simulating the excitation and inhibition of neurons. Spiking neurons receive spikes developing the membrane potential through time-following differential equations. A spike is emitted when the membrane potential crosses a threshold [264]. This approach offers advantages in terms of analog computing, low-power consumption, fast reasoning, event-driven processing, online learning, and large-scale parallelism, as it has demonstrated superior performance [267].

Diversity: The brain's composition is not uniform; it relies on various neuron types to achieve its complex functions. The cerebral cortex, for example, is organized into four major structures: the occipital lobe, temporal lobe, parietal lobe, and frontal lobe

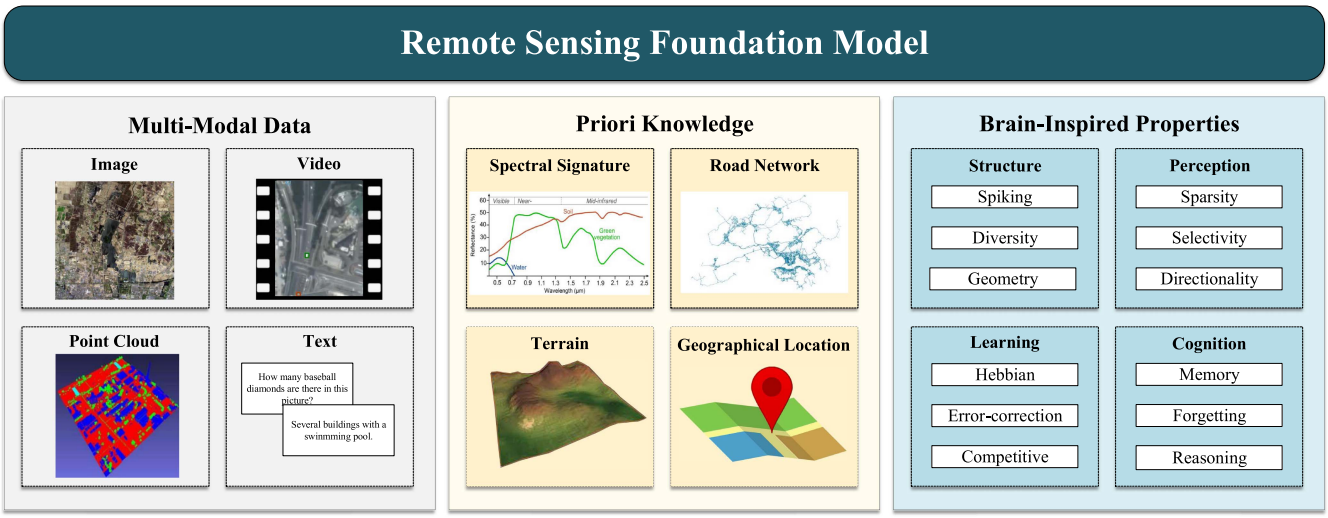


Fig. 14. Overall framework of the brain-inspired RSFM.

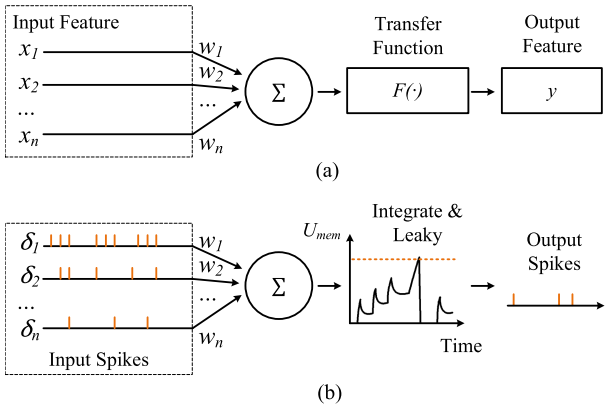


Fig. 15. Illustration of (a) classical artificial neural unit and (b) spiking neurons. The image is reproduced following [264].

[268]. Recent studies have revealed a rich diversity of neurons in the brain. For instance, Yao et al. [269] analyzed over 500 000 individual cells in the mouse primary motor cortex and identified 56 highly replicable neural types. Ed Lein et al. demonstrated the abundance of neuronal species in the cerebral cortex [270] using techniques, such as patch clamp. This diversity is essential for the realization of different modalities and functional differentiations in the brain. Similarly, our neural network needs to incorporate a variety of neuron types to facilitate the realization of different tasks and modalities in the model.

Geometry: Traditionally, the brain's complex functions were thought to arise from intricate inter-regional connections. However, neural field theory suggests that the brain's geometry may represent a more fundamental dynamical constraint. Studies by Caucheteux et al. [271] using human magnetic resonance imaging data demonstrated that cortical and subcortical activity can be understood as arising from the excitation of fundamental resonant modes of the brain's geometry (i.e., its shape). This implies that geometric constraints play a crucial role in shaping brain functions in addition to neural connections. Therefore, our model should also take into account the potential role of geometric constraints in shaping functions.

Discussion: SNNs have been studied a lot in deep learning. Yao et al. [272] proposed attention SNN. It integrates the attention mechanism into a million-scale SNN. On the ImageNet-1K dataset, it has achieved performance equivalent to that of the traditional artificial neural networks for the first time, and its theoretical energy efficiency is 31.8 times that of the artificial neural networks with the same structure. Therefore, brain-inspired SNNs have many potentials. For large-scale basic models, SNNs will have more potential. However, neuron diversity and geometric constraints have not been studied in the current model. The model's functional design of different neurons and geometric constraints combined with dynamics research will help improve the robustness of the basic model.

2) Perception: Perception is the process through which humans obtain information from the external world. For the brain, this input information includes visual, auditory, tactile, and other sensory data. Similarly, in RS, different data, such as visible light and SAR, provide multimodal information. To design an effective FM, we need to mimic the human brain's characteristics, such as sparsity, selectivity, and directionality, to enhance the model's efficiency in perceiving information.

Sparsity: The brain exhibits a hierarchical, sparse, and periodic structure [273]. Sparsity plays a crucial role in biological brains as it allows for the representation and processing of information using only a small number of activated neurons or saliences. This sparsity is a property of neural coding that enhances the brain's efficiency, robustness, and flexibility. Studies have shown that sparse representations in the cerebral cortex's V1 may satisfy the optimality criteria of information theory [274]. As we move to the higher levels of neurons, the receptive fields become larger, and the sparsity becomes stronger. Recent research has also indicated that neural circuits are organized in

a sparse yet efficient manner. Higher levels of intelligence have been associated with more direct information processing and less cortical activity during reasoning, highlighting the importance of sparsity in the brain's efficient perception [273].

Selectivity: Selectivity, often referred to as attention mechanism, is a key feature of the brain's ability to focus on specific objects and control areas of attention [275]. The brain receives a vast amount of information simultaneously but cannot process it with equal priority. Therefore, it employs selective attention to filter and prioritize information [276]. Selective attention exists widely in the human visual system and is regulated by both bottom-up and top-down mechanisms. Bottom-up selectivity responds to salient stimuli from the environment, such as changes in target brightness or motion. On the other hand, top-down selectivity allows humans to process relevant information based on the current behavior and intentions while ignoring irrelevant information, forming a close integration between attention and cognition [277].

Directionality: Directionality is the brain's ability to perceive its own position and orientation. The brain has azimuthal and oblique angle cells that provide orientation and position information. When the head faces a particular direction, the corresponding direction cells are activated [278]. Neural coding patterns of egocentric spatial orientation have been discovered in the medial temporal lobe of the human brain, supporting vector representations of egocentric spaces by encoding distances to reference points [279]. Building a multiscale directional network in the FM aligns with the biological basis and significance of directionality in perception.

Discussion: Overall, the application of sparsity, directionality, and selectivity in deep learning can lead to more efficient and effective neural networks that are better able to generalize and learn from data. Child et al.'s [280] sparse transformers can be used in long sequences for better performance in density modeling. Networks based on multiscale geometric structures, such as Ridgelet neural network and contourlet neural network, all use characteristics, such as directionality, to achieve a more sparse representation [43]. These methods can provide theoretical support for the perception of brain-inspired FMs.

3) Learning: Learning is a fundamental process for humans to acquire memories, knowledge, and practical skills. From a neuroscience perspective, neurons possess plasticity, enabling them to learn by modifying their connections and weights [281]. In this section, we introduce three brain-inspired learning models: Hebbian learning, error-correction learning, and competitive learning.

Hebbian Learning: In 1949, Hebbian presents a postulate: "cells that fire together wire together." This means that when an axon of cell is near enough to excite another cell or repeatedly or persistently takes part in firing it, the connection weight between these two cells will be increased. Inspired by this postulate, the correlation-based learning rules are generally called Hebbian learning [283]. Hebbian learning is a form of unsupervised learning, as it does not rely on external feedback or error signals. Instead, it captures statistical correlations between inputs and outputs, forming associative memories.

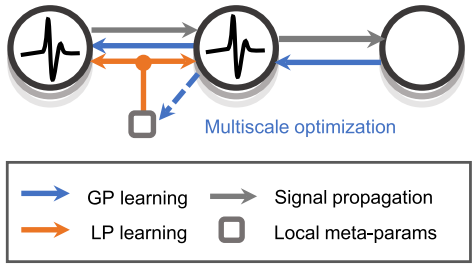


Fig. 16. Learning paradigm integrating global error-driven learning (called "global plasticity,") and local correlation-driven learning (called "LP,"). Image from [282].

Error-Correction Learning: Error-correction learning follows the principle that neurons should learn from their mistakes. In this form of learning, the connection weight between two neurons is adjusted based on the difference between the actual output and the desired output (i.e., the error). Error-correction learning is a form of supervised learning, as it requires a teacher or target to provide the desired output. By minimizing the error, this learning mechanism improves the performance of the network. The most popular learning algorithm for use with error-correction learning is the backpropagation algorithm [284]. Through setting the target loss, the gradient can be backpropagated across the neural network and the weight can be updated [285].

Competitive Learning: Competitive learning operates on the principle that neurons should compete for activation. Only a few neurons in a layer are allowed to be active at a time, while the others are inhibited. The weights of connections between inputs and active neurons are strengthened, whereas the weights of connections between inputs and inactive neurons are weakened [286]. Similar to Hebbian learning, competitive learning is unsupervised and does not rely on external feedback. It enables the discovery of features or clusters in the input data, leading to the formation of sparse representations. Activation learning [287] is a type of competitive learning. It injects competition within and among neurons and show the capacity of learning plentiful local features from few shots of input patterns.

Discussion: The learning method represented by error-correction learning of backpropagation has excellent performance on specific tasks. Therefore, it occupies the mainstream position in the current model learning. But its supervised learning is still far from the way the brain learns. Wu et al. [282] proposed a neuromorphic global–local synergic learning model, as shown in Fig. 16. It can metalearn local plasticity (LP) and receive top-down supervision information. By combining different learning methods, the ability of the model in few-shot learning, continual learning, and other scenarios is improved. Therefore, organically combining a variety of brain-inspired learning methods into the learning of the FM will help improve the learning efficiency.

4) Cognition: Cognition refers to the process through which human beings acquire knowledge, apply that knowledge, and

process information, representing the most fundamental psychological process in human beings. The human brain receives information input from the external world, processes it, converts it into internal psychological activities, and then controls actions accordingly. In this section, we will primarily introduce the mechanisms of brain memory, forgetting, and reasoning.

Memory: Memory is fundamental to human beings' ability to solve complex problems. The human brain can encode and store past experiences to form memories, which can be searched and retrieved when needed. Engram cells serve as fundamental evidence that the brain forms memories, providing the necessary conditions [288] for their emergence. Based on their differentiation and actions, memories can be divided into short-term memory and long-term memory.

Short-term memory, also known as working memory, usually lasts only seconds or minutes [289]. Two theories about short-term memory are "activity-silent neural networks" and "sustained activity." Activity-dependent synaptic plasticity enables the formation of a transient nervous system [290], leading to transient increases or decreases in neurotransmitter signals on synapses, forming a dynamic neural network [291]. Short-term memory is formed through the strengthening or weakening of these transient signals. The theory of sustained activity suggests that short-term memory is maintained by continuous action potential discharge. Studies by the authors in [292] and [293] have also shown functional magnetic resonance imaging and electroencephalography results indicating a sustained increase in brain activity during the delayed period of memory tasks.

The hippocampus, located in the temporal lobe, is a crucial part of the limbic system responsible for forming new memories and converting short-term memory into long-term memory in the human brain. This process, often called consolidation, involves gene activation and the formation of new synaptic connections between neurons in the brain. Research by Yap et al. [294] revealed that the hippocampus expresses sparse populations of neurons activated by novel experiences. These neurons may fine-tune their inputs to form persistent networks that provide a coordinated response to an experience, leading to long-term memory consolidation.

Forgetting: Forgetting is the opposite process of remembering. Forgetting occurs for different reasons and occurs at different stages of memory formation, storage, and retrieval. Currently known forgetting mechanisms include passive forgetting and active forgetting [295]. Passive forgetting is the nonspontaneous memory loss process of the human brain. Over time, memories become difficult to retrieve without context. In addition, similar memories can form interference and, thus, be lost. Finally, the instability of the biological memory mechanism will cause the memory to fade naturally over time. In contrast, active forgetting is considered to be a spontaneous memory extinction process. Active forgetting usually includes motivated forgetting and retrieval-induced forgetting. Motivated forgetting refers to the forgetting process that is under our own cognitive control. For example, when certain memories affect one's positive image or are inconsistent with beliefs, motivational forgetting will actively abandon this part of the memory. Retrieval-induced forgetting describes that when memory is consolidated through learning, it may weaken the same type of memory that has not been practiced, and only retrieval of certain memories may cause the forgetting of other memories. All in all, forgetting plays a vital role in human beings. It allows us to focus on what is retained in memory and to protect ourselves from adverse memories.

Reasoning: As a complex intelligent system, the brain's causal inference ability is one of the main manifestations of its intelligence. When the brain processes multisensor information, it exhibits the ability to make causal inferences. In particular, in situations where information from multiple sensors differs, the brain is able to infer whether the signals come from the same source or independent sources and does not integrate signals that are unlikely to come from the same source. Reuben et al. [296] show how interactions between different types of neurons lead to optimal integration and causal inference. Many neurons that receive input from both modalities are congruent neurons with similar tuning for both modalities, enabling multisensory integration. There are also heterotropic neurons, capable of detecting signals from different sources. The collaboration of coherent and heterotropic neurons may be what enables the brain to form causal inferences.

Discussion: Memory and forgetting play an important role in the cognitive process. Memorizing can help the model form knowledge, while forgetting can clear unnecessary information. Memory-based models have been heavily proposed [297] in current time-series modeling. In addition, reasoning has also received extensive attention, and a large number of models with reasoning have been proposed [298]. However, the current basic model still does not involve the exploration of these cognitive abilities, and its performance on complex tasks is still very limited.

VII. Open Problems

As the basis of RS field, the FMs have received extensive attention and research. There are still many challenging open problems to be solved in this field. In this section, we analyze 12 open problems about RSFMs and propose potential solutions, as shown in Fig. 17.

A. Brain-Inspired FMs

Neural networks have their roots in the study of the brain's neuron structure [299]. While AI research has made significant progress, the current FMs still fall short of capturing the brain's remarkable capabilities. The brain is exceptionally complex, yet it can achieve functions, such as recognition, cognition, and decision making, while consuming minimal power. Developing brain-inspired FMs can drive AI toward higher performance and lower costs.

The design of brain-inspired FMs can be approached from two main perspectives: brain structure and brain characteristics. The SNN, for example, emulates the human brain's activation and inhibition of signals through accumulation [300]. Another approach, the capsule network, simulates a set of neurons and uses vectors to represent feature pose information [301]. The brain's inherent characteristics, such as sparseness, selectivity,

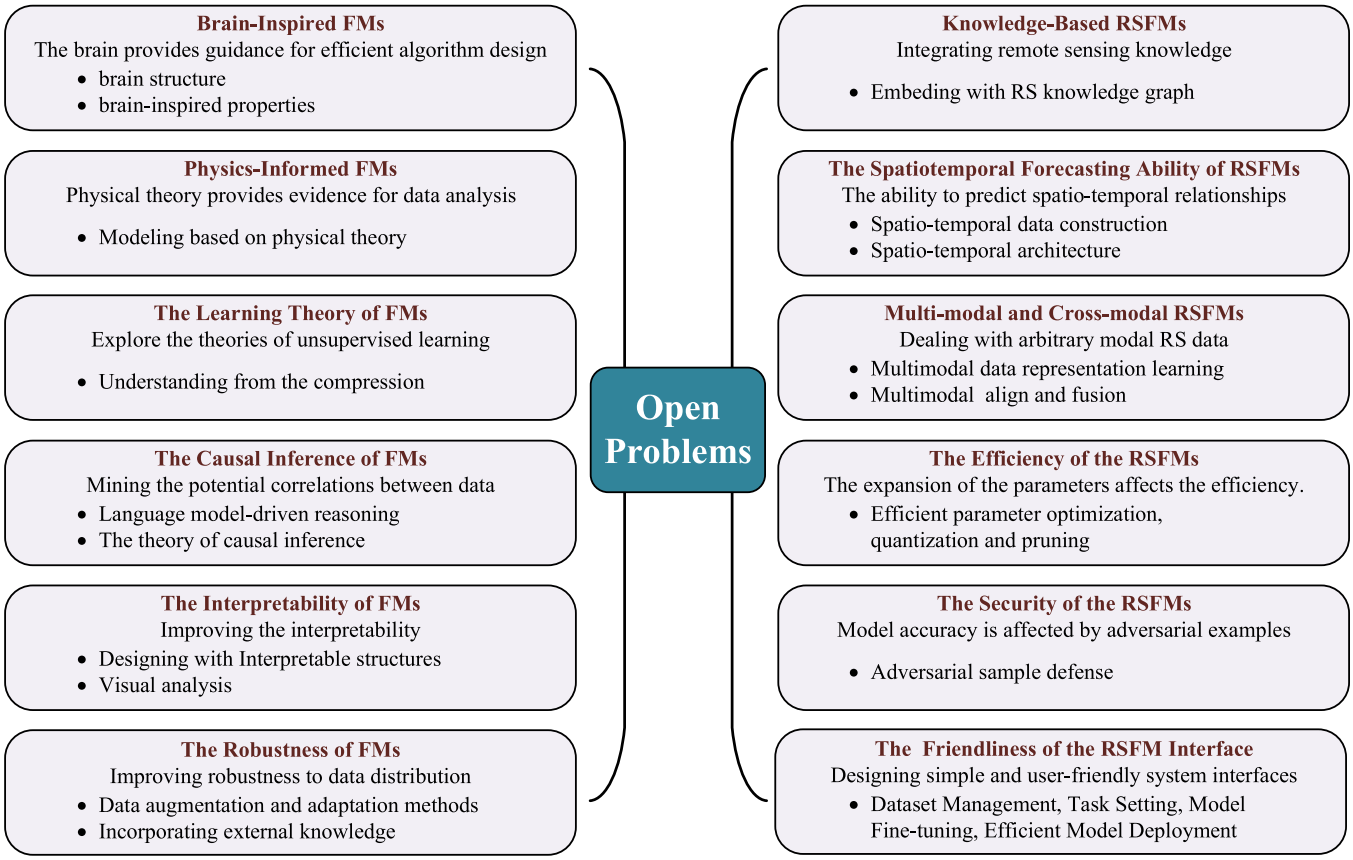


Fig. 17. Top 12 open problems of FMs.

and directionality, offer valuable insights for designing foundational models that can better mimic the brain's efficiency and effectiveness.

B. Physics-Informed FMs

In tasks related to RS interpretation, the data are collected from the physical world. However, current FMs often focus solely on data-driven approaches, neglecting the underlying physical characteristics of the data. By incorporating these physical characteristics, we can effectively unearth latent features within the data and extract robust and sparse representations.

A prime example of physics-informed data interpretation is the use of the Fourier transform for the spectrum analysis of 1-D signals and wavelet transform for 2-D signals. These methods realize a sparse representation of a complex signal. Yang et al. [134] employed lifting wavelet to extract multiscale frequency-domain features, leading to more robust feature representations. Wave MLP represents each token as a wave function with phase and amplitude in quantum mechanics, which enhances token representation [302]. In addition, Li et al. [303] proposed utilizing electromagnetic scattering information for SAR automatic target recognition. This approach characterizes the electromagnetic scattering properties of the target, providing crucial physical structure information. By integrating the existing physical research, it becomes possible to effectively describe target characteristics and improve FM performance.

C. Learning Theory of FMs

A typical characteristic of the FM is that it is trained using SSL, so the model is task agnostic [5]. However, the learning theory of the FMs is not well studied. SSL is a type of unsupervised learning. When a model is optimized for a certain goal (such as image reconstruction), the model can also achieve excellent performance on an unoptimized goal (such as image classification). The results obtained in this experiment do not have as good theoretical support. Therefore, it is an important issue to study the learning theory of the FM.

Sutskever et al. [304] proposed the use of compression to explain unsupervised learning. He regarded the information obtained by unsupervised learning as the common structure noticed by the compressor. The better the compressor, the more common structures it can extract. Correspondingly, the common patterns learned by a model in the data of unsupervised tasks can be used to help perform supervised tasks. Chen et al. [305] demonstrated that a good sequence predictor can also achieve powerful unsupervised learning in the image domain. The current theoretical research on FMs is still in the initial stage of development, but there is still a long way to go to theoretically guide the design of learning algorithms.

D. Causal Inference of FMs

Despite the significant breakthroughs of FMs, the current pretraining models face challenges in learning potential causal

relationships within data. The existing research reveals a critical shortcoming of correlation-based FMs in terms of causal inference [306]. However, systems, such as AutoGPT and BMTools [307], have shown promising potential by enabling models to learn how to call various tools to complete tasks.

Incorporating the underlying theory of causal inference into FMs can greatly enhance their understanding. For instance, the structural causal model allows for counterfactual reasoning, enabling inferences about how other variables would be affected if certain variables were to change. In addition, the structural causal model can determine the conditional independence between variables, assessing whether other variables remain independent when certain variables' values are given [308]. By combining FMs with the reasoning capabilities of the structural causal model, we can significantly improve the model's ability to analyze potential variable relationships.

E. Interpretability of FMs

Improving model interpretability is a crucial step toward understanding the internal algorithm logic of FMs. Utilizing interpretable model structures can be an effective approach in this regard. One such technique is wavelet decomposition, which can decompose features from the frequency domain. This allows for the learning of interpretable geometric texture features, making it easier to understand how the model processes and represents different patterns in the data [309]. By incorporating wavelet decomposition into the model, researchers can gain insights into the specific features and characteristics that the model focuses on during the interpretation process.

Furthermore, constructing models based on the existing physical principles can also enhance interpretability. Wu et al. [310] treat the forward process of neural network calculations as a diffusion process from a given initial state, explicitly utilizing dependencies between samples during forward calculations. By doing so, they explicitly utilize dependencies between samples during forward calculations, which can help in better understanding how the model propagates information and makes decisions based on the input data.

In addition, the attention correlations in the transformer structure can be leveraged to visualize feature associations. By visualizing these associations, researchers can gain insights into which parts of the input data are most relevant to the model's predictions. This enhances the interpretability of the model to some extent, providing a glimpse into the decision-making process of the FM.

F. Robustness of FMs

The training of FMs on large and diverse unlabeled datasets brings advantages in terms of generalization and feature extraction capabilities. However, it is practically impossible to collect a dataset that covers all possible distributions of scenes, making the robustness of the model closely related to the distribution of the training data [5].

To enhance the robustness of FMs, data augmentation is a simple yet effective method. Applying various transformations to the training data can introduce diversity into the training dataset and help the model learn to handle different variations of the data. The model can avoid overfitting to specific patterns and improves the ability to handle unseen data [83].

Moreover, adaptation methods are valuable for enhancing the robustness of FMs when applied to downstream tasks. Fine-tuning a small number of parameters while freezing others in the model can improve the model's performance on out-of-distribution samples. This process allows the model to adapt to the specific characteristics of the target task while retaining the general knowledge learned during pretraining [5].

In addition, incorporating external knowledge into the model's input can further boost its robustness. By providing relevant information as additional input, such as prior knowledge about the scene, domain-specific attributes, or environmental context, the model gains a broader understanding of the data and handles diverse and complex scenes well.

G. Knowledge-Based RSFMs

The capability of FMs in visual language has significantly advanced due to large-scale image and text alignment data, enhancing their understanding and analysis abilities. However, in the RS field, collecting abundant image and text-paired data for training is challenging. To address this, Deng et al. [311] created a dataset called GeoSignal, enabling fine-tuning of large language models specifically for Earth science-related queries. Similarly, fusing RS interpretation with large language models requires constructing fine-tuning datasets with expert knowledge to enhance the model's accuracy.

While underlying models demonstrate impressive performance, they are often limited in capturing and exploiting common-sense errors, leading to potential risks. To improve the reliability and interpretability of FMs, incorporating knowledge graphs can be advantageous [312]. Knowledge graphs serve as structured knowledge databases that provide additional information for model reasoning. By combining the FM with knowledge graphs, the model gains access to key task-related knowledge through in-contextual learning without requiring retraining. The FMs robust knowledge understanding and processing capabilities can further enhance its accuracy and performance.

H. Spatiotemporal Forecasting Ability of RSFMs

FMs driven by large-scale data have shown remarkable capabilities in feature extraction and understanding complex images. However, current RSFMs, using masked image modeling for SSL, fail to fully exploit the potential time-varying relationships in time-series data. Consequently, a significant amount of data remains untapped. To address this limitation, exploring a multitemporal FM pretraining method can be beneficial. By training the model with input from RS time-series data, we can analyze the dynamic temporal information within the data. This approach enables the FM to perform spatiotemporal analysis and prediction, significantly enhancing its ability to monitor disasters and other time-sensitive phenomena.

I. Multimodal and Cross-Modal RSFMs

RS data exhibits a wide variety, with different sensors providing data with varying temporal and spatial references and formats. This diversity makes it challenging to collect and align high-quality multimodal sample data, thereby limiting the development of multimodal RSFMs. One potential solution is to establish a standardized interface, which would enhance the efficiency of data aggregation.

Furthermore, different tasks often exploit distinct physical properties. For example, sound separation requires frequency information, while sound content recognition relies on timing information. Therefore, the data feature extraction models for different modalities need to be constructed based on their unique characteristics [313]. Introducing texture features, scattering center features, phase coherence features, spectral correlation features, etc., can effectively process multimodal data. In addition, selecting suitable feature spaces, such as Euclidean space, Hilbert space, and Unitary space, can help mitigate information loss.

To address the challenge of inconsistent data space–time benchmarks and alignment difficulties, an independent feature extraction network can be designed for each modality [131]. Subsequently, a small set of alignment features can be utilized to align the features, enabling FMs to work effectively across different data modalities. This approach ensures that the extracted features retain their unique properties while being compatible across different modalities.

J. Efficiency of RSFMs

As the amount of data and model parameters increases, there is growing concern about the training and reasoning efficiency of FMs. Large-scale language models have already achieved training with billion-scale parameters, and the number of parameters in RS vision FMs is expected to continue rising. However, the processing of vast amounts of RS image data requires significant computational resources. Therefore, the PEFT algorithms can help reduce training overhead.

Quantization and pruning are also essential techniques to consider. By reducing the precision of individual weight numerical representations, the model's operating efficiency can be significantly improved. For instance, transformer models can be quantized from FP32 to INT8, a widely adopted practice in training large language models. Pruning, on the other hand, involves removing elements of the network, ranging from individual weights to higher granularity component channels. Researchers, such as Spyrison et al. [314], have proposed sparse pruning methods, enabling GPT-series models to achieve 50% sparsity in a single pass without the need for retraining. In addition, research has demonstrated full parameter fine-tuning of a 65 billion parameter model on eight 3090 GPUs [315]. As such, continuously integrating the latest technology with RSFMs can improve training and reasoning efficiency while saving valuable computational resources.

Integrating the latest technology advancements with RSFMs can lead to improved training and reasoning efficiency, allowing for more efficient use of valuable computational resources. By continually exploring and implementing parameter-efficient techniques, researchers can ensure that RSFMs remain scalable, powerful, and capable of handling the ever-increasing volume of RS data with minimal computational overhead.

K. Security of RSFMs

The issue of hallucination in large language models is indeed a significant concern [42], especially in the visual domain, where it can resemble the phenomenon of adversarial examples. In certain cases, the model may confidently predict incorrect information or assign high probabilities to nontarget areas, which can lead to serious consequences in practical applications, particularly in RS, where model misidentification can result in critical decision-making errors.

Adversarial sample defense has been explored by researchers to enhance the security of models, both in general contexts [316] and specifically in RS [317]. However, given the large number of parameters in FMs, the effectiveness of conventional adversarial attack mitigation strategies may require further investigation and development.

Moreover, while adversarial attacks are commonly studied and implemented on digital images, it is vital to consider the potential impact of such attacks on high-resolution RS images. In RS, even minor local changes in real scenes could lead to significant model misidentification, highlighting the need for robust defense mechanisms that can handle such real-world variations.

L. Friendliness of RSFM Interfaces

RSFMs offer a significant advancement in the application of RS interpretation. However, their adoption still presents a high threshold for users. To promote the widespread use of RSFMs, it is crucial to design simple and user-friendly system interfaces. These interfaces should facilitate various RS interpretation applications and enable RSFMs to adapt with only a small number of samples.

The RSFM system should support the following key functionalities to enhance user experience.

1) *Dataset Management:* The system should allow users to easily upload and construct datasets. This feature enables users to input their own data for specific tasks, making it convenient to work with their own RS data.
2) *Task Setting:* Users should be able to set up interpretation tasks effortlessly. The interface should provide intuitive options to define the specifics of the tasks they want to perform using RSFMs.
3) *Model Fine-Tuning:* RSFMs should offer users the ability to fine-tune models with their dataset to achieve better performance on specific tasks. This fine-tuning process should be straightforward and require minimal expertise.
4) *Efficient Model Deployment:* Once the model is ready, the system should enable high-precision and efficient deployment. Users should be able to deploy their customized RSFM models quickly and easily for practical applications.

By providing a user-friendly interface, RSFMs can be readily applied to a broader range of scenarios and tasks. Users, even those with limited expertise in RS or deep learning, can utilize RSFMs to accomplish their interpretation objectives through simple and efficient operations. This will significantly enhance the practicality and accessibility of RSFMs in the field of RS interpretation.

VIII. Concluding Remarks and Discussion

FMs have emerged as a promising direction in RS research. In this article, we provided a comprehensive survey of the current development of RSFMs. We started by explaining the key technologies underlying FMs, including transformer structures, self-supervised pretraining methods, and efficient parameter optimization techniques. Then, the latest developments in FMs across various domains are presented, including language, vision, visual language, and RSFMs. We explored core applications in RS interpretation, including classification, location, and understanding tasks.

After that, performance comparison experiments are conducted from three aspects: global representation, local representation, and target localization. Through the experiments, we observed that while RSFMs demonstrate potential, they still face challenges in achieving significant advantages over natural FMs due to limited RS data and certain structural design limitations.

Through the above research and analysis, this article summarizes the research and development of the RS FM. From the perspective of the development process of the FM, models are updating quickly. As mentioned in this article, a model, such as metatransformer [131], applies the data with 12 modalities to one FM, and similar research will quickly follow up in the field of RS. In the near future, multimodal and cross-modal RSFMs will receive a lot of research. A large amount of data in the field of RS will be more fully mined. The barriers of multimodal data will also be gradually broken down.

However, the large-scale computing behind these studies means that this form of research is difficult to follow. A large number of calculations and lack of theoretical research support will be the shortcomings of the current RSFM. To this end, this article further elaborates a valuable research direction of the FM, that is, the brain-inspired RSFM. Different from the current research ideas of the FM, the brain-inspired properties will provide FMs with a theoretical foundation from a biological background, reliable performance, and higher data utilization efficiency. This framework provides a novel perspective to guide the development of future models and applications in RS interpretation.

Finally, we identified 12 open problems in RSFM research, encompassing areas, such as brain-inspired modeling, physical information integration, and knowledge-based learning. Addressing these open problems will drive the proposal and adoption of innovative methods in RS interpretation.

In summary, RSFMs hold immense potential and continue to be an active area of research. By addressing the identified challenges and exploring new avenues inspired by brain characteristics, we can unlock the full potential of RSFMs.

References

[1] 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. 16000–16009.

[2] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12104–12113.

[3] M. Dehghani et al., "Scaling vision transformers to 22 billion parameters," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 7480–7512.

[4] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[5] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[6] 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 Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[7] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.

[8] C. Schuhmann et al., "LAION-400M: Open dataset of clip-filtered 400 million image-text pairs," in *Proc. NeurIPS Workshop Datacentric AI*, 2021.

[9] C. Xinlei, X. Saining, and H. Kaiming, "An empirical study of training self-supervised visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9620–9629, doi: 10.1109/ICCV48922.2021.00950.

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

[11] L. Jiao et al., "Brain-inspired remote sensing interpretation: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2992–3033, Feb. 2023.

[12] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[13] L. Jiao et al., "Transformer meets remote sensing video detection and tracking: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1–45, Jun. 2023.

[14] C. Wen, Y. Hu, X. Li, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," 2023, *arXiv:2305.05726*.

[15] G. Mai et al., "On the opportunities and challenges of foundation models for geospatial artificial intelligence," 2023, *arXiv:2304.06798*.

[16] V. C. Gomes, G. R. Queiroz, and K. R. Ferreira, "An overview of platforms for big Earth observation data management and analysis," *Remote Sens.*, vol. 12, no. 8, 2020, Art. no. 1253.

[17] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315, doi: 10.1109/TGRS.2022.3222818.

[18] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5612822, doi: 10.1109/TGRS.2022.3194732.

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

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

[21] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "GFM: Building geospatial foundation models via continual pretraining," 2023, *arXiv:2302.04476*.

[22] 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. 4088–4099.

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

[24] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 22243–22255.

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

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

[27] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9694–9705.

[28] 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. [Online]. Available: https://openreview.net/forum?id=Ee277P3AYC

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023.

[30] 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.

[31] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[32] 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, May 2022, Art. no. 5608020.

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

[34] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[35] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3558–3568.

[36] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[37] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[38] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.

[39] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[40] 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, Apr. 2021.

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

[42] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, 2023, Art. no. 248.

[43] L. Jiao, R. Shang, F. Liu, and W. Zhang, *Brain and Nature-Inspired Learning, Computation and Recognition*. Amsterdam, The Netherlands: Elsevier, 2020.

[44] S. Schmidgall et al., "Brain-inspired learning in artificial neural networks: A review," 2023, *arXiv:2305.11252*.

[45] X.-L. Zou, T.-J. Huang, and S. Wu, "Towards a new paradigm for brain-inspired computer vision," *Mach. Intell. Res.*, vol. 19, no. 5, pp. 412–424, 2022.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017.

[47] Y. Yang et al., "Transformers meet visual learning understanding: A comprehensive review," 2022, *arXiv:2203.12944*.

[48] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, 2019.

[49] D. Zhang and D. Wang, "Relation classification: CNN or RNN?," in *Proc. Int. Conf. Comput. Process. Oriental Lang.*, 2016, pp. 665–675.

[50] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[51] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.

[52] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS: A survey of transformer-based pretrained models in natural language processing," 2021, *arXiv:2108.05542*.

[53] E. Yang et al., "Transformer versus traditional natural language processing: How much data is enough for automated radiology report classification?," *Brit. J. Radiol.*, vol. 96, 2023, Art. no. 20220769.

[54] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[55] E. A. V. Dis, J. Bollen, W. Zuidema, R. V. Rooij, and C. L. Bockting, "ChatGPT: Five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[56] A. Tlili et al., "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learn. Environ.*, vol. 10, no. 1, 2023, Art. no. 15.

[57] N. Savage, "Drug discovery companies are customizing ChatGPT: Here's how," *Nature Biotechnol.*, vol. 41, pp. 585–586, 2023.

[58] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[59] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, 2022, Art. no. 200.

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

[61] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.

[62] K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*. New York, NY, USA: Springer 2020, pp. 603–649.

[63] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[64] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, pp. 681–694, 2020.

[65] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?," *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, 2023.

[66] D. Y. H. Wu, D. Lin, V. Chen, and H.-H. Chen, "Associated learning: An alternative to end-to-end backpropagation that works on CNN, RNN, and transformer," in *Proc. Int. Conf. Learn. Representations*, 2021.

[67] H. Mohapatra and S. R. Mishra, "Exploring the sector-specific influence and response of AI tools: A critical review," 2023, *arXiv:2307.05909*.

[68] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.

[69] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang, "Semi-supervised vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 605–620.

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

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

[72] R. Balestriero et al., "A cookbook of self-supervised learning," 2023, *arXiv:2304.12210*.

[73] T. B. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[74] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[75] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009, pp. 9–41.

[76] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[77] Y. Wang, C. Albrecht, N. A. A. Braham, L. Mou, and X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.

[78] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[79] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[80] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.

[81] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2018.

[82] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[83] 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.

[84] 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. 9729–9738.

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

[86] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.

[87] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

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

[89] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[90] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.

[91] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[92] B. Peng et al., "Urban flood mapping with bitemporal multispectral imagery via a self-supervised learning framework," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2001–2016, Dec. 2020.

[93] Q. Jin, Y. Ma, F. Fan, J. Huang, X. Mei, and J. Ma, "Adversarial autoencoder network for hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 4555–4569, Aug. 2023.

[94] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2021, Art. no. 5521213.

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

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

[97] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625513.

[98] J. González-Santiago, F. Schenkel, and W. Middelmann, "Self-supervised image colorization for semantic segmentation of urban land cover," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 3468–3471.

[99] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," 2023, *arXiv:2303.15647*.

[100] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3045–3059.

[101] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics / 11th Int. Joint Conf. Natural Lang. Process. (Volume 1: Long Papers)*, 2021, pp. 4582–4597.

[102] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 709–727.

[103] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19113–19122.

[104] C. Oh et al., "BlackVIP: Black-box visual prompting for robust transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24224–24235.

[105] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[106] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 506–516.

[107] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2021.

[108] Q. Zhang et al., "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=lq62uWRJjiY

[109] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient fine-tuning of quantized LLMs," 2023, *arXiv:2305.14314*.

[110] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[111] R. Nakano et al., "WebGPT: Browser-assisted question-answering with human feedback," 2021, *arXiv:2112.09332*.

[112] M. Chen et al., "Evaluating large language models trained on code," 2021, *arXiv:2107.03374*.

[113] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[114] A. Zeng et al., "GLM-130B: An open bilingual pre-trained model," in *Proc. 11th Int. Conf. Learn. Representations*, 2022.

[115] T. L. Scao et al., "BLOOM: A 176B-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[116] H. W. Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[117] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-1144.html

[118] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 27730–27744.

[119] 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.

[120] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3202–3211.

[121] R. Wang et al., "BEVT: BERT pretraining of video transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14733–14743.

[122] L. Wang et al., "VideoMAE V2: Scaling video masked autoencoders with dual masking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14549–14560.

[123] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," 2023, *arXiv:2304.00685*.

[124] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.

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

[126] Y. Wang et al., "InternVideo: General video foundation models via generative and discriminative learning," 2022, *arXiv:2212.03191*.

[127] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.

[128] OpenAI, "Gpt-4 technical report," 2023.

[129] X. Chen et al., "PaLI: A jointly-scaled multilingual language-image model," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=mWVoBz4W0u

[130] H. Xu et al., "mplug-2: A modularized multi-modal foundation model across text, image and video," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023.

[131] Y. Zhang et al., "Meta-transformer: A unified framework for multimodal learning," 2023, *arXiv:2307.10802*.

[132] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023.

[133] L. Jiao, J. Gao, X. Liu, F. Liu, S. Yang, and B. Hou, "Multi-scale representation learning for image classification: A survey," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 23–43, Feb. 2023.

[134] Y. Yang et al., "An explainable spatial-frequency multiscale transformer for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5907515.

[135] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, Jan. 2021.

[136] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "GCSANet: A global context spatial attention deep learning network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1150–1162, Jan. 2022.

[137] L. Li et al., "A multiscale self-adaptive attention network for remote sensing scene classification," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2209.

[138] F. Liu, X. Qian, L. Jiao, X. Zhang, L. Li, and Y. Cui, "Contrastive learning-based dual dynamic GCN for SAR image scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, May 20, 2022, early access, doi: 10.1109/TNNLS.2022.3174873.
[139] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 143–154, 2022.
[140] H. Huang, Z. Mou, Y. Li, Q. Li, J. Chen, and H. Li, "Spatial-temporal invariant contrastive learning for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 6509805.
[141] Z. Li et al., "Contrastive learning based on multiscale hard features for remote-sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5614413.
[142] J. Xu, Y. Li, Q. Shi, and L. He, "Occluded scene classification via cascade supervised contrastive learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4565–4578, May 2023.
[143] W. Miao, J. Geng, and W. Jiang, "Semi-supervised remote-sensing image scene classification using representation consistency Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616614.
[144] C. Peng, Y. Li, L. Jiao, and R. Shang, "Efficient convolutional neural architecture search for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6092–6105, Jul. 2021.
[145] X. Qian et al., "A hybrid network with structural constraints for SAR image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5202717.
[146] K. Xu, P. Deng, and H. Huang, "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618715.
[147] M. Zhao, Q. Meng, L. Zhang, X. Hu, and L. Bruzzone, "Local and long-range collaborative learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5606215, doi: 10.1109/TGRS.2023.3265346.
[148] X. Lu et al., "Weak-to-strong consistency learning for semisupervised image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5510715.
[149] 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, Jun. 2021.
[150] M. Yang et al., "Coarse-to-fine contrastive self-supervised feature learning for land-cover classification in SAR images with limited labeled data," *IEEE Trans. Image Process.*, vol. 31, pp. 6502–6516, Oct. 2022.
[151] V. Marsocci and S. Scardapane, "Continual Barlow twins: Continual self-supervised learning for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5049–5060, May 2023.
[152] Z. Xue, X. Yu, A. Yu, B. Liu, P. Zhang, and S. Wu, "Self-supervised feature learning for multimodal remote sensing image land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5533815.
[153] Z. Xue, B. Liu, A. Yu, X. Yu, P. Zhang, and X. Tan, "Self-supervised feature representation and few-shot land cover classification of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5541618.
[154] X. Lu et al., "Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5543516.
[155] W. Li, H. Gao, Y. Su, and B. M. Momanyi, "Unsupervised domain adaptation for remote sensing semantic segmentation with transformer," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4942.
[156] R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, and R. Stolkin, "Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images," *Remote Sens.*, vol. 12, no. 5, 2020, Art. no. 872.
[157] P. He et al., "MANet: Multi-scale aware-relation network for semantic segmentation in aerial scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5624615.
[158] M. Luo and S. Ji, "Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 105–128, 2022.
[159] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5535317.
[160] W. Chen, S. Ouyang, J. Yang, X. Li, G. Zhou, and L. Wang, "JAGAN: A framework for complex land cover classification using gaofen-5 AHSI images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1591–1603, Jan. 2022.
[161] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5400314.
[162] G. Lenczner, A. Chan-Hon-Tong, B. L. Saux, N. Luminari, and G. L. Besnerais, "DIAL: Deep interactive and active learning for semantic segmentation in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3376–3389, Apr. 2022.
[163] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5626915.
[164] M. Ghanbari, L. Xu, and D. A. Clausi, "Local and global spatial information for land cover semisupervised classification of complex polarimetric SAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3892–3904, Apr. 2023.
[165] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, "Self-supervised representation learning for remote sensing image change detection based on temporal prediction," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1868.
[166] P. Chen, L. Guo, X. Zhang, K. Qin, W. Ma, and L. Jiao, "Attention-guided siamese fusion network for change detection of remote sensing images," *Remote Sens.*, vol. 13, no. 22, 2021, Art. no. 4597.
[167] W. Zhang, L. Jiao, F. Liu, S. Yang, and J. Liu, "Adaptive contourlet fusion clustering for SAR image change detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2295–2308, Mar. 2022.
[168] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 4409818.
[169] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4412712.
[170] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5224713.
[171] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5620014.
[172] W. Zhang, L. Jiao, F. Liu, S. Yang, W. Song, and J. Liu, "Sparse feature clustering network for unsupervised SAR image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5226713.
[173] H. Dong et al., "Deep shearlet network for change detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5241115.
[174] H. Dong, W. Ma, L. Jiao, F. Liu, and L. Li, "A multiscale self-attention deep clustering for change detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5207016.
[175] H. Zheng et al., "HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108717.
[176] J. Wang et al., "SSCFNet: A spatial-spectral cross fusion network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4000–4012, Apr. 2023.
[177] M. Zhang, Z. Liu, J. Feng, L. Liu, and L. Jiao, "Remote sensing image change detection based on deep multi-scale multi-attention Siamese transformer network," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 842.
[178] Y. Li, B. He, F. Melgani, and T. Long, "Point-based weakly supervised learning for object detection in high spatial resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5361–5371, 2021.
[179] T. Zhang et al., "Foreground refinement network for rotated object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5610013.
[180] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5604114.
[181] J. Bai et al., "Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405316.
[182] Y. Liu, S. Zhang, Z. Wang, B. Zhao, and L. Zou, "Global perception network for salient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5617212.
[183] G. Cheng, X. Xie, W. Chen, X. Feng, X. Yao, and J. Han, "Self-guided proposal generation for weakly supervised object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625311.

[184] Y. Ye et al., "An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 516.

[185] T. Zhang, X. Zhang, P. Zhu, X. Jia, X. Tang, and L. Jiao, "Generalized few-shot object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 353–364, 2023.

[186] G. Wang, X. Zhang, Z. Peng, X. Jia, X. Tang, and L. Jiao, "MOL: Towards accurate weakly supervised remote sensing object detection via multi-view nOisy learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 457–470, 2023.

[187] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5602011.

[188] T. Zhang et al., "Object-centric masked image modeling-based self-supervised pretraining for remote sensing object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5013–5025, May 2023.

[189] Y. Han, W. Meng, and W. Tang, "Capsule-inferenced object detection for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5260–5270, Apr. 2023.

[190] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.

[191] J. Zhang, X. Jia, J. Hu, and J. Chanussot, "Online structured sparsity-based moving-object detection from satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6420–6433, Sep. 2020.

[192] J. Feng et al., "Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 116–130, 2021.

[193] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5185–5198, Sep. 2022.

[194] Z. Pi et al., "Very low-resolution moving vehicle detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624517.

[195] C. Xiao et al., "DSFNet: Dynamic and static fusion network for moving object detection in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3510405.

[196] J. Feng, Y. Liang, X. Zhang, J. Zhang, and L. Jiao, "SDANet: Semantic-embedded density adaptive network for moving vehicle detection in satellite videos," *IEEE Trans. Image Process.*, vol. 32, pp. 1788–1801, Mar. 2023.

[197] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution Siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.

[198] W. Zhang, L. Jiao, F. Liu, S. Yang, and J. Liu, "DFAT: Dynamic feature-adaptive tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 43–58, Jan. 2023.

[199] Y. Cui, B. Hou, Q. Wu, B. Ren, S. Wang, and L. Jiao, "Remote sensing object tracking with deep reinforcement learning under occlusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5605213.

[200] W. Zhang, L. Jiao, F. Liu, L. Li, X. Liu, and J. Liu, "MBLT: Learning motion and background for vehicle tracking in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703315.

[201] W. Song et al., "A joint siamese attention-aware network for vehicle object tracking in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625617.

[202] Y. Li, C. Bian, and H. Chen, "Object tracking in satellite videos: Correlation particle filter tracking method with motion estimation by Kalman filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5630112.

[203] S. Chen et al., "Vehicle tracking on satellite video based on historical model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7784–7796, Aug. 2022.

[204] Y. Li and C. Bian, "Object tracking in satellite videos: A spatial-temporal regularized correlation filter tracking method with interacting multiple model," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 6511105.

[205] Y. Nie, C. Bian, and L. Li, "Object tracking in satellite videos based on Siamese network with multidimensional information-aware and temporal motion compensation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 6517005.

[206] X. Li et al., "A collaborative learning tracking network for remote sensing videos," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1954–1967, Mar. 2023.

[207] R. Zhang, L. Jiao, L. Li, X. Liu, F. Liu, and S. Yang, "A quantum evolutionary learning tracker for video," *IEEE Trans. Evol. Comput.*, to be published, doi: 10.1109/TEVC.2023.3264641.

[208] J. Yang, Z. Pan, Z. Wang, B. Lei, and Y. Hu, "SiamMDM: An adaptive fusion network with dynamic template for real-time satellite video single object tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608619.

[209] W. Ao, Y. Fu, X. Hou, and F. Xu, "Needles in a haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Trans. Image Process.*, vol. 29, pp. 1944–1957, 2020.

[210] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5619513.

[211] J. Zhang, X. Zhang, Z. Huang, X. Cheng, J. Feng, and L. Jiao, "Bidirectional multiple object tracking based on trajectory criteria in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5603714.

[212] L. Kong, Z. Yan, Y. Zhang, W. Diao, Z. Zhu, and L. Wang, "CF-Tracker: Multi-object tracking with cross-frame connections in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5611214.

[213] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, "Visual grounding in remote sensing images," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 404–412.

[214] Y. Zhan, Z. Xiong, and Y. Yuan, "RSVG: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5604513.

[215] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "RRSIS: Referring remote sensing image segmentation," 2023, *arXiv:2306.08625*.

[216] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.

[217] X. Ye et al., "A joint-training two-stage method for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709616.

[218] Y. Li et al., "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608816.

[219] S. Wang et al., "Multi-label semantic feature fusion for remote sensing image captioning," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 1–18, 2022.

[220] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 939.

[221] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "TypeFormer: Multi-scale transformer with type controller for remote sensing image caption," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 6514005.

[222] J. Wang, Z. Chen, A. Ma, and Y. Zhong, "CapFormer: Pure transformer for remote sensing image caption," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 7996–7999.

[223] Q. Yang, Z. Ni, and P. Ren, "Meta captioning: A meta learning based remote sensing image captioning framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 186, pp. 190–200, 2022.

[224] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multiinteraction network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2154–2165, Feb. 2022.

[225] U. Zia, M. M. Riaz, and A. Ghafoor, "Transforming remote sensing images to textual descriptions," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102741.

[226] X. Zhang et al., "Multi-source interactive stair attention for remote sensing image captioning," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 579.

[227] S. Chang and P. Ghamisi, "Changes to captions: An attentive network for remote sensing change captioning," 2023, *arXiv:2304.01091*.

[228] Z. Zhang et al., "A spatial hierarchical reasoning network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 4400815.

[229] C. Chappuis, V. Zermatten, S. Lobry, B. L. Saux, and D. Tuia, "Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1372–1381.

[230] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, "From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5623111.

[231] Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, M. A. Al Zuair, and F. Melgani, "Bi-modal transformer-based approach for visual question answering in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 4708011.

[232] M. M. Al Rahhal et al., "Open-ended remote sensing visual question answering with transformers," *Int. J. Remote Sens.*, vol. 43, no. 18, pp. 6809–6823, 2022.

[233] C. Chappuis, V. Mendez, E. Walt, S. Lobry, B. Le Saux, and D. Tuia, "Language transformers for remote sensing visual question answering," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 4855–4858.

[234] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5630613.

[235] L. Bashmal, Y. Bazi, F. Melgani, R. Ricci, M. M. Al Rahhal, and M. Zuair, "Visual question generation from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3279–3293, Mar. 2023.

[236] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.

[237] L. Jiao et al., "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2022.

[238] Y. Li et al., "Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 181–212, Dec. 2022.

[239] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

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

[241] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[242] 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.

[243] 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.

[244] R. Hänsch et al., "The 2022 IEEE GRSS data fusion contest: Semisupervised learning [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 334–337, Mar. 2022.

[245] J. Castillo-Navarro, B. L. Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in Earth observation: The miniFrance suite, dataset analysis and multi-task network study," *Mach. Learn.*, vol. 111, pp. 3125–3160, 2022.

[246] ISPRS, "2D semantic labeling challenge," 2016. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[247] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618515.

[248] 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.

[249] ISPRS, "2D semantic labeling challenge," 2016. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[250] 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.

[251] 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. 11976–11986.

[252] M. Guo, C. Lu, Z. Liu, M. Cheng, and S. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, pp. 733–752, 2023.

[253] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4804–4814.

[254] 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. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[255] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.

[256] Y. Rao et al., "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18082–18091.

[257] 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. 3520–3529.

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

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

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

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

[262] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[263] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

[264] B. Yin, F. Corradi, and S. M. Bohté, "Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks," *Nature Mach. Intell.*, vol. 3, no. 10, pp. 905–913, 2021.

[265] S. Herculano-Houzel, "The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost," *Proc. Nat. Acad. Sci.*, vol. 109, no. supplement_1, pp. 10661–10668, 2012.

[266] S. Zheng, L. Qian, P. Li, C. He, X. Qin, and X. Li, "An introductory review of spiking neural network and artificial neural network: From biological intelligence to artificial intelligence," 2022, *arXiv:2204.07519*.

[267] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Front. Neurosci.*, vol. 12, 2018, Art. no. 774.

[268] W. Penfield and E. Boldrey, "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation," *Brain*, vol. 60, no. 4, pp. 389–443, 1937.

[269] Z. Yao et al., "A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex," *Nature*, vol. 598, no. 7879, pp. 103–110, 2021.

[270] J. Berg et al., "Human neocortical expansion involves glutamatergic neuron diversification," *Nature*, vol. 598, no. 7879, pp. 151–158, 2021.

[271] J. C. Pang et al., "Geometric constraints on human brain function," *Nature*, vol. 618, pp. 566–574, 2023.

[272] M. Yao et al., "Attention spiking neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9393–9410, Aug. 2023.

[273] E. Genç et al., "Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 1905.

[274] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[275] P. R. Roelfsema, "Attention voluntary control of brain cells," *Science*, vol. 332, no. 6037, pp. 1512–1513, 2011.

[276] D. E. L. Lockhofen and C. Mulert, "Neurochemistry of visual attention," *Front. Neurosci.*, vol. 15, 2021, Art. no. 643597.

[277] A. Thiele and M. A. Bellgrove, "Neuromodulation of attention," *Neuron*, vol. 97, no. 4, pp. 769–785, 2018.

[278] A. Finkelstein, D. Derdikman, A. Rubin, J. N. Foerster, L. Las, and N. Ulanovsky, "Three-dimensional head-direction coding in the bat brain," *Nature*, vol. 517, no. 7533, pp. 159–164, 2015.

[279] L. Kunz et al., "A neural code for egocentric spatial maps in the human medial temporal lobe," *Neuron*, vol. 109, no. 17, pp. 2781–2796, 2021.

[280] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.
[281] S. Ikeda, "Functional recovery on stroke, stroke model and rehabilitation(progress in regenerative medicine)," *Japanese J. Psychosomatic Med.*, vol. 53, no. 8, pp. 742–747, 2013.
[282] Y. Wu et al., "Brain-inspired global-local learning incorporated with neuromorphic computing," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 65.
[283] W. Gerstner and W. M. Kistler, "Mathematical formulations of Hebbian learning," *Biol. Cybern.*, vol. 87, no. 5, pp. 404–415, 2002.
[284] P. J. Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, vol. 1. Hoboken, NJ, USA: Wiley, 1994.
[285] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[286] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cogn. Sci.*, vol. 9, no. 1, pp. 75–112, 1985.
[287] H. Zhou, "Activation learning by local competitions," 2022, *arXiv:2209.13400*.
[288] S. A. Josselyn and S. Tonegawa, "Memory engrams: Recalling the past and imagining the future," *Science*, vol. 367, no. 6473, 2020, Art. no. eaaw4325.
[289] R. D. Fields, "The enigma of working memory: Changing views," *Neuroscientist*, vol. 28, no. 5, pp. 420–424, 2022.
[290] P. S. Goldman-Rakic, "Cellular basis of working memory," *Neuron*, vol. 14, no. 3, pp. 477–485, 1995.
[291] G. Mongillo, O. Barak, and M. Tsodyks, "Synaptic theory of working memory," *Science*, vol. 319, no. 5869, pp. 1543–1546, 2008.
[292] S. M. Courtney, L. G. Ungerleider, K. Keil, and J. V. Haxby, "Transient and sustained activity in a distributed neural system for human working memory," *Nature*, vol. 386, no. 6625, pp. 608–611, 1997.
[293] J. J. Foster, D. W. Sutterer, J. T. Serences, E. K. Vogel, and E. Awh, "The topography of alpha-band activity tracks the content of spatial working memory," *J. Neuriophysiol.*, vol. 115, no. 1, pp. 168–177, 2016.
[294] E.-L. Yap et al., "Bidirectional perisomatic inhibitory plasticity of a FOS neuronal network," *Nature*, vol. 590, no. 7844, pp. 115–121, 2021.
[295] R. L. Davis and Y. Zhong, "The biology of forgetting a perspective," *Neuron*, vol. 95, no. 3, pp. 490–503, 2017.
[296] R. Rideauxa, K. R. Storrsc, G. Maielloc, and A. E. Welchmanb, "How multisensory neurons solve causal inference," *Proc. Nat. Acad. Sci.*, vol. 118, no. 32, 2021, Art. no. e2106235118.
[297] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: A survey," *Big Data*, vol. 9, no. 1, pp. 3–21, 2021.
[298] C. Helwe, C. Clavel, and F. Suchanek, "Reasoning with transformer-based models: Deep learning, but shallow reasoning," in *Proc. Int. Conf. Automated Knowl. Base Construct.*, 2021.
[299] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.
[300] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997.
[301] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.
[302] Y. Tang, K. Han, J. Guo, C. Xu, Y. Li, and Y. Wang, "An image patch is a wave: Phase-aware vision MLP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10925–10934.
[303] C. Li, L. Du, Y. Li, and J. Song, "A novel SAR target recognition method combining electromagnetic scattering information and GCN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 4508705.
[304] I. Sutskever, "An observation on generalization," 2023. [Online]. Available: https://simons.berkeley.edu/talks/ilya-sutskever-openai-2023-08-14
[305] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.
[306] Z. Jin et al., "Can large language models infer causation from correlation?," 2023, *arXiv:2306.05836*.
[307] Y. Qin et al., "Tool learning with foundation models," 2023, *arXiv:2304.08354*.
[308] M. Glymour, J. Pearl, and N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.
[309] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund, "A review of road extraction from remote sensing images," *J. Traffic Transp. Eng. (English Ed.)*, vol. 3, no. 3, pp. 271–282, 2016.
[310] Q. Wu, C. Yang, W. Zhao, Y. He, D. Wipf, and J. Yan, "DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=j6zUzrapY3L
[311] C. Deng et al., "Learning a foundation language model for geoscience knowledge understanding and utilization," 2023, *arXiv:2306.05064*.
[312] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," 2023, *arXiv:2306.08302*.
[313] X. Wang et al., "Large-scale multi-modal pre-trained models: A comprehensive survey," *Mach. Intell. Res.*, vol. 20, pp. 447–482, 2023.
[314] N. Spyrison, D. Cook, and K. Marriott, "A study on a user-controlled radial tour for variable importance in high-dimensional data," 2022, *arXiv:2301.00077*.
[315] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, and X. Qiu, "Full parameter fine-tuning for large language models with limited resources," 2023, *arXiv:2306.09782*.
[316] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.
[317] B. Peng, Bo Peng, J. Zhou, J. Xie, and L. Liu, "Scattering model guided adversarial examples for SAR target recognition: Attack and defense," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5236217.

Licheng Jiao (Fellow, IEEE) received the B.S. degree in electrical engineering and computer science from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. degree in electric engineering and the Ph.D. degree in signals, circuits, and systems from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Distinguished Professor with the School of Electronic Engineering, Xidian University, Xi'an, where he is currently the Director of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. His research interests include machine learning, deep learning, natural computation, remote sensing, image processing, and intelligent information processing.

Dr. Jiao is the Chairperson of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Foreign Member of the Academia European and Russian Academy of Natural Sciences, a Fellow of IEEE, The Institution of Engineering and Technology (IET), Chinese Association for Artificial Intelligence (CAAI), China Computer Federation (CCF), and Chinese Association of Automation (CAA), a Councilor of the Chinese Institute of Electronics(CIE), a Committee Member of the Chinese Committee of Neural Networks, and an expert of the Academic Degrees Committee of the State Council.

Zhongjian Huang (Student Member, IEEE) received the B.S. degree in intelligent science and technology in 2018 from Xidian University, Xi'an, China, where he is currently working toward the Ph.D. degree in computer science and technology.

He is currently a member of the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, International Research Center for Intelligent Perception and Computation, and Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, China. His current research interests include video tracking and satellite videos analysis.

Xiaoqiang Lu (Graduate Student Member, IEEE) received the B.S. degree in information countermeasure technique in 2020 from Xidian University, Xi'an, China, where he is currently working toward the Ph.D. degree in computer science and technology with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, School of Artificial Intelligence.

His research interests include machine learning, deep learning, object detection, and semantic segmentation.

Xu Liu (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from the North University of China, Taiyuan, China, in 2013, and the Ph.D. degree in electronic circuit and system from Xidian University, Xi'an, China, in 2019.

He is currently an Associate Professor of Huashan elite and Postdoctoral Researcher of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. From 2015 to 2019, he was the Chair of the IEEE Xidian University Student Branch. His current research interests include machine learning and image processing.

Yuting Yang (Graduate Student Member, IEEE) received the B.S. degree in electronic information science and technology from Northwest University, Xi'an, China, in 2018. She is currently working toward the Ph.D. degree in computer science and technology with Xidian University, Xi'an, China.

She is currently a member of the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, International Research Center for Intelligent Perception and Computation, and Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, China. Her research interests include computer vision, the interpretability of deep learning, and multiscale geometric analysis.

Jiaxuan Zhao (Graduate Student Member, IEEE) received the B.S. degree in materials science and engineering in 2019 from Xidian University, Xi'an, China, where she is currently working toward the Ph.D. degree in computer science and technology with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence.

Her research interests include multimodal fusion, evolutionary computing, and image understanding.

Jinyue Zhang received the B.S. degree in intelligent science and technology in 2018 from Xidian University, Xian, China, where she is currently working toward the Ph.D. degree in computer science and technology.

Her current research interests include video tracking and satellite video analysis.

Biao Hou (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2003.

Since 2003, he has been with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University, where he is currently a Professor. His research interests include compressive sensing and synthetic aperture radar image interpretation.

Shuyuan Yang (Senior Member, IEEE) received the B.A. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively.

She has been a Professor with the School of Artificial Intelligence, Xidian University. Her research interests include machine learning and multiscale geometric analysis.

Fang Liu (Senior Member, IEEE) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree from Xidian University, Xi'an, China, in 1995, both in computer science and technology.

She is currently a Professor with the School of Computer Science, Xidian University. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

Wenping Ma (Senior Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2003 and 2008, respectively.

She is currently an Associate Professor with the School of Artificial Intelligence, Xidian University. Her research interests include natural computing and intelligent image processing.

Dr. Ma is a member of CIE.

Lingling Li (Senior Member, IEEE) received the B.S. degree in electronic and information engineering and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

She is currently an Associate Professor with the School of Artificial Intelligence, Xidian University. From 2013 to 2014, she was an Exchange Ph.D. Student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Leioa, Spain. Her research interests include quantum evolutionary optimization, and deep learning.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer application technology from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2006.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. From January 2015 to March 2016, she was a Visiting Scientist with Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.

Puhua Chen (Senior Member, IEEE) received the B.S. degree in environmental engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D. degree in circuit and system from Xidian University, Xi'an, China, in 2016.

She is currently a Lecturer with the School of Artificial Intelligence, Xidian University. Her research interests include machine learning, pattern recognition, and remote sensing image interpretation.

Zhixi Feng (Member, IEEE) received the B.A. degree in automation from the Lanzhou University of Technology, Lanzhou, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2018.

He is currently an Associate Professor of artificial intelligence with Xidian University. His research interests include machine learning and remote sensing information processing.

Xu Tang (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and system from Xidian University, Xi'an, China, in 2007, 2010, and 2017, respectively.

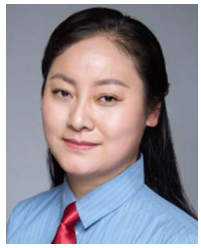
From 2015 to 2016, he was a joint Ph.D. student along with Prof. W. J. Emery with the University of Colorado at Boulder, Boulder, CO, USA. He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. His research interests include remote sensing image content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, object detection, etc.

Yuwei Guo (Senior Member, IEEE) was born in Shaanxi, China, on March 1988. She is currently working toward the M.S. and Ph.D. degrees in circuit and system with Xidian University, Xi'an, China.

She is also an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University. Her research interests include rough set theory, data mining, and image processing.

Dou Quan (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2015 and 2021, respectively.

From 2019 to 2020, she was a joint Ph.D. student along with Prof. Jocelyn Chanussot with the Research Center of Inria Grenoble-Rhone-Alpes, France. She is currently a Lecturer with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University. Her research interests include machine learning, deep learning and metric learning, image matching, image registration, and image classification.

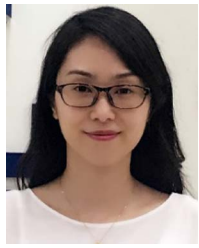


Shuang Wang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in circuits and systems from Xidian University, Xi'an, China, in 2000, 2003, and 2007, respectively.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University. Her research interests include sparse representation, image processing, synthetic aperture radar (SAR) automatic target recognition, remote sensing image captioning, and polarimetric SAR data analysis.

Weibin Li received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1998 and 2000, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2004.

He is currently a Professor with Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an. His research interests include spatio-temporal intelligence, GNSS navigation system, remote sensing image processing, industrial Internet, and industrial intelligence.



Jing Bai (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Zhengzhou University, Zhengzhou, China, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2009.

She is currently a Professor with Xidian University. Her research interests include image processing, machine learning, and intelligent information processing.

Yangyang Li (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science and technology and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2001, 2004, and 2007, respectively.

She is currently a Professor with the School of Artificial Intelligence, Xidian University. Her research interests include quantum-inspired evolutionary computation, artificial immune systems, and deep learning.

Ronghua Shang (Senior Member, IEEE) received the B.S. degree in information and computation science and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2003 and 2008, respectively.

She is currently a Professor with Xidian University. Her research interests include evolutionary computation, image processing, and data mining.

Jie Feng (Senior Member, IEEE) received the B.S. degree in electronic information engineering from Chang'an University, Xi'an, China, in 2008, and the Ph.D. degree in electronic circuit and system from Xidian University, Xi'an, in 2014.

She is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University. Her research interests include remote sensing image processing, deep learning, and machine learning.