HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model

Di Wang*, Meiqi Hu*, Yao Jin*, Yuchun Miao*, Jiaqi Yang*, Yichu Xu*, Xiaolei Qin*, Jiaqi Ma*, Lingyu Sun*, Chenxing Li*, Chuan Fu, Hongruixuan Chen, Chengxi Han†, Naoto Yokoya, *Member, IEEE,* Jing Zhang†, *Senior Member, IEEE,* Minqiang Xu, Lin Liu, Lefei Zhang, *Senior Member, IEEE,* Chen Wu†, *Member, IEEE,* Bo Du†, *Senior Member, IEEE,* Dacheng Tao, *Fellow, IEEE* and Liangpei Zhang†, *Fellow, IEEE*

Abstract—Accurate hyperspectral image (HSI) interpretation is critical for providing valuable insights into various earth observation-related applications such as urban planning, precision agriculture, and environmental monitoring. However, existing HSI processing methods are predominantly task-specific and scene-dependent, which severely limits their ability to transfer knowledge across tasks and scenes, thereby reducing the practicality in real-world applications. To address these challenges, we present HyperSIGMA, a vision transformer-based foundation model that unifies HSI interpretation across tasks and scenes, scalable to over one billion parameters. To overcome the spectral and spatial redundancy inherent in HSIs, we introduce a novel sparse sampling attention (SSA) mechanism, which effectively promotes the learning of diverse contextual features and serves as the basic block of HyperSIGMA. HyperSIGMA integrates spatial and spectral features using a specially designed spectral enhancement module. In addition, we construct a large-scale hyperspectral dataset, HyperGlobal-450K, for pre-training, which contains about 450K hyperspectral images, significantly surpassing existing datasets in scale. Extensive experiments on various high-level and low-level HSI tasks demonstrate HyperSIGMA's versatility and superior representational capability compared to current state-of-the-art methods. Moreover, HyperSIGMA shows significant advantages in scalability, robustness, cross-modal transferring capability, real-world applicability, and computational efficiency. The code and models will be released at HyperSIGMA.

Index Terms—Remote sensing, Hyperspectral image, Foundation model, Attention, Vision transformer, Large-scale dataset

1 Introduction

The rapid advancements in aeronautical engineering, sensor technology, and computer science have made it possible to acquire massive hyperspectral remote sensing images (hereafter referred to as HSIs) with fine spectral resolution [1]–[3]. HSIs cover a spectral range from visible near-infrared to short-wave and mid-infrared, capturing target features through continuous and fine spectral bands. This results in nearly continuous spectral curves that provide detailed surface information, enabling the distinction of subtle spectral differences between substances for precise land cover interpretation [4]. Hyperspectral imagery significantly enhances our capability for comprehensive, accurate, and timely earth observation and monitoring [5]–[7]. It offers crucial scientific insights and decision-making support for fields such as urban planning [8], precision agriculture [9], and environmental monitoring [10].

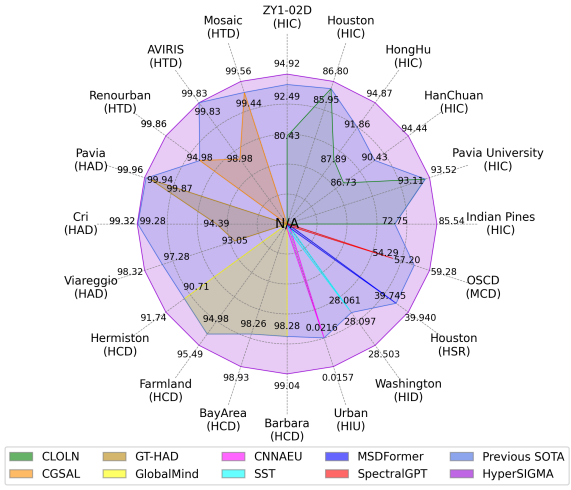


Fig. 1. HyperSIGMA offers a universal solution for HSI processing, demonstrating superior performance across 20 datasets, including both high-level and low-level hyperspectral tasks, as well as multispectral scenes. It outperforms advanced models like SpectralGPT, even those specifically designed for these tasks. HIC: Hyperspectral Image Classification. HTD: Hyperspectral Target Detection. HAD: Hyperspectral Anomaly Detection. HCD: Hyperspectral Change Detection. HIU: Hyperspectral Image Unmixing. HID: Hyperspectral Image Denoising. HSR: Hyperspectral Super-Resolution. MCD: Multispectral Change Detection.

The challenges of processing HSIs primarily stem from their inherent characteristics: *high dimensionality, data redundancy, and spatial variability* [11]. High dimensionality

D. Wang, M. Hu, Y. Jin, Y. Miao, J. Yang, Y. Xu, X. Qin, J. Ma, L. Sun, C. Li, W. Jiang, C. Han, J. Zhang, L. Zhang, C. Wu, B. Du and L. Zhang are with the Wuhan University, China; C. Fu is with the Chongqing University, China; H. Chen and N. Yokoya are with The University of Tokyo, Japan; M. Xu and L.Liu are with the National Engineering Research Center of Speech and Language Information Processing, China; D. Tao is with the Nanyang Technological University, Singapore. ({d_wang, meiqi.hu, yao.jin, miaoyuchun, jqyang, xuyichu, qinxlei, jiaqima, lingyu.sun, chenxing.li, chengxihan, zhanglefei, chen.wu, dubo, zlp62}@whu.edu.cn; fuchuan@cqu.edu.cn; {qschrx, jingzhang.cv, dacheng.tao}@gmail.com; yokoya@k.u-tokyo.ac.jp; {mqxu7, linliu}@iflytek.com).
**: Equal contribution; †: Corresponding author.*

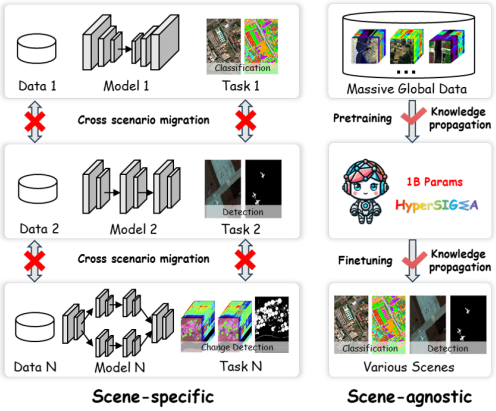


Fig. 2. Previous HSI models are trained separately on different scenes, limiting cross-scene knowledge transfer. In contrast, our model acquires universal, scene-agnostic knowledge through pre-training with a large dataset of global HSIs, enabling effective transfer to various scenes through fine-tuning.

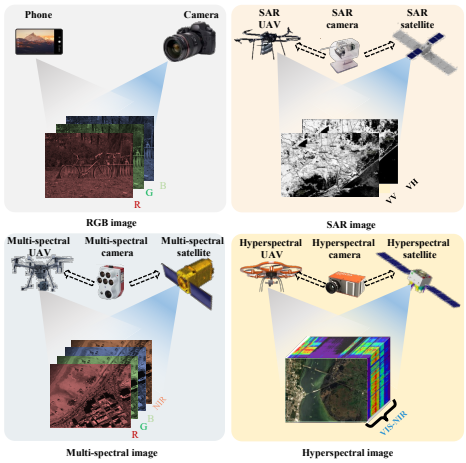


Fig. 3. Comparison of RGB, Synthetic-aperture radar (SAR), multi-spectral, and hyperspectral images.

often leads to the *Hughes phenomenon* [12], where machine learning models overfit as the number of channels increases, especially with limited training samples. Data redundancy, both spectral and spatial, results in unnecessary computations due to uninformative bands and pixels, often linked to the sensors' spectral and spatial resolutions. Spatial variability, influenced by imaging conditions like atmosphere, lighting, and topography, causes mismatches between object categories and spectral curves. To address these challenges, traditional HSI interpretation strategies typically involve two steps: *dimensionality reduction* [13] and *feature extraction* [14], with feature extraction being a critical research focus due to its impact on subsequent tasks' performance. The HSI community continually seeks effective feature extractors, leading to numerous proposed methods [15]–[17]. Among these, deep learning techniques have become predominant for their ability to automatically extract strong feature representations. To tackle spatial variability and enhance contextual feature representation, significant progress has been made with HSI interpretation methods based on convolutional neural networks (CNNs) [18], recurrent neural networks (RNNs) [19], transformers [20], and state space models [21][1]. Despite these advancements, most HSI methods remain scene-dependent, meaning models are trained and tested on the same specialized scenarios with limited cross-scene knowledge transfer, even for similar tasks. For instance, models for classifying Indian Pines and Pavia University scenes are trained separately, restricting their applicability. Therefore, developing universal HSI processing methods is crucial for advancing the hyperspectral community. Fig. 2 illustrates the differences between traditional and universal HSI processing schemes.

Foundation models, developed through training on vast datasets and optimized using substantial computational resources, have recently been transforming the field of artificial intelligence [22], [23]. Most foundation models utilize the transformer architecture [24], known for its scalability and flexibility, allowing for rapid parameter scaling through the stacking of transformer blocks. In the field of computer vision (CV), numerous large-scale foundation models [25]–[27] based on vision transformers [28], [29] have demonstrated exceptional performance across various tasks. Similarly, current remote sensing (RS) foundation models [30], [31] have proven capable of effectively handling multiple tasks, including scene classification, semantic segmentation, object detection, and change detection.

Intuitively, developing hyperspectral foundation models may provide a promising solution for unified HSI interpretation across tasks and scenes. However, our literature review reveals a lack of foundation models specifically designed for HSI interpretation, despite their ability to capture terrestrial objects at a finer wavelength scale than aerial RGB, SAR, and multispectral images (see Fig. 3). The challenges to developing large-scale foundation models for HSI processing primarily lie in the unique characteristics of hyperspectral data, which complicate data collection, pre-training, and model design. Specifically, large-scale pre-training requires significant computational resources, while acquiring and processing hyperspectral data is labor- and time-intensive. Additionally, effective pre-training strategies and model architectures are expected to be tailored to the distinct characteristics of HSIs. These combined obstacles hinder the progress of large-scale hyperspectral foundation models in the RS community.

To address these challenges, we introduce **HyperSIGMA** (**HyperS**pectral **I**ntelli**G**ence co**M**prehension found**A**tion model), the first step towards hyperspectral foundation models tailored for HSI interpretation. HyperSIGMA integrates spatial and spectral features using a specially designed spectral enhancement module. To tackle the issues of spectral and spatial redundancy in HSIs, we introduce a novel sparse sampling attention (SSA) mechanism, which effectively promotes the learning of diverse contextual features and serves as the foundational block of HyperSIGMA. In addition, we have constructed a large-scale hyperspectral dataset, HyperGlobal-450K (**Hyper**spectral **Global** Image dataset), for pre-training. This dataset comprises about 450K hyperspectral images, equivalent to over 20 million trispectral images with non-overlapping channels. The appendix

1. The development history of HSI processing technology can be found in the appendix.

presents detailed comparisons between HyperGlobal-450K and existing large-scale RS datasets. Drawing inspiration from the scaling law [32] and the success of existing large-scale RS foundation models [3], [33], [34], we have scaled HyperSIGMA to over 1 billion parameters, supported by HyperGlobal-450K. Unlike existing RS foundation models focusing primarily on high-level tasks such as semantic segmentation and object detection, with minimal exploration of low-level tasks like image denoising and super-resolution, HyperSIGMA offers a unified solution to both high-level and low-level tasks (see Fig. 1). We hope this study provides valuable insights into developing foundation models for HSIs and anticipate that HyperSIGMA's strong representation capability will advance the hyperspectral RS field across diverse applications.

The main contributions of this paper are four-fold:

- We construct a global hyperspectral image dataset, HyperGlobal-450K, facilitating large-scale pre-training of hyperspectral foundation models. HyperGlobal-450K surpasses existing multispectral and hyperspectral datasets in volume by orders.
- We develop a hyperspectral intelligence comprehension foundation model, HyperSIGMA, with over 1 billion parameters. It is the first billion-level foundation model specifically designed for HSI interpretation, offering a unified solution to both high-level and low-level tasks.
- We propose a novel attention mechanism, sparse sampling attention, addressing challenges inherent to hyperspectral images by effectively extracting strong feature representations with diverse contexts.
- Extensive experiments across diverse HSI tasks provide valuable insights into HyperSIGMA's remarkable versatility and superior representational capability compared to current state-of-the-art methods. Moreover, HyperSIGMA demonstrates significant advantages in scalability, robustness, cross-modal transferring capability, real-world applicability and computational efficiency.

2 Related Work

Numerous foundation models have been proposed for natural images, with the vision transformer adopted as the mainstream network structure. Among them, scaling is an effective strategy. Typically, ViT [29] and Swin Transformer [28] have been scaled up to 1 or 10 billion parameters [27], [35] to explore the potential of large-scale vision models. Another available approach is combining the advantages of both convolution and attention. For this purpose, ViTAE [25] introduces the inductive bias of CNN into vision transformers. In addition to vision transformers, CNN-based large-scale foundation models have also been explored [36].

Compared to the large model size in the CV field, the foundation models in the RS field are usually smaller, focusing more on the pre-training strategy on RS images [37]. Due to the high cost of annotating RS datasets and the abundance of unlabeled images, pre-training via self-supervised learning is more prevalent. Among them, contrastive learning [38] leverages RS-specific characteristics such as seasonal [39], temporal [40], and geographic variations [41]. Besides, many models [3], [30], [31], [33], [42]–[44] employ masked image modeling (MIM) [45], [46]. For instance, SatMAE [42] extends MAE [46] to multispectral and multitemporal data. SpectralGPT [3] partitions multispectral images into 3-D cubes to create visual embeddings for MAE. While SMLFR [43] develops pioneering generative CNN foundation models for RS scenes by pre-training on approximately 9 million images, achieving superior performance on multiple RS downstream tasks. In addition to pre-training from scratch, some RS foundation models are constructed based on existing vision models. GFM [44] utilizes a powerful universal foundation model to guide the learning of RS features, while MTP [47] performs continuous multi-task supervised pre-training on existing RS or CV foundation models. In contrast, UPetu [48] introduces the first unified RS parameter-efficient fine-tuning framework, achieving performance comparable to full fine-tuning while optimizing just 0.73% of the parameters.

Compared to existing approaches that primarily focus on RGB and multispectral images, HyperSIGMA is the first billion-level foundation model specifically for HSI interpretation. Leveraging pre-training on the large-scale HyperGlobal-450K dataset, HyperSIGMA employs a sparse sampling mechanism to tackle spectral and spatial redundancy in HSIs, offering a unified approach to both high-level and low-level hyperspectral vision tasks. Further detailed comparison between HyperSIGMA and existing representative CV and RS foundation models can be found in the appendix, where more discussions about the related works of HSI processing techniques, MHSA-based vision transformer networks, and large-scale RS datasets are included.

3 The HyperGlobal-450K Dataset

3.1 Data Source

In view of a series of significant criteria such as global coverage and free access, we select both the Earth Observing One (EO-1) and Gaofen5 (GF-5) satellites as the data source of HyperGlobal-450K. The details of the motivation and workflow in selecting sensors and images, and the basic information of these satellites are presented in the appendix.

3.2 Data Acquisition and Processing

We collect all EO-1 images acquired during 2011-2017 and additionally compensate the GF-5 images from China. Then, we design a set of standards to achieve image selection from cloud contents, locations, and bands. After clipping, we obtain a large-scale hyperspectral dataset with global coverage (see Fig. 4), containing 447,072 HSI patches in size of 64×64, i.e., HyperGlobal-450K, including 247,072 EO-1 patches and 200,000 GF-5 patches. Detailed data acquisition and processing pipelines can be found in the appendix.

4 Methodology

The construction of HyperSIGMA involves three main steps: initializing model weights through pre-training, enhancing model structure with SSA, and fusing spatial-spectral features. Following typical HSI processing strategies [19], we

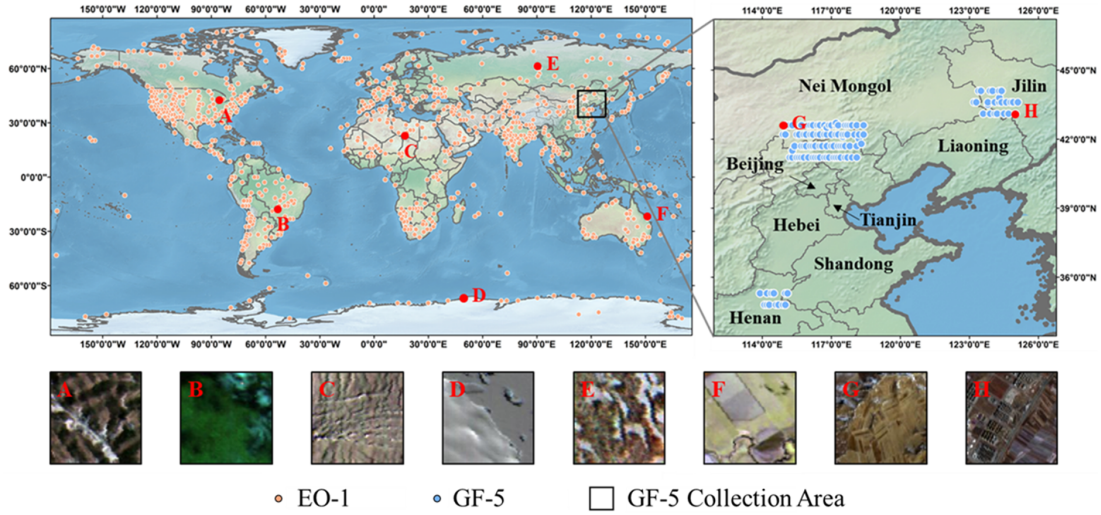


Fig. 4. The distribution of HyperGlobal-450K samples across the globe. The sampled patches of typical landscapes from different regions, including forests, grasslands, barelands, and croplands, clearly exhibit the characteristics of their respective geographical regions.

use two parallel subnetworks to extract spatial and spectral features. First, inspired by [31], [49], we obtain the model weights using MAE [46] pre-training on HyperGlobal-450K. Importantly, the spatial and spectral networks are pre-trained separately. Next, SSA is integrated into the model to enhance its structure. Lastly, we fuse the spatial and spectral information to enhance the representation of the extracted features, resulting in the final HyperSIGMA model. Technical details will be presented in the following sections.

4.1 Model Pre-training

4.1.1 Masked Image Modeling

MAE [46] is a widely used MIM-based self-supervised learning technique. In MAE, an image is divided into non-overlapping patches, some of which are masked. The network then predicts and reconstructs these masked patches using the visible ones. The loss function is determined by comparing the network's predictions to the ground truth of the masked areas. MAE is particularly effective for pre-training ViTs [29] on large-scale unlabeled datasets. In this study, we employ MAE to pre-train both spatial and spectral subnetworks on HyperGlobal-450K.

4.1.2 Pre-training of Spatial Subnetwork

Similar to many multispectral foundation models [3], [42], we use ViT [29] as the backbone for the spatial subnetwork. Following successful practices [3], [31], [42], [49], we employ MAE [46] for pre-training. The only modification from the original implementation on natural images is adjusting the input channel of the patch embedding layers to match the number of channels in the input HSIs.

4.1.3 Pre-training of Spectral SubNetwork

For the spectral subnetwork, similar to [50], we extend the concept of spatial tokenization in ViTs to the spectral domain, generating spectral tokens by embedding channels. To implement spectral MAE, we adapt the channel tokenization process. Specifically, for a 3-D HSI cube $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$, we first aggregate adjacent channels through average clustering along the channel dimension, resulting in $\mathbf{X}' \in \mathbb{R}^{H\times W\times N_{spec}}$, where N_{spec} is the desired token number. Next, $\mathbf{X}'$ is reshaped into a 2-D matrix of shape $\mathbb{R}^{N_{spec}\times(H\cdot W)}$ by dimensional permutation and flattening. It is then projected into D-dimension embedding space via linear mapping, resulting in $\mathbf{X} \in \mathbb{R}^{N_{spec}\times D}$. $\mathbf{X}$ serves as the spectral channel embedding, analogous to the spatial patch embedding in standard ViT. After spectral tokenization, $\mathbf{X}$ is processed by the subsequent ViT blocks, and the remaining pre-training steps follow those of the MAE. Here, it should be noted that, although we use the HSIs with DN or radiance value for pre-training, the spectral contexts of real-world data that adopts reflectance value still can be captured, because our spectral subnetwork perceives the relationships between channels regardless of the data type.

4.1.4 Implementation

Data Pre-processing Given the high-dimensional nature of HSIs, we perform dimensionality reduction by randomly selecting continuous channels to preserve the original spectral order and enhance the diversity of pre-training data. Since HSIs in the HyperGlobal-450K dataset are acquired by different sensors, our network cannot simultaneously process the images with varying numbers of channels in the same batch. Therefore, the number of selected channels remains consistent across all HSIs. The specific details can be found in the appendix.

Experimental Settings The mask ratio is a crucial hyper-parameter in MIM algorithms, as it influences the difficulty of recovering masked regions and, consequently, the effectiveness of pre-training. A high mask ratio makes the reconstruction task overly challenging, potentially hindering restoration, while a low mask ratio may result in ineffective model weights due to the ease of the task. Thus, selecting an appropriate mask ratio R is essential for ensuring pre-training quality. For spatial MAE, we adopted $R_{spat} = 0.75$, a setting proven effective for both natural images [46] and remote sensing (RS) images [31]. In contrast, research on

HSI channel masking and recovery is limited. For example, [50] independently determined relevant values for different datasets in HSI and LiDAR joint classification. Therefore, we conducted a series of ablation studies to identify a suitable R_{spec} for spectral MAE. We provide the details of determining R_{spec} (0.75 by default) in the appendix.

After determining the mask ratios, we retrained the spatial and spectral subnetworks using the default settings as previously described. Following [31], [46], we extended the pre-training to 1,600 epochs for sufficient training. Given the image size of $\mathbf{X}_0$, the patch size P for the spatial ViT is set to 8, resulting in $N_{spat} = \frac{H \cdot W}{P^2} = 64$ tokens. The other training parameters of the spatial subnetwork, including batch size, learning rate, and weight decay, match those of the spectral subnetwork. We pre-trained different ViT versions - base, large, and huge - for both spatial and spectral networks, denoted as SpatViT and SpecViT, respectively. This allows HyperSIGMA's model size to reach the billion level when using the ViT-Huge backbone for both subnetworks. All experiments were conducted on NVIDIA V100 GPUs. The appendix provides more details about the pre-training cost.

4.2 Model Structure

4.2.1 Preliminaries of Vision Transformer

In this section, we briefly introduce ViTs [29] as they form the foundational structures of HyperSIGMA. Inspired by the transformer network used in NLP, where words are mapped into vectors, ViT operates by first splitting an image into non-overlapping patches. Each patch is then mapped to a 1-D token vector. The details of mapping spatial patches and spectral channels of HSI are detailed in Sec. 4.1.2, Sec. 4.1.3, and Sec. 4.1.4. These tokens, combined with learnable positional embeddings $\mathbf{E}$, are processed by a series of transformer blocks $f = \{f_1, f_2, \cdots, f_d\}$, where d is the network depth. This process can be formulated as:

$$\begin{aligned} \mathbf{U}_0 &= \mathbf{X} + \mathbf{E}, \quad \mathbf{X} \in \mathbb{R}^{N \times D}, \mathbf{E} \in \mathbb{R}^{N \times D}, \\ \mathbf{U}_i &= f_i(\mathbf{U}_{i-1}), \quad \mathbf{U}_i \in \mathbb{R}^{N \times D}, i = 1, \cdots, d, \\ \mathbf{Z} &= \mathrm{LN}(\mathbf{U}_n), \quad \mathbf{Z} \in \mathbb{R}^{N \times D}. \end{aligned} \tag{1}$$

Here, $\mathbf{U}_i$ is the output of the ith block f_i, $\mathbf{Z}$ is the final output feature of the ViT, while LN represents the layer normalization [51]. The computation process of f_i is formulated as:

$$\begin{aligned} \mathbf{U}'_{i-1} &= \mathrm{MHSA}(LN(\mathbf{U}_{i-1})) + \mathbf{U}_{i-1}, \\ \mathbf{U}_i &= \mathrm{FFN}(\mathrm{LN}(\mathbf{U}'_{i-1})) + \mathbf{U}'_{i-1}. \end{aligned} \tag{2}$$

Here, FFN denotes the feed-forward networks containing two linear layers. MHSA is the core module of transformer blocks, containing multiple parallel self-attentions (SAs). Each SA can be formulated as:

$$\mathrm{SA}(\mathbf{U}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D'}})\mathbf{V}, \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{N \times D'}, \mathbf{K} \in \mathbb{R}^{N \times D'}, \mathbf{V} \in \mathbb{R}^{N \times D'}$ are the query, key and value generated from $\mathbf{U} \in \mathbb{R}^{N \times D}$ by three linear layers, respectively. The output feature of MHSA is obtained by concatenating all the outputs of SAs along the channel dimension:

$$\mathrm{MHSA}(\mathbf{U}) = [Concat(\mathrm{SA}_1(\mathbf{U}), \cdots, \mathrm{SA}_h(\mathbf{U}))\mathbf{W}]^T. \tag{4}$$

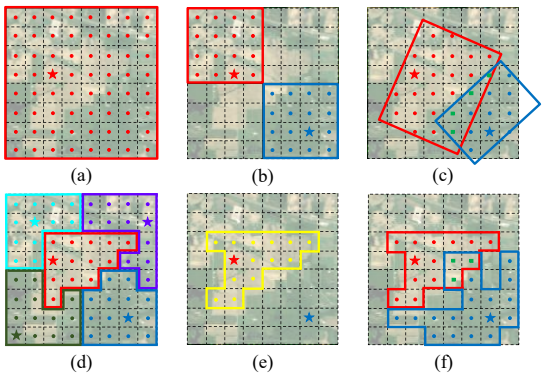


Fig. 5. Comparison of different attention mechanisms: (a) Full SA [29], (b) WMHSA [28], (c) RVSA [31], (d) NLSA [52], (e) DMHA [53], (f) SSA. Stars represent queries, with dots surrounded by corresponding colored lines indicating the attention regions of captured contexts. Green rectangles in (c) and (f) denote common areas shared by both queries. In DMHA, all queries share the same keys in the yellow region.

Here, $\mathrm{SA}_h(\cdot)$ denotes the h-th SA head. $\mathbf{W} \in \mathbb{R}^{hD' \times D}$ denotes the weight matrix of a linear layer to recover the original embedding dimension D. In practice, $D = h \cdot D'$ for convenience.

4.2.2 Sparse Sampling Attention

In this section, we introduce the proposed SSA, designed to efficiently learn diverse contextual features by addressing the spatial and spectral redundancy of HSIs. Given $\mathbf{Q} = \{q_1, \cdots, q_N\}$, $\mathbf{K} = \{k_1, \cdots, k_N\}$, and $\mathbf{V} = \{v_1, \cdots, v_N\}$, we predict N_p offsets $[(\Delta x_1, \Delta y_1), \cdots, (\Delta x_{N_p}, \Delta y_{N_p})]$ using a linear layer $\boldsymbol{W}_p \in \mathbb{R}^{D' \times 2N_p}$ for each query vector q at coordinates (c_x, c_y). Next, we sample new key k' and value v' from the original key and value matrices $\mathbf{K}$ and $\mathbf{V}$ at these positions using bilinear interpolation. Note that $\mathbf{K}$ and $\mathbf{V}$ have been reshaped into 2-D feature maps of shape $\mathbb{R}^{H' \times W' \times D'}$, where $N = H'W'$. This process can be formulated as:

$$\begin{aligned} k'_j = \sum_{(o_x, o_y)} & \max(0, 1 - |o_x - (c_x + \Delta x_j)|) \\ & \max(0, 1 - |o_y - (c_y + \Delta y_j)|)\mathbf{K}[o_x, o_y, :], \\ v'_j = \sum_{(o_x, o_y)} & \max(0, 1 - |o_x - (c_x + \Delta x_j)|) \\ & \max(0, 1 - |o_y - (c_y + \Delta y_j)|)\mathbf{V}[o_x, o_y, :]. \end{aligned} \tag{5}$$

Here, $j = 1, ..., N_p$, $\mathbf{K}[o_x, o_y, :]$ is a vector extracted at (o_x, o_y) from $\mathbf{K}$, where (o_x, o_y) represents all coordinates. We totally sample $N \cdot N_p$ points and obtain $\mathbf{K}', \mathbf{V}' \in \mathbb{R}^{N \times N_p \times D'}$. Consequently, SSA can be formulated as:

$$\mathrm{SSA}(\mathbf{U}) = softmax(\frac{\mathbf{Q} \otimes \mathbf{K}'}{\sqrt{D'}}) \otimes \mathbf{V}'. \tag{6}$$

Here, $\otimes$ contains a series of operations including matrix broadcasting, Hadamard product, and removing dimensions through accumulation.

Comparison to Other Attention Methods Fig. 5 illustrates the differences between the proposed SSA and various attention methods. Full SA (Fig. 5 (a)) captures global context with quadratic complexity relative to token sequence length, whereas window-based attention methods [28], [31] (Fig. 5

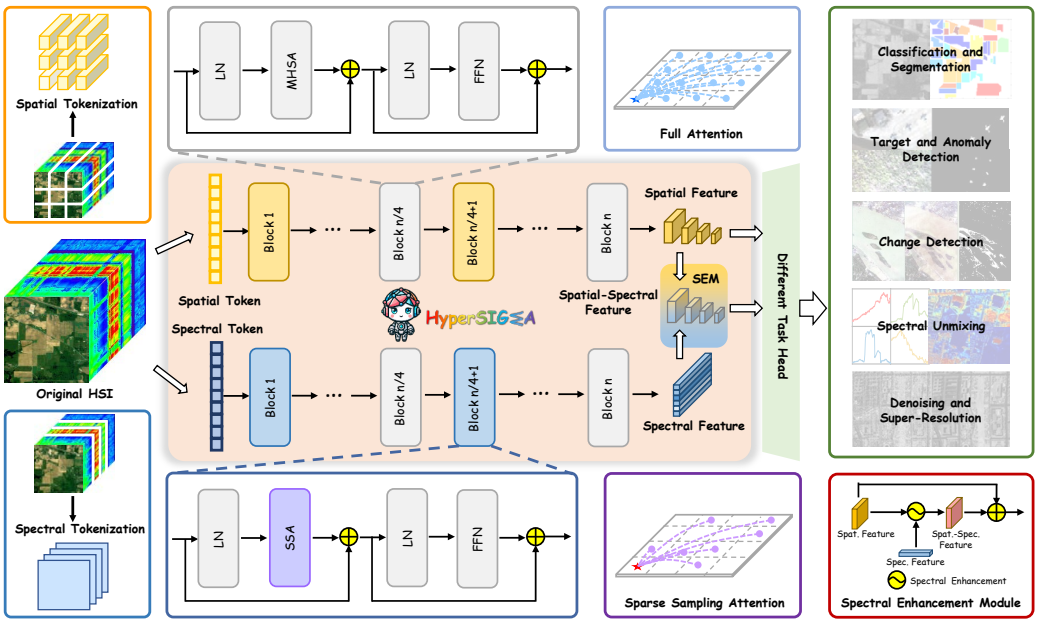


Fig. 6. The HyperSIGMA model consists of two subnetworks tailored for spatial and spectral feature extraction (Sec. 4.1). Initially, spatial patches or spectral channels are tokenized and processed through multiple transformer blocks, where some SAs are substituted with the proposed Sparse Sampling Attention (SSA) (Sec. 4.2.2). Spatial and spectral features are then generated and fused by the Spatial-Spectral Fusion Module (SEM) (Sec. 4.2.3). These fusion features (or spatial features) are fed into task-specific heads for various high-level and low-level HSI tasks.

(b)-(c)) enhance efficiency by focusing on local rectangular regions. NLSA [52] partitions the feature map into non-overlapped attention buckets by spherical locality-sensitive hashing, i.e., each token is assigned to a specific group, where the number of buckets is pre-defined. For efficient parallel computing, it requires the buckets to be further divided into fixed-size chunks. Deformable-DETR [54] introduces deformable attention across features, enhancing context diversity by using adaptive sampling points for each query, primarily for object detection. Inspired by this, DMHA [53] (Fig. 5 (e)) integrates deformable attention into vision transformer backbones, but all queries capture context from the same regions, even when different queries might need different contexts. In contrast, our SSA (Fig. 5 (f)) learns a unique set of sampling positions for each query, further enriching the context. Additionally, unlike deformable attention which uses linear layers to predict attention weights, SSA calculates attention weights through token interaction, enhancing the ability to capture inter-region relationships.

4.2.3 Structure of HyperSIGMA

After obtaining the pre-trained weights of SpatViT and SpecViT (Sec. 4.1), we construct the HyperSIGMA model by replacing the original full SA with the proposed SSA (Sec. 4.2.2) and fusing spatial-spectral features. The overall structure of HyperSIGMA is shown in Fig. 6.

Attention Replacement Following the approach in [31], [49], we retain only the full SA in the $i\cdot(n/4)$th $(i = 1, 2, 3, 4)$ layers, while replacing the full SA in other layers with the proposed SSA. This approach allows the network to effectively extract both local and global contexts.

Feature Fusion In traditional HSI dual-branch processing pipelines [19], it is common practice to combine spatial and spectral features in the late stage. Following this practice, we fuse the features extracted from spatial and spectral subnetworks. It is noteworthy that, in the spectral ViT, each channel is transformed into a 1-D token vector, disrupting the original spatial structures. Our preliminary experiments have shown that attempting to reproject these tokens into 2-D features is ineffective. Therefore, we retain the 1-D shape features for the spectral subnetwork. In our implementation, we fuse the spatial and spectral features using a specially designed approach, the Spectral Enhancement Module (SEM), which enhances spatial features with spectral information.

Specifically, given a spatial feature $\mathbf{F}_{spat} \in \mathbb{R}^{H'\times W'\times D}$ and a spectral feature $\mathbf{F}_{spec} \in \mathbb{R}^{N_{spec}\times D}$, we first compress their dimensions using a linear layer to obtain $\mathbf{F}_{spat} \in \mathbb{R}^{H'\times W'\times D_1}$ and $\mathbf{F}_{spec} \in \mathbb{R}^{N_{spec}\times D_1}$. Next, we reduce the spatial dimension of $\mathbf{F}_{spec}$ through average aggregation to create a 1D vector $\mathbf{V} \in \mathbb{R}^{N_{spec}}$. Finally, we apply another linear layer to align the dimension of $\mathbf{V}$ with the channel number of $\mathbf{F}_{spat}$. The spectral enhancement process is formulated as follows:

$$\mathbf{F}^{\star} = (1 + \mathbf{V}) * \mathbf{F}_{spat}, \tag{7}$$

Here, we reuse the symbol $\mathbf{V}$. $*$ denotes channel-wise product. Note that a skip-connection is employed to retain the original spatial information. In this way, the extracted spatial features are calibrated using spectral information.

SpatSIGMA Since both $\mathbf{F}_{spat}$ and $\mathbf{F}^{\star}$ offer potential for interpreting HSIs, alongside HyperSIGMA, the spatial subnetwork termed SpatSIGMA also forms a foundational model for solving various high-level and low-level HSI tasks.

Remarks The proposed method introduces three key novelties: (1) *Unsupervised Pre-training*: By eliminating the reliance on labeled data, unsupervised pre-training significantly reduces annotation costs. A random channel selection strategy is employed during preprocessing, effectively mitigating the high dimensionality of HSI and enabling adaptation to data

from mixed sensors. (2) *Spatial-Spectral Fusion*: The SEM is designed and integrate spectral information into spatial representations through channel-wise corrections, offering a robust spatial-spectral feature representation. (3) *Adaptive Attention*: The SSA mechanism adaptively determines optimal positions for sampling keys and values, addressing spatial and spectral redundancy in HSIs. By replacing the full attention mechanism in both SpatViT and SpecViT, SSA improves contextual diversity, enhances representational capability, and lowers computational complexity.

5 Experiments

In this section, we thoroughly evaluate HyperSIGMA's performance on representative HSI high-level interpretation tasks: image classification (Sec. 5.1), target detection (Sec. 5.2), anomaly detection (Sec. 5.2), and change detection (Sec. 5.3), and low-level tasks: spectral unmixing (Sec. 5.4), image denoising (Sec. 5.5) and image super-resolution (Sec. 5.6). Additionally, we conduct experiments to examine HyperSIGMA's scalability (Sec. 5.7.1), robustness (Sec. 5.7.2), cross-modal transferring capability (Sec. 5.7.3), real-world applicability (Sec. 5.7.4) and computational efficiency (Sec. 5.7.5).

HyperSIGMA in Fine-tuning Considering the variability in HSIs, where the number of channels differs across datasets, pre-trained weights for the spatial patch embedding layer may not be usable during fine-tuning if the channel numbers do not match. In such cases, we randomly initialize the embedding layer and retrain it during fine-tuning. Notably, the number of channels remains consistent during fine-tuning and inference, utilizing all channels to fully leverage the spectral information. Similarly, pre-trained weights for the spectral tokenization layer cannot be used if the spatial size of inputs for the spectral subnetwork is not consistent during pre-training and fine-tuning. Besides, pre-trained positional embeddings in spatial (spectral) subnetwork will be adjusted through interpolation to fit the spatial size (spectral channel) of the fine-tuning data.

5.1 Hyperspectral Image Classification

Image classification is a fundamental task in HSI interpretation. In this section, we begin with a series of ablation studies (Sec. 5.1.2), followed by fine-tuning pre-trained models on common HSI classification datasets (Sec. 5.1.3).

5.1.1 Experiment Settings

Dataset In addition to the widely used Indian Pines (IP) [55] and Pavia University (PU)[2] datasets, we also utilize three challenging datasets: HanChuan (HC) [56], HongHu (HH) [56], and Houston (HU) [57]. Furthermore, we introduce a new dataset from the Yellow River Delta, captured by the ZY1-02D satellite (ZY) [58].

Implementation The classification task focuses on classifying all pixels in a hyperspectral image, similar to the semantic segmentation tasks in natural or aerial RGB images. This can be achieved through either patch-level or image-level classification [59]. Patch-level classification selects a patch centered around each pixel and classifies the center pixel based on this patch, while image-level classification corresponds directly to semantic segmentation. For computational efficiency, we primarily employ semantic segmentation unless otherwise specified.

2. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

TABLE 1
OA (%) of various methods across HSI classification datasets. **Best** and **2nd-best** results are highlighted.

Method	IP	PU	HC	HH	HU	ZY
MSDN [60]	57.54	76.48	73.40	78.55	72.18	88.57
SSFCN [61]	41.93	78.88	63.35	71.62	72.39	82.46
FullyContNet [59]	71.11	80.31	78.80	67.12	51.07	84.47
SpectralFormer [20]	50.02	75.37	82.60	85.33	77.21	72.03
HSIC-FM [62]	36.02	77.28	66.21	70.47	54.43	76.98
SSGRN [63]	69.58	81.45	90.43	82.19	68.62	77.46
CSIL [64]	66.53	88.23	88.55	91.86	66.11	92.49
IDCN [65]	71.12	91.64	84.15	89.19	85.34	92.29
CLOLN [66]	72.75	93.11	86.73	87.89	85.95	80.43
SpatSIGMA	**85.08**	**93.36**	**94.03**	**94.35**	**87.33**	**94.72**
HyperSIGMA	**85.54**	**93.52**	**94.44**	**94.87**	**86.80**	**94.92**

Classification performance is assessed using the common overall accuracy (OA). We compare HyperSIGMA against several classical and advanced classification and segmentation methods, including MSDN [60], SSFCN [61], FullyContNet [59], SpectralFormer [20], HSIC-FM [62], SSGRN [63], CSIL [64], IDCN [65], and CLOLN [66]. The details related to network structure and experimental configurations are presented in the appendix.

5.1.2 Ablation Studies

To determine effective hyperparameter settings, we conduct a series of ablation experiments on the IP and PU datasets. For simplicity, we use only the spatial subnetwork. According to the ablation study (see the appendix), the number of sampling points is set to 8. Additionally, we prove the effectiveness of the proposed SSA and the necessity for pre-training on the constructed large-scale hyperspectral dataset, i.e., HyperGlobal-450K.

5.1.3 Results and Analyses

Table 1 presents the classification accuracies of different methods. Our models consistently outperform state-of-the-art approaches. For example, on the HanChuan dataset, containing 16 visually similar agricultural categories, models were trained with 50 labeled samples per class (about 0.22% of the entire image, please see the appendix). Despite these constraints, our approach still outperforms the recent graph convolutional network [67] based method SSGRN [63], which achieves an OA of 90.43% but lags 4% behind HyperSIGMA. Apart from the Houston dataset, HyperSIGMA consistently outperforms SpatSIGMA across diverse scenes, leveraging spectral information to improve accuracy. In summary, our models exhibit significant advancements in extracting robust and universal representations for HSI classification, demonstrating notable performance improvements. The appendix provides more qualitative classification results to demonstrate the performance of HyperSIGMA more clearly.

5.2 Hyperspectral Target and Anomaly Detection

Unlike typical object detection for optical images, which identifies objects in natural or aerial images by predict-

TABLE 2
AUC (%) of various methods for the HTD and HAD tasks. **Best** and **2nd-best** results are highlighted.

Target Detection				Anomaly Detection			
Method	Mosaic	AVIRIS	Renourban	Method	Pavia	Cri	Viareggio
CEM [76]	97.92	**99.83**	95.52	RX [83]	99.82	96.75	95.25
STD [77]	99.30	93.98	82.30	KIFD [84]	93.33	**99.28**	96.79
CSCR [78]	98.95	99.35	87.60	CRD [85]	99.33	91.64	95.87
SRBBH [79]	93.33	84.43	92.89	GTVLRR [86]	99.49	88.25	94.54
HTD-Net [80]	89.70	98.80	94.43	Auto-AD [87]	99.79	97.86	97.28
HTD-IRN [81]	99.30	98.98	97.65	RGAE [88]	**99.94**	97.09	79.68
CGSAL [82]	99.44	98.98	99.34	GT-HAD [89]	99.87	94.39	93.05
SpatSIGMA	**99.75**	**99.78**	**99.79**	SpatSIGMA	**99.94**	98.14	**98.44**
HyperSIGMA	**99.56**	**99.83**	**99.86**	HyperSIGMA	**99.96**	**99.32**	**98.32**

ing bounding boxes, hyperspectral detection tasks utilize unique spectral information from HSIs to identify target regions of interest. These tasks commonly include hyperspectral target detection (HTD) and hyperspectral anomaly detection (HAD). HTD aims to locate areas with spectral characteristics similar to known target spectra, whereas HAD identifies anomalies based on spectral differences from the surrounding environment, without requiring prior target spectra. Both tasks share the goal of detecting specific regions of interest. In this section, to demonstrate HyperSIGMA's transferability, we conduct fine-tuning experiments on both HTD and HAD tasks.

5.2.1 Experiment Settings

Dataset We adopt Mosaic [68], AVIRIS [69], and Renourban [70] datasets for the HTD tasks, and Pavia [71], Cri [72], and Viareggio [73] for the HAD tasks.

Implementation Following HTD-ViT [74], we tackle the hyperspectral detection task using coarse detection labels. Unlike HTD-ViT, which employs per-pixel classification, we draw inspiration from promptable segmentation [75] to design a segmentation network utilizing HyperSIGMA as a feature extractor. By reframing hyperspectral detection as a segmentation problem, we evaluate HyperSIGMA's performance in both HTD and HAD tasks.

We compare our models, SpatSIGMA and HyperSIGMA, with existing classical and advanced HTD methods: CEM [76], STD [77], CSCR [78], SRBBH [79], HTD-Net [80], HTD-IRN [81], and CGSAL [82], as well as HAD methods including RX [83], KIFD [84], CRD [85], GTVLRR [86], Auto-AD [87], RGAE [88], and GT-HAD [89]. The area under the ROC curve (AUC) is used as the evaluation metric. Please refer to the appendix for detailed network structure and experimental configurations.

5.2.2 Results and Analyses

The results, shown in Table 2, demonstrate that our models outperform all existing methods in both HTD and HAD tasks. HyperSIGMA, with enhanced spectral information, shows further accuracy improvements over SpatSIGMA in most cases. In summary, our models achieve state-of-the-art performance in hyperspectral detection tasks, demonstrating their effectiveness of transferability. The appendix offers further analyses of the detection results.

5.3 Hyperspectral Change Detection

In this section, we fine-tune pre-trained models for hyperspectral change detection, which parallels conventional remote sensing tasks using aerial RGB images [90], [91]. Our focus here is primarily on classical bi-temporal scenarios.

TABLE 3
F1 (%) of various methods on hyperspectral change detection tasks. **Best** and **2nd-best** results are highlighted.

Method	Hermiston	Farmland	BA	SB
CVA [93]	81.03	92.49	87.09	83.76
ISFA [94]	72.62	93.01	89.05	85.35
GETNET [95]	90.54	94.39	96.07	96.59
ML-EDAN [96]	89.03	93.93	97.41	97.80
BIT [97]	71.28	84.56	94.69	96.19
EMS-Net [98]	85.67	92.97	97.67	97.32
CSA-Net [99]	85.29	93.27	97.80	98.24
SST-Former [92]	88.80	94.34	97.23	97.58
GlobalMind [100]	90.71	94.98	98.26	98.28
SpatSIGMA	**92.08**	**95.31**	**98.89**	**98.95**
HyperSIGMA	**91.74**	**95.49**	**98.93**	**99.04**

5.3.1 Experiment Settings

Dataset We tested on four common benchmark datasets: Hermiston, Farmland, Bay Area (BA), Santa Barbara (SB), which are publicly available[3][4].

Implementation Following the successful practice of SST-Former [92], a transformer-based method for hyperspectral change detection, we adopt a classification network. We compared our models against several established methods: CVA [93], ISFA [94], GETNET [95], ML-EDAN [96], BIT [97], EMS-Net [98], CSA-Net [99], SST-Former [92], and GlobalMind [100]. Model performance was evaluated using F1 score. Please refer to the appendix for detailed network structure and hyperparameter settings.

5.3.2 Results and Analyses

As shown in Table 3, our models achieve the highest F1 scores across all datasets, demonstrating their clear superiority over other methods. Compared to SpatSIGMA, HyperSIGMA further improves its performance on three datasets. Overall, our models deliver finer and more complete detection results, demonstrating HyperSIGMA's strong feature representation for hyperspectral change detection tasks. Please refer to the appendix for more results.

5.4 Hyperspectral Unmixing

After applying HyperSIGMA to high-level downstream tasks, we extended its use to low-level tasks, starting with hyperspectral unmixing [101]. This task aims to address the complex spectral mixtures in hyperspectral data by decomposing each pixel's spectral signature into pure spectral signatures (endmembers) and their corresponding proportions (abundances) [102]–[104]. This facilitates the identification and quantification of various components in each pixel.

5.4.1 Experiment Settings

Dataset We adopt the Urban dataset[5] to evaluate the performance of HyperSIGMA for hyperspectral unmixing.

Implementation Following [108], we employ a simple encoder-decoder structure for hyperspectral unmixing. Seven advanced linear and nonlinear unmixing approaches were chosen for comparison. These include three classical methods: FCLS [105], ELMM [106], and SUnSAL [107],

3. https://citius.usc.es/investigacion/datasets/hyperspectral-change-detection-dataset
4. https://rslab.ut.ac.ir/data
5. https://rslab.ut.ac.ir/data

TABLE 4
Quantitative comparison of endmember and abundance prediction performance across various methods on the Urban dataset. **Best** and **2nd-best** results are highlighted.

Method	Abundance	Endmember
FCLS [105]	0.1406	0.4373
ELMM [106]	0.0334	-
SUnSAL [107]	0.1859	-
CNNAEU [108]	0.0216	0.0865
CyCU [109]	0.0821	0.2952
DeepTrans [110]	0.0488	0.1512
EGUnet [111]	0.1732	0.5796
SpatSIGMA	**0.0176**	**0.0598**
HyperSIGMA	**0.0157**	**0.0584**

TABLE 5
Quantitative comparison of different HSI denoising methods on the WDC Mall dataset with non-i.i.d Gaussian, impulse, stripe, and deadline noises. **Best** and **2nd-best** results are highlighted.

Method	PSNR	SSIM	SAM
Noisy	10.984	0.351	0.719
LLRT [118]	21.933	0.824	0.221
NGMeet [119]	23.452	0.872	0.260
LRTFL0 [120]	25.558	0.907	0.198
E-3DTV [121]	25.966	0.919	0.117
DS2DP [122]	26.741	0.931	0.152
QRNN3D [123]	28.097	0.945	0.132
SST [124]	28.061	0.957	0.111
SpatSIGMA	**28.103**	**0.959**	**0.109**
HyperSIGMA	**28.503**	**0.961**	**0.106**

as well as recent deep learning-based methods: CNNAEU [108], CyCU-Net [109], DeepTrans [110], and EGU-Net [111]. We used two common quantitative evaluation indices: the mean spectral angle distance (mSAD) to compare the similarity between the learned endmembers and references, and the mean squared error (MSE) to measure the quality of the obtained abundance map. Note that lower values indicate better performance. More details about the task setting, network structure, and hyperparameter settings are presented in the appendix.

5.4.2 Results and Analyses

The evaluation results are shown in Table 4, where we only list the average value across all endmembers. It can be seen that HyperSIGMA achieves the best performance for both endmembers and abundances, demonstrating its superior feature representation capability. HyperSIGMA consistently outperforms SpatSIGMA with the help of spectral information, which is significant for the unmixing task. These findings highlight the potential of our models to enhance hyperspectral low-level tasks. The appendix provides more results and analyses for each endmember.

5.5 Hyperspectral Image Denoising

Denoising, a fundamental low-level vision task, is widely discussed across various data forms, such as natural [112], [113], raw [114], [115] and thermal images [116], [117]. In this section, we consider a classic hyperspectral image restoration task - HSI denoising, which aims at recovering a clean HSI from its noisy observation.

TABLE 6
Quantitative comparison of different methods on the Houston dataset at 8× scale factors. **Best** and **2nd-best** results are highlighted.

Method	PSNR	SSIM	SAM	CC	RMSE	ERGAS
Bicubic	38.108	0.899	4.671	0.918	0.015	5.123
GDRRN [125]	38.259	0.909	4.905	0.914	0.014	4.914
SwinIR [126]	39.401	0.920	4.059	0.937	**0.013**	4.333
SSPSR [127]	39.284	0.916	4.267	0.935	**0.013**	4.421
RFSR [128]	39.490	0.921	3.840	0.938	**0.013**	4.297
Gelin [129]	39.639	0.921	3.923	0.940	**0.013**	4.245
MSDFormer [130]	39.745	0.923	3.661	0.942	**0.012**	4.211
SpatSIGMA	**39.895**	**0.926**	**3.628**	**0.943**	**0.012**	**4.121**
HyperSIGMA	**39.940**	**0.927**	**3.552**	**0.944**	**0.012**	**4.101**

5.5.1 Experiment Settings

Dataset For the HSI denoising experiment, we utilize the widely used Washington DC Mall (WDC Mall) dataset[6], which contains images with 191 spectral bands and a spatial resolution of 1,208×307.

Implementation The comparison methods include classical model-based approaches (LLRT [118], NGMeet [119], LRTFL0 [120], and E-3DTV [121]), and recent deep learning-based methods (DS2DP [122], QRNN3D [123], and SST [124]). We use three common evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structure SIMilarity (SSIM), and Spectral Angle Mapper (SAM). Higher PSNR and SSIM values and lower SAM values indicate better denoising performance. The appendix provides more details about network structure and experimental settings.

To demonstrate the effectiveness of our model, we consider a challenging mixed noise case that contains four types of noises, incorporating Gaussian noise, impulse noise, stripes, and deadlines. Details can be found in the appendix.

5.5.2 Results and Analyses

As shown in Table 5, our proposed methods greatly outperform other approaches. For instance, HyperSIGMA achieves a higher PSNR than the second-best method, SST [124]. Even SpatSIGMA, the simplified version using only spatial information from the pre-trained RS model, performs comparably to SST. We attribute these results to the superior feature representations by pre-training models, especially the utilization of spectral information. For additional results in various noisy cases, please see the appendix.

5.6 Hyperspectral Image Super-Resolution

In addition to HSI denoising, we investigate another fundamental low-level task: HSI super-resolution, which aims to produce a clearer HSI from its low-resolution observation.

5.6.1 Experiment Settings

Dataset We employ the Houston dataset[7] that contains 48 channels with a spatial resolution of 4,172×1,202 for the super-resolution task.

Implementation In our experiments, we first acquire training patches. These patches are uniformly downsampled using bicubic interpolation. We then upsample the downsampled patches back to the original size using various super-resolution methods. Specifically, we compare our SpatSIGMA and HyperSIGMA with several advanced methods,

6. http://lesun.weebly.com/hyperspectral-data-set.html
7. https://hyperspectral.ee.uh.edu/?pageid=1075

TABLE 7
Fine-tuning accuracies of SpatSIGMA and HyperSIGMA with different ViT backbones on Indian Pines and Xiongan datasets. Base: ViT-Base. Large: ViT-Large. Huge: ViT-Huge.

Method	SpatViT	SpecViT	OA	AA	Kappa
Segmentation on Indian Pines					
SpatSIGMA	Base	-	85.08	78.30	83.04
SpatSIGMA	Large	-	86.55	81.70	84.63
SpatSIGMA	Huge	-	86.74	75.02	84.89
HyperSIGMA	Base	Base	85.54	76.68	83.58
HyperSIGMA	Base	Large	86.81	79.49	85.00
HyperSIGMA	Base	Huge	87.38	83.24	85.63
HyperSIGMA	Huge	Huge	87.70	79.19	86.01
Classification on Xiongan					
1DCNN [18]	-	-	55.89	71.19	52.26
2DCNN [18]	-	-	27.32	38.79	23.61
HybridSN [131]	-	-	89.83	74.72	88.39
A^2S^2KResNet [132]	-	-	90.32	76.85	88.93
SpatSIGMA	Base	-	93.68	79.80	92.75
SpatSIGMA	Large	-	94.59	82.95	93.78
SpatSIGMA	Huge	-	94.73	82.43	93.94
HyperSIGMA	Base	Base	95.06	82.43	94.32
HyperSIGMA	Base	Large	95.30	82.77	94.58
HyperSIGMA	Base	Huge	95.65	83.85	94.98
HyperSIGMA	Huge	Huge	95.78	84.19	95.13

including Bicubic, SwinIR [126], GDRRN [125], SSPSR [127], RFSR [128], GELIN [129], and MSDFormer [130]. Here, we adopt a difficult case: 8× super-resolution to better test the model performance.

We evaluate the models using six popular metrics to comprehensively assess performance in both spatial and spectral dimensions: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Spectral Angle Mapper (SAM), Cross Correlation (CC), Root-Mean-Squared Error (RMSE), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS). While PSNR, SSIM, and RMSE are standard metrics for natural image restoration, CC, SAM, and ERGAS are common in hyperspectral image fusion tasks. Superior super-resolution performance is indicated by higher PSNR, SSIM, and CC values, and lower SAM, RMSE, and ERGAS values. The appendix offers details of network structure and experimental settings.

5.6.2 Results and Analyses

The quantitative results of HSI super-resolution are shown in Table 6. We can find that the proposed HyperSIGMA significantly outperforms other methods. Notably, it surpasses the recent advanced approach, MSDFormer [130], across all metrics. These findings further demonstrate the capability of our models in HSI low-level tasks. The appendix showcases the super-resolution results, and more results under other upsampling ratios.

5.7 Further Investigations and Analyses

5.7.1 Model Scalability

In previous sections, we have shown that HyperSIGMA consistently outperforms state-of-the-art methods, even when using the ViT-Base backbone. To evaluate the scalability of HyperSIGMA, we fine-tuned SpatSIGMA and HyperSIGMA using different ViT backbones on the Indian Pines dataset and the more challenging Xiongan dataset [133], which is 1,580×3,750 pixels with 256 channels and 19 categories, mainly comprising cash crops. We also implemented some classical methods for comparison, including 1DCNN and 2DCNN [18], HybridSN [131], and A^2S^2KResNet [132].

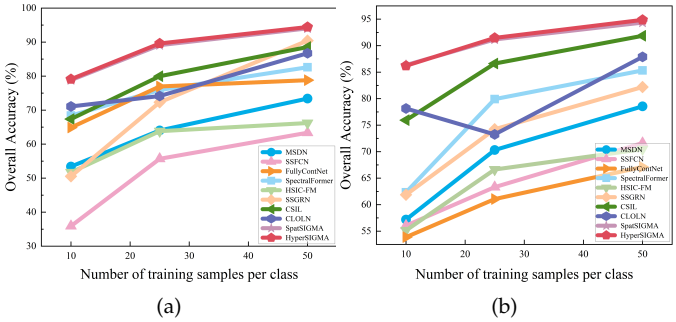


Fig. 7. Overall accuracies of different methods with limited training samples: (a) HanChuan dataset. (b) HongHu dataset.

Table 7 presents the results. It is clear that using larger ViT backbones leads to better performance for both SpatSIGMA and HyperSIGMA, underscoring their good scalability. Notably, we observed that ViT-Huge in SpatSIGMA did not perform as well as lighter versions used in HyperSIGMA, highlighting the importance of utilizing spectral information. The largest HyperSIGMA, with over 1 billion parameters, achieved 87.70% OA and 95.78% OA on Indian Pines and Xiongan, respectively, surpassing other methods significantly. The experimental settings and more analyses related to the model scalability can be found in the appendix.

5.7.2 Model Robustness

Fine-Tuning with Limited Training Samples To evaluate model robustness, we conducted experiments using limited samples for training. Annotating HSI is challenging and costly due to varying ambient light, spectral nonlinearities, and diverse object types, leading to a scarcity of labeled samples [134]. Therefore, it is crucial to assess model performance with limited training samples. We used the HanChuan and HongHu datasets, with 10, 25, and 50 samples per class for training.

Figure 7 presents the accuracies of different methods on these datasets with limited training samples. Our models consistently achieved the highest accuracy regardless of the number of training samples. For example, on the HanChuan dataset, when only 10 samples per class were used (0.04% of the total samples), HyperSIGMA achieved 79.10% accuracy, which is 8% higher than the second-best approach, CLOLN [66]. Moreover, our methods present minimal accuracy decreases under the reduction of labels. In summary, our model demonstrates strong robustness even with limited training samples.

Robustness to Adversarial Attack Previous research has shown that HSI interpretation methods using deep neural networks are vulnerable to adversarial attack [135], [136]. This motivates us to evaluate the stability of our proposed models against such attacks.

We use the Indian Pines dataset, commonly utilized in adversarial attack studies, for our experiment. For training, we select 50 samples per class to better highlight accuracy variations, with the remainder used as the clean testing set. We employ two classical attack methods: FGSM [139] and PGD [140], with perturbation budgets (ϵ) uniformly set to 0.1. Our models are compared with common classification methods as well as those specially designed to defend

TABLE 8
OA (%) of different methods under various attacks on the Indian Pines dataset. CTS: Clean Test Set. **Best** and **2nd-best** results are highlighted.

Method	CTS	FGSM	PGD
SSFCN [61]	82.57	60.26	1.48
SACNet [135]	88.63	80.19	30.18
RCCA [137]	94.51	91.33	82.96
S^3ANet [138]	96.23	95.48	42.01
SpatSIGMA	**96.28**	**96.19**	**95.23**
HyperSIGMA	**96.76**	**96.62**	**95.92**

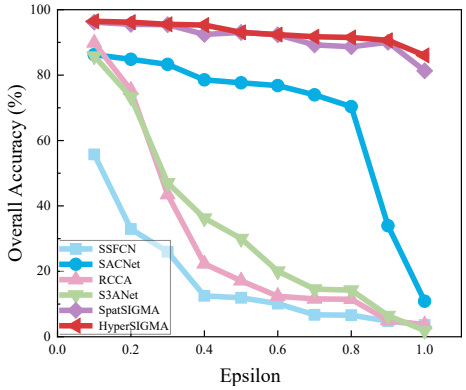


Fig. 8. Classification performance of various methods under different FGSM attack perturbation values on the Indian Pines dataset.

against adversarial attacks, including SSFCN [61], SACNet [135], RCCA [137], and S^3ANet [138]. We fine-tune using the AdamW optimizer, with an initial learning rate of 0.0005 and a weight decay of 0.00005, for 1,000 epochs.

Table 8 presents the classification accuracy on the original clean testing set and under various attacks. Our models show minimal accuracy decreases and consistently outperform others, regardless of the attack type. Compared to SpatSIGMA, HyperSIGMA, which leverages spectral information, achieves even higher accuracy. To further assess model stability, we increase the FGSM perturbation budget ϵ and plot the accuracy changes in Figure 8. Most comparison methods' accuracies decline rapidly when ϵ reaches 0.4. In contrast, our models maintain stability, with only slight accuracy changes even at ϵ of 1.0. These results clearly demonstrate the robustness of our models against adversarial attacks.

Robustness to Image Degradation We further evaluate the stability of HyperSIGMA by testing its performance on degraded images. Specifically, we use two methods to degrade hyperspectral images. First, we apply image compression for saving storage space and transmission bandwidth. Evaluating model performance on compressed HSIs is of practical significance, as current HSI approaches under image compression remain under-explored. We conduct experiments using JPEG compression at a bit rate of 0.33901, a widely-used standard for image compression [141]. Second, we consider complex imaging conditions such as sunlight, atmosphere, and terrain, which often result in noisy HSIs. To simulate real-world scenarios, we manually add i.i.d zero-mean Gaussian noise with a variance of 70. Image compression may eliminate valuable data, whereas noise can cause interference, thereby reducing image quality.

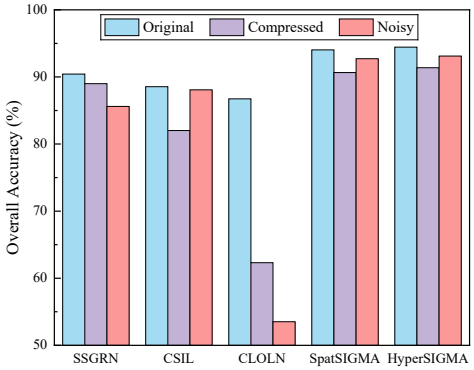


Fig. 9. Classification performance of various methods under different image degradation strategies on the HanChuan dataset.

We selected competitive approaches, SSGRN [63], CSIL [64] and CLOLN [66] for comparison in the classification task on the HanChuan dataset. The results in Fig. 9 show a noticeable decline in performance for comparison methods when subjected to image compression or noise addition. In contrast, the OAs of SpatSIGMA and HyperSIGMA remain stable. Even under degraded image conditions, our models consistently outperform other state-of-the-art methods, highlighting the robustness of HyperSIGMA in practical settings. It should be noted that unlike previous adversarial attacks, which typically affect only the testing set, the data degradation in this study alters the entire HSI, demonstrating the comprehensive robustness of our approach.

5.7.3 *Cross-modal Transferability*

Given the similarity between HSIs and multispectral images (MSIs), we hypothesized that HyperSIGMA could be effectively applied to multispectral tasks. To test this, we conducted experiments on two related tasks: Hyperspectral-Multispectral Collaborative Classification and Multispectral Change Detection.

Hyperspectral-Multispectral Collaborative Classification We investigate the integration of MSIs with HSIs for classification tasks. Specifically, we utilize both types of imagery collaboratively and compares its performance against several state-of-the-art methods, including CGCDL [142], CDLS [143], Cospace1 [144], Cospace2 [145], J-C [146] and SPDDA [147], using the YRE_2 dataset [147] (see Table 9). It can be seen that both SpatSIGMA and HyperSIGMA achieve competitive performances, demonstrating the superior abilities of our models in adapting the out-of-modal data.

Multispectral Change Detection In this task, we utilize the OSCD dataset [148], which consists of Sentinel-2 images. We opted for a simple approach [149] to highlight HyperSIGMA's feature extraction capabilities, avoiding the influence of a complex change decoder. Table 10 compares HyperSIGMA's performance with leading methods, including FC-EF [150], FC-Siam-Diff [150], FC-Siam-Conc [150], SiamCRNN [151], SNUNet [152], DSIFN [153], BIT [97], ChangeFormer [154], and ChangeMamba [155]. We also compared our method with SpectralGPT [3], a recent model tailored for multispectral data. HyperSIGMA surpassed the state-of-the-art ChangeMamba, achieving an F1 score of 58.53% versus 57.20%. Further enhancement using

TABLE 9
Collaborative classification results of different methods on the YRE_2 dataset. **Best** and **2nd-best** results are highlighted.

Method	OA	AA	Kappa
CGCDL [142]	69.01	60.10	58.42
CDLS [143]	50.42	47.76	36.95
Cospace1 [144]	70.42	63.14	60.62
Cospace2 [145]	71.81	63.49	62.16
J-C [146]	72.82	63.91	63.15
SPDDA [147]	75.29	66.02	66.46
SpatSIGMA	**75.51**	**78.66**	**67.69**
HyperSIGMA	**75.67**	**78.49**	**67.83**

TABLE 10
Change detection results of different methods on the OSCD dataset. SpectralGPT accuracy values are sourced from [3]. **Best** and **2nd-best** results are highlighted.

Method	OA	Kappa	F1	Precision	Recall
FC-EF [150]	94.80	45.10	47.83	49.67	46.12
FC-Siam-Diff [150]	94.06	45.26	48.38	43.91	53.85
FC-Siam-Conc [150]	94.55	47.57	50.43	47.60	53.62
SiamCRNN [151]	95.53	52.08	54.42	57.58	51.60
SNUNet [152]	93.68	39.70	43.03	40.29	46.16
I3PE [149]	95.57	50.27	52.57	58.79	47.54
DSIFN [153]	**96.04**	52.22	54.21	**67.32**	45.38
BIT [97]	94.88	43.82	41.19	50.56	38.66
ChangeFormer [154]	95.40	52.15	54.57	55.66	53.51
ChangeMamba [155].	**96.04**	54.82	57.20	56.08	**58.36**
SpectralGPT [3]	-	-	54.29	52.39	57.20
SpatSIGMA	**96.08**	**56.49**	**58.53**	**64.59**	53.50
HyperSIGMA	95.78	**57.06**	**59.28**	59.12	**59.45**

spectral information boosted HyperSIGMA's performance to an F1 score of 59.28%, marking a 4.99% improvement over SpectralGPT. The implementation details of the cross-modal transferability experiment can be found in the appendix.

5.7.4 Real-world Applicability

We further evaluated the model's performance in practical applications, specifically focusing on offshore oil leak detection. As a case study, we examined the Gulf of Mexico oil spill on April 20, 2010, where over 200 million gallons of crude oil were released, marking the largest offshore oil disaster in U.S. history. HSIs are advantageous for monitoring such oil spills due to their rich spectral information. We utilized a hyperspectral dataset related to this event [156], covering the Gulf of Mexico (GM) area. Following the same implementation procedure as previous segmentation tasks, we used 50 pixel-level samples each for both the oil spill area and the background seawater as training data. Fig. 10 presents the results, including pseudo-color images of the study area, our detection results, and ground truth. The results demonstrate that SpatSIGMA and HyperSIGMA effectively distinguish between seawater and oil film, even in areas with minimal oil leakage, enabling precise detection. Notably, HyperSIGMA outperformed SpatSIGMA due to its use of spectral information. These findings strongly support the universality and practical potential of our models.

5.7.5 Model Computational Efficiency

Finally, we analyze the computational cost of our models. As discussed in Section 2, vision transformer has become a primary choice for RS foundation models, with many methods directly adopting or advancing the original ViT structure [3], [27], [29], [157], [158] through modifications

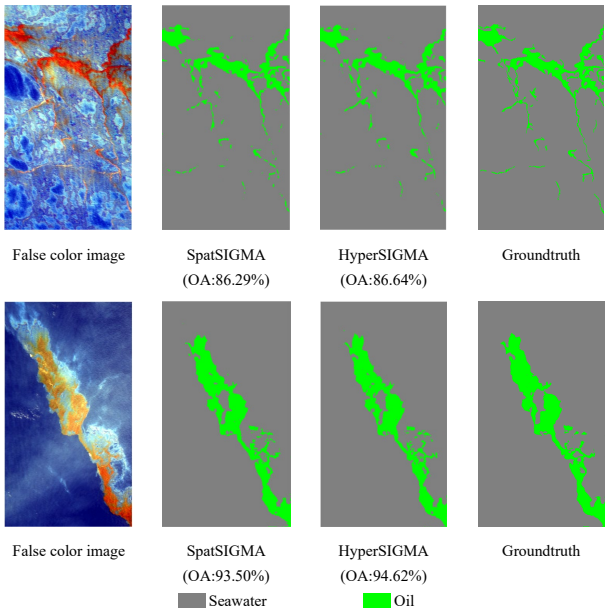


Fig. 10. Visualization of detected oil leakage in various Gulf of Mexico regions by our models. The first row: GM13. The second row: GM18.

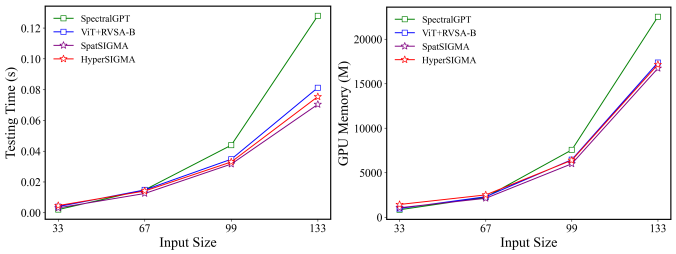


Fig. 11. Inference time and memory footprint of different models with varying input patch sizes, where inference time is measured for a single patch. All methods use a ViT-B backbone, and experiments are conducted on an NVIDIA V100 GPU. (a) Inference time. (b) GPU memory.

[31], [159]. Since HyperSIGMA is also built on ViT, the main distinction between our models and existing methods lies in the attention mechanism. Theoretically, as outlined in Section 4.2.2, the proposed SSA has a complexity of $O(3N^2D' + 12NN_pD' + NN_p)$ (ignoring the number of heads), which is comparable to window attention [28], VSA [159] and RVSA [31]. To validate this, we measure the inference time and GPU memory footprint of our models and typical RS foundation models, including SpectralGPT [3] and RVSA [31], as shown in Fig. 11. The results show that our models exhibit similar inference time and memory footprint, with a slight advantage for larger patches, even for the two-branch HyperSIGMA, when compared to ViT-B+RVSA. These advantages are particularly evident when compared to SpectralGPT, which employs full attention, especially with larger patches. These findings underscore the efficiency of our models. Further SSA complexity analysis and experimental details are provided in the appendix.

6 Conclusion

In this paper, we present HyperSIGMA, the first hyperspectral foundation model with over 1 billion parameters for hyperspectral image (HSI) interpretation. Furthermore, we curate HyperGlobal-450K, the largest hyperspectral dataset to date, composed of HSIs from around the world, establishing a solid basis for self-supervised pre-training research. To address HSI redundancy, we propose a novel sparse sampling attention mechanism, enabling adaptive perception of relevant contextual regions with few learnable sampling points. We also design a spectral enhancement module to achieve effective spatial-spectral feature fusion. Comprehensive evaluations on both high-level and low-level hyperspectral tasks demonstrate HyperSIGMA's superior performance. Further analysis reveals its excellent scalability, robustness, cross-modal transferability, and computational efficiency, making it suitable for various real-world applications.

We find that HyperSIGMA offers only limited improvements over SpatSIGMA in some cases. This may be due to the challenges of recovering complete channels for pre-training the spectral subnetwork. Further investigation will delve into refining both the pre-training pretext task and network architecture to enhance the spectral foundation model. Despite this, we believe our findings provide valuable guidance to the hyperspectral community on advancing foundation models. We hope that HyperSIGMA, with its superior performance and excellent properties, will be widely adopted in various applications.

Acknowledgement

The model pre-training was supported by The Dawning Information Industry Co., Ltd., with fine-tuning conducted on the supercomputing system at the Supercomputing Center of Wuhan University. Thanks Wentao Jiang from the School of Computer Science at Wuhan University for implementing the accelerated version of HyperSIGMA (Sec. 5.7.5).

References

[1] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[2] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and f. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[3] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.

[4] B. Manifold, S. Men, R. Hu, and D. Fu, "A versatile deep learning architecture for classification and label-free prediction of hyperspectral images," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 306–315, 2021.

[5] X. Sun, D. Yin, F. Qin, H. Yu, W. Lu, F. Yao, Q. He, X. Huang, Z. Yan, P. Wang *et al.*, "Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery," *Nature Communications*, vol. 14, no. 1, p. 1444, 2023.

[6] A. M. Lechner, G. M. Foody, and D. S. Boyd, "Applications in remote sensing to forest ecology and management," *One Earth*, vol. 2, no. 5, pp. 405–412, 2020.

[7] D. M. Griffith, K. B. Byrd, L. D. Anderegg, E. Allan, D. Gatziolis, D. Roberts, R. Yacoub, and R. R. Nemani, "Capturing patterns of evolutionary relatedness with reflectance spectra to model and monitor biodiversity," *Proceedings of the National Academy of Sciences*, vol. 120, no. 24, p. e2215533120, 2023.

[8] S. van der Linden and P. Hostert, "The influence of urban structures on impervious surface maps from airborne hyperspectral data," *Remote Sensing of Environment*, vol. 113, no. 11, pp. 2298–2305, 2009.

[9] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4117–4128, 2016.

[10] W. Obermeier, L. Lehnert, M. Pohl, S. Makowski Gianonni, B. Silva, R. Seibert, H. Laser, G. Moser, C. Müller, J. Luterbacher, and J. Bendix, "Grassland ecosystem services in a changing environment: The potential of hyperspectral monitoring," *Remote Sensing of Environment*, vol. 232, p. 111273, 2019.

[11] J. Xia, L. Bombrun, T. Adalı, Y. Berthoumieu, and C. Germain, "Spectral–spatial classification of hyperspectral images using ica and edge-preserving filter via an ensemble strategy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4971–4982, 2016.

[12] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[13] C.-I. Chang, Q. Du, T.-L. Sun, and M. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2631–2641, 1999.

[14] L. Bruce, C. Koger, and J. Li, "Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2331–2338, 2002.

[15] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

[16] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4186–4201, 2015.

[17] J. Li, X. Wang, H. Zhao, and Y. Zhong, "Learning a cross-modality anomaly detector for remote sensing imagery," *IEEE Transactions on Image Processing*, vol. 33, pp. 6607–6621, 2024.

[18] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[19] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5893–5909, 2018.

[20] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[21] Y. Liu, J. Xiao, Y. Guo, P. Jiang, H. Yang, and F. Wang, "HSID-Mamba: Exploring bidirectional state-space models for hyperspectral denoising," *arXiv preprint arXiv:2404.09697*, 2024.

[22] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang *et al.*, "Towards artificial general intelligence via a multimodal foundation model," *Nature Communications*, vol. 13, no. 1, p. 3094, 2022.

[23] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, vol. 30, 2017.

[25] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive bias," in *NeurIPS*, vol. 34, 2021, pp. 28 522–28 535.

[26] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recog-

nition and beyond," *International Journal of Computer Vision*, pp. 1–22, 2023.

[27] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *CVPR*, June 2022, pp. 12 104–12 113.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[30] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[31] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[32] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[33] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–17, 2024.

[34] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, H. He, J. Wang, J. Chen, M. Yang, Y. Zhang, and Y. Li, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *CVPR*, 2024, pp. 27 672–27 683.

[35] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling up capacity and resolution," in *CVPR*, June 2022, pp. 12 009–12 019.

[36] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *CVPR*, 2023, pp. 14 408–14 419.

[37] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

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

[39] O. Mañas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *ICCV*, 2021, pp. 9414–9423.

[40] U. Mall, B. Hariharan, and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *CVPR*, 2023, pp. 5261–5270.

[41] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, "GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization," in *NeurIPS*, vol. 36, 2023, pp. 8690–8701.

[42] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *NeurIPS*, vol. 35, 2022, pp. 197–211.

[43] Z. Dong, Y. Gu, and T. Liu, "Generative convnet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[44] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *ICCV*, 2023, pp. 16 806–16 816.

[45] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *CVPR*, 2022, pp. 9653–9663.

[46] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[47] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao, and L. Zhang, "MTP: Advancing remote sensing foundation model via multi-task pretraining," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–24, 2024.

[48] Z. Dong, Y. Gu, and T. Liu, "Upetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[49] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *ECCV*, 2022, pp. 280–296.

[50] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[51] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv e-prints*, p. arXiv:1607.06450, 2016.

[52] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *CVPR*, 2021, pp. 3517–3526.

[53] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *CVPR*, 2022, pp. 4794–4803.

[54] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[55] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3," *Purdue University Research Repository*, vol. 10, no. 7, p. 991, 2015.

[56] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf," *Remote Sensing of Environment*, vol. 250, p. 112012, 2020.

[57] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.

[58] W. Sun, K. Liu, G. Ren, W. Liu, G. Yang, X. Meng, and J. Peng, "A simple and effective spectral-spatial method for mapping large-scale coastal wetlands using china zy1-02d satellite hyperspectral images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 104, p. 102572, 2021.

[59] D. Wang, B. Du, and L. Zhang, "Fully contextual network for hyperspectral scene parsing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[60] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9201–9222, 2019.

[61] Y. Xu, B. Du, and L. Zhang, "Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperpsectral image classificaiton," *IEEE Transactions on Big Data*, pp. 1–1, 2019.

[62] J. Yang, B. Du, and L. Zhang, "Overcoming the barrier of incompleteness: A hyperspectral image classification full model," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[63] D. Wang, B. Du, and L. Zhang, "Spectral-spatial global graph reasoning for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[64] J. Yang, B. Du, and L. Zhang, "From center to surrounding: An interactive learning framework for hyperspectral image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 145–166, 2023.

[65] X. Li, M. Ding, Y. Gu, and A. Pižurica, "An end-to-end framework for joint denoising and classification of hyperspectral images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 7, pp. 3269–3283, 2023.

[66] C. Li, B. Rasti, X. Tang, P. Duan, J. Li, and Y. Peng, "Channel-layer-oriented lightweight spectral-spatial network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[67] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[68] A. Giannandrea, N. Raqueno, D. W. Messinger, J. Faulring, J. P. Kerekes, J. Van Aardt, K. Canham, S. Hagstrom, E. Ontiveros, A. Gerace *et al.*, "The share 2012 data campaign," in *Algorithms*

and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIX, vol. 8743, 2013, pp. 94–108.

[69] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1030–1043, 2013.

[70] D. Zhu, P. Zhong, B. Du, and L. Zhang, "Attention-based sparse and collaborative spectral abundance learning for hyperspectral subpixel target detection," *Neural Networks*, vol. 178, p. 106416, 2024.

[71] N. Billor, A. S. Hadi, and P. F. Velleman, "Bacon: blocked adaptive computationally efficient outlier nominators," *Computational statistics & data analysis*, vol. 34, no. 3, pp. 279–298, 2000.

[72] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: a unified framework for sparse dimension reduction," *Data Mining and Knowledge Discovery*, vol. 22, pp. 340–371, 2011.

[73] S. Sun, J. Liu, and S. Sun, "Hyperspectral subpixel target detection based on interaction subspace model," *Pattern Recognition*, vol. 139, p. 109464, 2023.

[74] H. Qin, W. Xie, Y. Li, and Q. Du, "Htd-vit: Spectral-spatial joint hyperspectral target detection with vision transformer," in *IGARSS*, 2022, pp. 1967–1970.

[75] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[76] C.-I. Chang, J. Liu, B. Chieu, H. Ren, C.-M. Wang, C. Lo, P.-C. Chung, C.-W. Yang, and D. Ma, "Generalized constrained energy minimization approach to subpixel target detection for multispectral imagery," *Optical Engineering*, vol. 39, no. 5, pp. 1275–1281, 2000.

[77] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 629–640, 2011.

[78] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Transactions on geoscience and remote sensing*, vol. 53, no. 3, pp. 1463–1474, 2014.

[79] Y. Zhang, B. Du, and L. Zhang, "A sparse representation-based binary hypothesis model for target detection in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1346–1354, 2014.

[80] G. Zhang, S. Zhao, W. Li, Q. Du, Q. Ran, and R. Tao, "HTD-Net: A deep convolutional neural network for target detection in hyperspectral imagery," *Remote Sensing*, vol. 12, no. 9, p. 1489, 2020.

[81] D. Shen, X. Ma, W. Kong, J. Liu, J. Wang, and H. Wang, "Hyperspectral target detection based on interpretable representation network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[82] D. Zhu, B. Du, M. Hu, Y. Dong, and L. Zhang, "Collaborative-guided spectral abundance learning with bilinear mixing model for hyperspectral subpixel target detection," *Neural Networks*, vol. 163, pp. 205–218, 2023.

[83] I. S. Reed and X. Yu, "Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution," *IEEE transactions on acoustics, speech, and signal processing*, vol. 38, no. 10, pp. 1760–1770, 1990.

[84] S. Li, K. Zhang, P. Duan, and X. Kang, "Hyperspectral anomaly detection with kernel isolation forest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 319–329, 2019.

[85] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Transactions on geoscience and remote sensing*, vol. 53, no. 3, pp. 1463–1474, 2014.

[86] T. Cheng and B. Wang, "Graph and total variation regularized low-rank representation for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 391–406, 2019.

[87] S. Wang, X. Wang, L. Zhang, and Y. Zhong, "Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[88] G. Fan, Y. Ma, X. Mei, F. Fan, J. Huang, and J. Ma, "Hyperspectral anomaly detection with robust graph autoencoders," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[89] J. Lian, L. Wang, H. Sun, and H. Huang, "GT-HAD: Gated transformer for hyperspectral anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024.

[90] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3867–3878, 2023.

[91] C. Han, C. Wu, M. Hu, J. Li, and H. Chen, "C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[92] Y. Wang, D. Hong, J. Sha, L. Gao, L. Liu, Y. Zhang, and X. Rong, "Spectral–spatial–temporal transformers for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[93] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *International Journal of Remote Sensing*, vol. 19, no. 3, pp. 411–426, 1998.

[94] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2858–2874, 2014.

[95] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-d cnn framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2019.

[96] J. Qu, S. Hou, W. Dong, Y. Li, and W. Xie, "A multilevel encoder–decoder attention network for change detection in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[97] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[98] M. Hu, C. Wu, and B. Du, "EMS-NET: Efficient multi-temporal self-attention for hyperspectral change detection," in *IGARSS*, 2023, pp. 6664–6667.

[99] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[100] M. Hu, C. Wu, and L. Zhang, "Globalmind: Global multi-head interactive self-attention network for hyperspectral change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 465–483, 2024.

[101] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon, "A bayesian model for joint unmixing and robust classification of hyperspectral images," in *ICASSP*, 2018, pp. 3399–3403.

[102] N. Keshava and J. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[103] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[104] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.

[105] D. C. Heinz *et al.*, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE transactions on geoscience and remote sensing*, vol. 39, no. 3, pp. 529–545, 2001.

[106] M. A. Veganzones, L. Drumetz, G. Tochon, M. Dalla Mura, A. Plaza, J. Bioucas-Dias, and J. Chanussot, "A new extended linear mixing model to address spectral variability," in *WHISPERS*, 2014, pp. 1–4.

[107] U. Kumar, C. Milesi, R. R. Nemani, S. Kumar Raja, S. Ganguly, and W. Wang, "Sparse unmixing via variable splitting and augmented lagrangian for vegetation and urban area classification using landsat data," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, pp. 59–65, 2015.

[108] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral–spatial hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 535–549, 2020.

[109] L. Gao, Z. Han, D. Hong, B. Zhang, and J. Chanussot, "CyCU-Net: Cycle-consistency unmixing network by learning cascaded autoencoders," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[110] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, and P. Scheunders, "Hyperspectral unmixing using transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[111] D. Hong, L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, and B. Zhang, "Endmember-guided unmixing network (egu-net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6518–6531, 2021.

[112] Q. Chen, Y. Wang, Z. Geng, Y. Wang, J. Yang, and Z. Lin, "Equilibrium image denoising with implicit differentiation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1868–1881, 2023.

[113] J. Ma, T. Cheng, G. Wang, Q. Zhang, X. Wang, and L. Zhang, "Prores: Exploring degradation-aware visual prompt for universal image restoration," *CoRR*, vol. abs/2306.13653, 2023.

[114] J. Ma, G. Wang, L. Zhang, and Q. Zhang, "Restoration and enhancement on low exposure raw images by joint demosaicing and denoising," *Neural Networks*, vol. 162, pp. 557–570, 2023.

[115] J. Ma, S. Yan, L. Zhang, G. Wang, and Q. Zhang, "Elmformer: Efficient raw image restoration with a locally multiplicative transformer," in *ACM MM*, 2022, pp. 5842–5852.

[116] A. A. Korosov, D. Demchev, N. Miranda, N. Franceschi, and J. Park, "Thermal denoising of cross-polarized sentinel-1 data in interferometric and extra wide swath modes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[117] X. Miao, Y. Zhang, and J. Zhang, "Thermal hyperspectral image denoising using total variation based on bidirectional estimation and brightness temperature smoothing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[118] Y. Chang, L. Yan, and S. Zhong, "Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising," in *CVPR*, 2017, pp. 5901–5909.

[119] W. He, Q. Yao, C. Li, N. Yokoya, and Q. Zhao, "Non-local meets global: An integrated paradigm for hyperspectral denoising," in *CVPR*, 2019, pp. 6868–6877.

[120] F. Xiong, J. Zhou, and Y. Qian, "Hyperspectral restoration via l_0 gradient regularized low-rank tensor factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 10 410–10 425, 2019.

[121] J. Peng, Q. Xie, Q. Zhao, Y. Wang, L. Yee, and D. Meng, "Enhanced 3dtv regularization and its applications on hsi denoising and compressed sensing," *IEEE Transactions on Image Processing*, vol. 29, pp. 7889–7903, 2020.

[122] Y.-C. Miao, X.-L. Zhao, X. Fu, J.-L. Wang, and Y.-B. Zheng, "Hyperspectral denoising using unsupervised disentangled spatiospectral deep priors." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[123] K. Wei, Y. Fu, and H. Huang, "3-d quasi-recurrent neural network for hyperspectral image denoising," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 363–375, 2021.

[124] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *AAAI*, 2023, pp. 1368–1376.

[125] Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *BigMM*, 2018, pp. 1–4.

[126] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *CVPR*, 2021, pp. 1833–1844.

[127] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[128] X. Wang, J. Ma, and J. Jiang, "Hyperspectral image super-resolution via recurrent feedback embedding and spatial–spectral consistency regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[129] X. Wang, Q. Hu, J. Jiang, and J. Ma, "A group-based embedding learning and integration network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[130] S. Chen, L. Zhang, and L. Zhang, "MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[131] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2020.

[132] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel resnet for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7831–7843, 2021.

[133] C. Yi, L. Zhang, X. Zhang, W. Yueming, Q. Wenchao, T. Senlin, and P. Zhang, "Aerial hyperspectral remote sensing classification dataset of xiongan new area (matiwan village)," *National Remote Sensing Bulletin*, vol. 24, no. 11, pp. 1299–1306, 2020.

[134] J. Yang, B. Du, D. Wang, and L. Zhang, "ITER: Image-to-pixel representation for weakly supervised hsi classification," *IEEE Transactions on Image Processing*, vol. 33, pp. 257–272, 2024.

[135] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8671–8685, 2021.

[136] C. Shi, Y. Dang, L. Fang, Z. Lv, and M. Zhao, "Hyperspectral image classification with adversarial attack," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[137] B. Tu, W. He, Q. Li, Y. Peng, and A. Plaza, "A new context-aware framework for defending against adversarial attacks in hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[138] Y. Xu, Y. Xu, H. Jiao, Z. Gao, and L. Zhang, "S^{3}ANet: Spatial–spectral self-attention learning network for defending against adversarial attacks in hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[139] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *ICLR*, 2017.

[140] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[141] G. K. Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[142] T. Liu, Y. Gu, and X. Jia, "Class-guided coupled dictionary learning for multispectral-hyperspectral remote sensing image collaborative classification," *Science China Technological Sciences*, vol. 65, no. 4, pp. 744–758, 2022.

[143] Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *CVPR*, 2016, pp. 5081–5090.

[144] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "Cospace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4349–4359, 2019.

[145] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1470–1474, 2020.

[146] B. Guo, T. Liu, and Y. Gu, "Integrating coupled dictionary learning and distance preserved probability distribution adaptation for multispectral–hyperspectral image collaborative classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[147] ——, "Structure preserved discriminative distribution adaptation for multi-hyperspectral image collaborative classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[148] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS*, 2018, pp. 2115–2118.

[149] H. Chen, J. Song, C. Wu, B. Du, and N. Yokoya, "Exchange means change: An unsupervised single-temporal change detection framework based on intra- and inter-image patch exchange," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 206, pp. 87–105, 2023.

[150] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*, 2018, pp. 4063–4067.

[151] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change Detection in Multisource VHR Images via Deep Siamese Convolutional Multiple-Layers Recurrent Neural Network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2020.

[152] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 3056416, 2022.
[153] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.
[154] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS*, 2022, pp. 207–210.
[155] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "Changemamba: Remote sensing change detection with spatio-temporal state space model," *arXiv preprint arXiv:2404.03425*, 2024.
[156] P. Duan, X. Kang, P. Ghamisi, and S. Li, "Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.
[157] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *ICCV*, 2023, pp. 4088–4099.
[158] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired foundation model for observing the earth crossing modalities," *arXiv preprint arXiv:2403.15356*, 2024.
[159] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "VSA: Learning varied-size window attention in vision transformers," in *ECCV*, 2022, pp. 466–483.

Di Wang (Member, IEEE) received the B.E. degree in surveying and mapping from China University of Petroleum (East China), Qingdao, China, in 2018, the M.E. degree from State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2020, where He is currently pursuing the Ph.D. degree at the School of Computer Science. His research interests include deep learning, computer vision and remote sensing.

Meiqi Hu received the B.S. degree in surveying and mapping engineering from the School of Geoscience and info-physics, Central South University, Changsha, China, in 2019, and received the Ph.D. degree in Photogrammetry and Remote Sensing from State Key Lab of Information Engineering in Surveying, Mapping and Remote sensing, Wuhan University, Wuhan, China, in 2024. Her research interests include deep learning, multitemporal remote sensing image change detection and spectral unmixing.

Yao Jin received the B.E. degree in geographic informaion science from Huazhong Agricultural University, Wuhan, China, in 2018. the M.E. degree from Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan, China, in 2023, where He is currently pursuing the Ph.D. degree at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China. His research interests include hyperspectral images information extraction.

Yuchun Miao received the B.E. degree in Mathematics and Applied Mathematics from the University of Electronic Science and Technology of China, Chengdu, China. He is currently pursuing the Ph.D. degree in the School of Computer Science, Wuhan University, China. His research interests include tensor modeling for high dimensional image and reinforcement learning for LLM.

Jiaqi Yang received Ph.D degree from Wuhan University, Wuhan, China, 2024. Her interests include deep learning, image interpretation, and HSI classification.

Yichu Xu (Student Member, IEEE) received the B.S. degree in geographic information system from Lanzhou University, Lanzhou, China, in 2018. He is currently pursuing the Ph.D. degree at the School of Computer Science, Wuhan University, Wuhan, China. His research interests include computer vision, remote sensing and hyperspectral image processing.

Xiaolei Qin received the B.S. degree in geographic information science from the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China, in 2021. She is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. Her research interests include multispectral images processing and intepretation.

Jiaqi Ma (Student Member, IEEE) received the B.S. degree in School of Information Engineering, Zhengzhou University, China, in 2018. He received the M.S. degree in the School of Computer Science, Wuhan University, China, in 2021. He was also under joint supervision at the National University of Singapore from 2023 to 2024. He is currently pursuing the Ph.D. degree in the School of Computer Science, Wuhan University, China. His research interests include low-level vision and computational imaging.

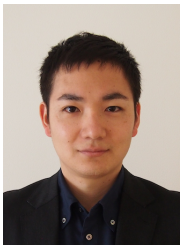
Lingyu Sun (Student Member, IEEE), received the B.S. degree in School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. She is currently pursuing the Ph.D. degree in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. Her research interests include time serie mapping and analysis.

Chenxing Li received the B.S. degree in China University of Geosciences, Wuhan, China, in 2023. He is currently pursuing the master's degree at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. His research interests include machine learning and hyperspectral target detection.

Chuan Fu received the B.S. degree in Printing engineering from the School of Printing and Packing at Wuhan University, Wuhan, China, in 2016, the Ph.D. degree in surveying and mapping engineering from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2023. He is currently a postdoctoral researcher with the Collge of Computer Science, Chongqing University, Chongqing, China. His research interests include deep learning, remote sensing, and image processing.

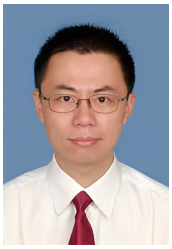
Hongruixuan Chen (Graduate Student Member, IEEE) received the B.E. degree from Anhui University, Hefei, China, in 2019, and the M.E. degree from Wuhan University, Wuhan, China, in 2022. He is currently pursuing the Ph.D. degree with the Graduate School of Frontier Science, The University of Tokyo, Chiba, Japan. He is now also an academic visitor in ETH Zurich, Zurich, Switzerland. His current research is motivated by how to better monitor, describe and understand changes in our planet's surface by studying machine learning and computer vision approaches, thereby contributing to urban planning, resource management, environmental protection, and sustainable development. He was a Trainee at the United Nations Satellite Centre (UNOSAT), United Nations Institute for Training and Research (UNITAR), Geneva, Switzerland. He also acts as a reviewer for 16 international journals, e.g., IEEE TIP, IEEE TGRS, and ISPRS J P&RS.

Chengxi Han (Student Member, IEEE) received the B.S. degree in remote sensing science and technology from Central South University, China, in 2018. He got the Ph.D. degree in photogrammetry and remote sensing from LIESMARS, Wuhan University, China, in 2024. Dr. Han has been a Student Chapter Coordinator in the IEEE GRSS AdCom since January 2024. His research interests include remote sensing image information extraction. His personal website is https://chengxihan.github.io/.



Naoto Yokoya (Member, IEEE) is currently an Associate Professor at the University of Tokyo and a Team Leader at the RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he leads the Geoinformatics Team. He was an Assistant Professor at the University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow, working at the German Aerospace Center (DLR), Oberpfaffenhofen, and Technical University of Munich (TUM), Munich, Germany. His research focuses on visual information processing of large-scale real-world scenes. Dr. Yokoya won first place in the 2017 IEEE GRSS Data Fusion Contest organized by the IADF TC. He was the Chair from 2019 to 2021, a Co-Chair of the IEEE GRSS IADF TC from 2017 to 2019, and also the Secretary of the IEEE GRSS All Japan Joint Chapter from 2018 to 2021. He was an Associate Editor for the IEEE JSTARS from 2018 to 2021. He has been an Associate Editor for the IEEE TGRS since 2021, as well as the J P&RS since 2024. He has been designated a Clarivate Highly Cited Researcher since 2022.

Jing Zhang (Senior Member, IEEE) is currently a professor at the School of Computer Science, Wuhan University, Wuhan, China. He previously served as a Research Fellow at the School of Computer Science, The University of Sydney. He has published over 100 papers in leading venues such as CVPR, ICCV, NeurIPS, IEEE TPAMI, and IJCV, with research focused on computer vision and deep learning. He is an Area Chair for ICPR, a Senior Program Committee member for AAAI and IJCAI, and a guest editor for IEEE TBD, while also regularly reviewing for top-tier journals and conferences.



Minqiang Xu is currently the Principal Scientist of iFLYTEK Digital. He received the Ph.D. degree in circuits and systems from University of Science and Technology of China in 2011. During 2018-2010, he studied at the Image Formation and Processing (IFP) Laboratory at the University of Illinois at Urbana-Champaign (UIUC), where he was supervised under Prof. Thomas S. Huang, a pioneer scientist in the field of computer vision. He has won the first place multiple times in international voiceprint recognition competitions including NIST SRE and VoxSRC. He participated in the development of the champion system (NEC-UIUC) for the ImageNet 2010 competition.

Lin Liu is currently a researcher at the National Engineering Research Center for Speech and Language Information Processing. He received the engineering doctorate degree in electronic information from the University of Science and Technology of China. He has long been engaged in scientific research and industrialization in artificial intelligence technologies such as speech recognition, remote sensing image recognition, and machine translation.

Lefei Zhang (Senior Member, IEEE) is currently a Professor with the School of Computer Science, Wuhan University, and also with the Hubei Luojia Laboratory, Wuhan. He was a Big Data Institute Visitor with the Department of Statistical Science, University College London, London, U.K., and a Hong Kong Scholar with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. His research interests include pattern recognition, image processing, and remote sensing. Dr. Zhang serves as a Topical Associate Editor for IEEE TGRS and an Associate Editor for IEEE GRSL.

Chen Wu (Member, IEEE) is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.

Bo Du (Senior Member, IEEE) is currently a Professor with the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University, where he is also the Director with the National Engineering Research Center for Multimedia Software. He has more than 80 research articles published in the IEEE TPAMI, TIP, TCYB, TGRS, JSTARS, and GRSL, and 13 of them are ESI hot papers or highly cited papers. His main research interests include pattern recognition, hyperspectral image processing, and signal processing. He serves as a Senior PC Member for IJCAI and AAAI and the Area Chair for ICPR. He received the Clarivate Highly Cited Researcher (2019–2022), the 2018 IJCAI Distinguished Paper Prize, the 2018 IEEE Data Fusion Contest Champion, the 2018 IEEE WHISPERS Best Paper Award, and the First Place in NLP Competitions GLUE and SuperGLUE in 2022. He serves as an Associate Editor for *Neural Networks*, *Pattern Recognition*, and *Neurocomputing*.

Dacheng Tao (Fellow, IEEE) is currently a Distinguished University Professor in the College of Computing & Data Science at Nanyang Technological University. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences, with best paper awards, best student paper awards, and test-of-time awards. His publications have been cited over 112K times and he has an h-index 160+ in Google Scholar. He received the 2015 and 2020 Australian Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a Fellow of the Australian Academy of Science, AAAS, ACM and IEEE.

Liangpei Zhang (Fellow, IEEE) is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He was a Principal Scientist of the China State Key Basic Research Project appointed by the Ministry of National Science and Technology of China to lead the remote sensing program from 2011 to 2016. He has published more than 700 research papers and five books. He is the Institute for Scientific Information (ISI) highly cited author. He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence. He is a fellow of the Institution of Engineering and Technology (IET). He is currently serving as an Associate Editor for the IEEE TGRS.