International Journal of Applied Earth Observations and Geoinformation 113 (2022) 103022

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag

WHU-OHS: A benchmark dataset for large-scale Hersepctral Image classification

Jiayi Li [a], Xin Huang [a,b,*], Lilin Tu [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China

A R T I C L E I N F O

Keywords:
Hyperspectral image classification
Deep learning
Sematic segmentation
Transfer learning

A B S T R A C T

Hyperspectral image (HSI) classification is one of the most important remote sensing techniques. Currently, the performances of most of the HSI classification networks on the public HSI datasets are overoptimistic (i.e., the overall accuracy exceeds 98 %). This deficiency is partly due to the very limited scale of these existing datasets, which impedes the network generalization and poses limitations for the future development. The existing hyperspectral datasets urgently need to be scaled up in size. Therefore, in this study, we built a dataset named the WHU-OHS dataset, which consists of about 90 million manually labeled samples of 7795 Orbita hyperspectral satellite (OHS) image patches (sized 512 × 512) from 40 Chinese locations. This dataset ranges from the visible to near-infrared range, with an average spectral resolution of 15 nm. The extensive geographical distribution, large spatial coverage, and widely used classification system make the WHU-OHS dataset a challenging benchmark. This dataset was validated by comprehensive experiments using several representative deep HSI classification networks. Furthermore, the transferability of the HSI classification networks under the conditions of the same/different HSI sensors was tested. In particular, when classifying the existing public HSI datasets, using initial parameters obtained by pre-training on the WHU-OHS dataset can further improve the inference accuracy as well as the training efficiency. The WHU-OHS dataset and a PyTorch toolbox for large-scale HSI classification are available at https://irsip.whu.edu.cn/resources/resources_v2.php and https://github.com/zjjerica/WHU-OHS-Pytorch, respectively.

1. Introduction

The ability to understand the Earth's surface has been significantly promoted by the availability of vast amounts of remote sensing data and advanced remote sensing technologies (Li et al., 2019a, 2019b). Among the various remote sensing data sources, hyperspectral image (HSI) data have the unique advantages of a high spectral resolution, continuous frequency, and wide range of electromagnetic spectrum (Wambugu et al., 2021). As a result, HSI classification is recognized as being one of the most important remote sensing techniques (Li et al., 2014; Datta et al., 2022; Jaiswal et al., 2021).

To investigate the current situation, we analyzed the journal articles published from Jan. 2016 to Jan. 2021 in the remote sensing community. Specifically, three representative journals (i.e., *ISPRS Journal of Photogrammetry and Remote Sensing*, *Remote Sensing of Environment*, and *IEEE Transactions on Geoscience and Remote Sensing*) were selected. Please note that we did not review all the remote sensing journals, as our aim was not to count the total number of hyperspectral papers, but to reflect the overall trends. Furthermore, we performed a *meta*-analysis of the classification algorithms used in the afore-mentioned journal articles (see Supplementary Material I). The analysis of all the HSI classification related articles in the three journals revealed the following characteristics (Table 1):

We found 312 journal articles on this topic over the last five years, indicating that HSI classification is a hot research topic in the remote sensing community. In total, 277 of these articles utilized the existing publicly labeled datasets (the first eight datasets in Table 2) to validate the effectiveness of the designed classification methods. In Table 2, from the visible to near-infrared spectrum (or at least the short-wave infrared region), each dataset has tens to hundreds of continuous bands with a dozen or so categories, and covers a localized landscape with a *small spatial coverage* (i.e., <350 km^2). All the datasets except for the Botswana dataset are made up of *aerial photographs*. As the flight plans of airborne equipment are carried out under clear sky conditions with good

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China.
E-mail address: xhuang@whu.edu.cn (X. Huang).

https://doi.org/10.1016/j.jag.2022.103022
Received 15 June 2022; Received in revised form 9 September 2022; Accepted 10 September 2022
Available online 18 September 2022

visibility and no clouds, the imaging conditions of airborne images are typically more stable than those of satellite images. Thus, when the training and test data are drawn from the same small HSI (Li et al., 2015b), most of the current methods (especially those that are more prone to spatial smoothing) are overoptimistic (i.e., the overall accuracy (OA) exceeds 98 %) (Xu et al., 2022).

Since pixels within a homogeneous parcel violate the sample independence assumption, the actual sample size cannot be regarded as the number of pixels. Although several datasets have millions of labeled pixels, when taking parcels as the basic unit, *the sample size of each dataset is quite limited* (Table 2). Furthermore, for most of the current public datasets, the reference samples are sparse, occupying only a small proportion of the whole image. The sparse (see the snapshots of each public dataset presented in Fig. 1) and limited reference samples are divided into a test set and a training set, leading to the following two shortcomings. Firstly, an insufficient test set can bias the accuracy evaluation and restrict the reliability of further applications. Thus, there is an immediate need to *build a dense and large HSI benchmark dataset*. Secondly, the small number of training samples is a well-known challenge for HSI classification (Li et al., 2015a; Gao et al., 2022), as the training models can be ill-posed. To deal with this problem, some recent machine learning paradigms, such as active learning (Xue et al., 2021) and semi-supervised learning (Hu et al., 2018; Li et al., 2020), assume that the same type of surfaces follow the same distribution in the feature space, and can obtain sematic information from the unlabeled samples.

While these efforts can relieve the small number of training samples issue, for the classification of an HSI image (i.e., the "target domain task"), the large amount of existing semantic information (i.e., the "knowledge" from the source domain) can be helpful. This is the core idea of transfer learning, which is one of the hot topics in the HSI classification field (Rao et al., 2022). As noted from the 312 surveyed articles, except for a few studies jointly using training samples from two similar datasets to investigate the role of transfer learning (Liang et al., 2021), most of the current studies still adopt simulated data. In this approach, several disjoint sub-images are cut from a single image, as source and target images, respectively. Thus, the superiority of transfer learning, which involves inheriting semantic information from external HSI data, has not been fully exploited. The main dilemma of transfer learning is how to overcome the large differences in the data modalities, sensors, landscapes, and categories among the multiple HSIs, which is a problem more often encountered in practical interpretation tasks.

We identified 111 deep learning related articles, representing *the deep learning era for HSI classification* (Audebert et al., 2019; Wambugu et al., 2021). The current research shows that the model complexity of the HSI classification networks is much lower than that of the very high spatial resolution remote sensing (VHR) image networks (Hu et al., 2021). To date, most of the complicated but well-performing deep models have been trained by the use of large numbers of samples. Under the condition of there being no publicly accessible large-scale benchmark datasets for HSI classification, there are three main techniques used to compensate for the inadequacy of the labeled samples.

(1) The most intuitive technique is data augmentation (Zhu et al., 2018), which involves synthesizing "pseudo" samples from a small number of real samples by, for instance, rotating the real samples (Zhang et al., 2021). However, it is generally believed that the quality of synthetic samples is far from that of real samples, and in HSI processing, the performance is even worse, due to the low signal-to-noise ratio of HSI images (Li et al., 2019a, 2019b).
(2) The second popular technique is to use a small number of target domain samples to fine-tune the existing deep models pre-trained from other fields with large numbers of samples (Windrim et al., 2018). However, the big gap between HSI datasets and datasets from other domains (e.g., natural (RGB) imagery, text data, and VHR imagery) poses a challenge to the transferability.
(3) The recent progress in self-supervised learning (Chen and Bruzzone, 2022), which involves directly learning the essential characteristics of data from unlabeled data, has been a milestone in many fields such as sematic segmentation (Li et al., 2021), change detection (Chen and Bruzzone, 2022), and scene classification (Zhao et al., 2020) through the use of VHR imagery. Thus, although the fine spectral bands of HSI data can precisely delineate the different land-cover types, the shortage of open-access large-scale HSI datasets, as well as validation samples, can hinder the use of HSI data with these new machine learning techniques (see Table 2). In addition, the size of the public hyperspectral datasets is much smaller than that of the VHR datasets, which severely restricts the development of HSI interpretation.

There is therefore an urgent need to construct a large-scale hyperspectral dataset, so as to further explore deep learning techniques in this area. The contributions of this paper can be summarized as follows:

(1) We present a large-scale land-use/land-cover (LULC) HSI classification dataset, namely, the WHU-OHS dataset, covering >150,000 km^2 in China, which is made up of Orbita hyperspectral satellite (OHS) images. To the best of our knowledge, this is the largest open-source hyperspectral remote sensing dataset.
(2) On the basis of the WHU-OHS dataset, we conducted systematic comparisons with several representative deep learning based HSI classification methods, with/without consideration of multi-scene heterogeneity.
(3) We investigate the potential of transferring from the WHU-OHS dataset (by pre-training) to the other current hyperspectral datasets (Table 2).
(4) A PyTorch toolbox (https://github.com/zjjerica/WHU-OHS-Pytorch) for large-scale HSI classification is introduced.

The rest of this paper is organized as follows. Section 2 describes the construction of the WHU-OHS dataset, which consists of 42 OHS satellite images and covers an area of >150,000 km^2 in China. Section 3 summarizes the current status of the deep learning based HSI classification methods. In Sections 4 and 5, the series of comprehensive experiments conducted on the WHU-OHS dataset are described, with/without consideration of multi-scene heterogeneity. In Section 6, we discuss the feasibility of transferring from the WHU-OHS dataset to the other existing and widely used hyperspectral datasets. In Section 7, we discuss the potential use of the WHU-OHS dataset in future work. Finally, Section 8 concludes the paper.

Table 1
Numbers of surveyed articles and their specific characteristics.

Total number	Transfer learning	Deep learning	Convolutional neural network (CNN)	Recurrent neural network (RNN)	Generative adversarial network (GAN)
312	18	111	96	10	12
Public data	Spectral-spatial	Auto-encoder	3D- CNN	Sematic segmentation	Capsule net
277	246	27	28	14	3

2. WHU-OHS dataset

2.1. Orbita hyperspectral micro-nano satellite images and pre-processing

According to the description in the user manual, the hyperspectral constellation composed of 10 Orbita hyperspectral micro-nano satellites was launched and is managed by Zhuhai Orbita Aerospace, China. The sun-synchronous orbits of each satellite are about 500 km in altitude, with an inclination of around 98°. The imagery has a spatial resolution of 10 m (nadir) and a swath width of 60 km (nadir). With the 10 satellites, the hyperspectral constellation can realize global observation within 2 days. With a spectral range of 400–1000 nm, there are 256 spectral channels in total, with an average spectral resolution of 2.5 nm. The OHS constellation can acquire Earth observation data with a high spatial-spectral-temporal resolution and a large spatial coverage. As a result, OHS data have been applied in LULC mapping, crop estimation and monitoring, and other applications.

For each image, the pre-processing, including relative radiometric calibration, atmospheric correction, and geometric registration, was conducted in the ENVI 5.3 environment. By using the Radiometric Calibration toolbox in ENVI 5.3, the relative radiometric calibration is aimed at eliminating the sensor error and transferring the digital number value of the original signal to an apparent radiance value. Based on the MODTRAN4 radiative transfer model (Berk et al., 1999), atmospheric correction was implemented using the Fast Line-of-sight Atmospheric Analysis of Hypercubes (FLAASH) tool in ENVI 5.3. In this model, GMTED2010 data (Danielson and Gesch, 2011) are used to estimate the average elevation. The atmospheric parameters and the aerosols in the MODTRAN4 model are mainly determined by longitude and latitude, imaging time, and land-cover types. Thus, for the proposed dataset, the atmospheric models included subarctic summer, mid-latitude summer, and tropical models, and the aerosols included urban aerosols and rural background aerosols. In addition, the spectral band with a central wavelength of 820 nm was also used for water vapor inversion. The effectiveness of the atmospheric correction step is discussed in Supplementary Material II. Next, using the rational polynomial coefficient (RPC) Orthorectification toolbox in ENVI 5.3, orthophoto correction without control points was carried out based on the GMTED2010 global digital elevation model (DEM) dataset (Danielson and Gesch, 2011). These data-processing procedures were conducted by Zhuhai Orbita Aerospace, China.

Considering the spectral details and data volume, Zhuhai Orbita Aerospace provides imagery with 32 spectral bands selected from the 256 channels. The central wavelength of each band is listed in Table 3. In view of the number of spectral bands, spectral bandwidths, and spectral range, the proposed WHU-OHS dataset basically meets the definition of hyperspectral imagery (Rinker, 1990; Manolakis et al., 2016).

2.2. Land-use/land-cover types

A hierarchical classification system with seven major classes and 24 classes was adopted for the proposed WHU-OHS dataset, based on the Chinese Land Use Classification Criteria (GB/T21010-2017), as shown in Fig. 2. The descriptions of the classes are provided in Table 4. The intra-class spectral diversity was calculated on the basis of the within-class scatter matrix (Altman et al., 1994). As shown in Eq. (1), for the ith class, given N_i samples, the scatter matrix $S_w^{(i)}$ can be denoted as:

$$S_w^{(i)} = \frac{1}{N}\sum_{k=1}^{N_i}\left(X_k^{(i)} - m^{(i)}\right)\left(X_k^{(i)} - m^{(i)}\right)^T \quad (1)$$

where $X_k^{(i)}$ and $\boldsymbol{m}^{(i)}$ refer to the spectral feature of the kth sample and the average spectral feature of the ith class, respectively. For the 42 OHS images, min-max scaling spectral feature normalization was carried out as follows:

$$X_k^{(i)} = X_{k,0}^{(i)} - \min_i/\max_i - \min_i \quad (2)$$

where $\min_i$ and $\max_i$ refer to the minimum and the maximum spectral reflectance values for all the ith class samples, respectively.$X_{k,0}^{(i)}$ refers to the original spectral features. The trace of $S_w^{(i)}$ then measures the spectral diversity of these N_i samples. Table 4 presents the intra-class diversity (ICD) for each class. According to this table, the spectral diversity of all the classes across the 42 OHS images is high.

2.3. Dataset construction for deep learning

The WHU-OHS dataset is made up of 42 OHS satellite images

Table 2
Public and newly released datasets for HSI classification.

Dataset	Platform and sensor	Spectral		Spatial		Reference labels		
		Bands (available bands)	Range (μm)	Size (pixels)	Resolution (meters)	Classes	Sample size	
							Pixels	Parcels**
Indian Pines =	Aerial: AVIRIS	224 (200)	0.4–2.5	145 × 145	20.0	16	10,249	40
Salinas Valley	Aerial: AVIRIS	224 (204)	0.4–2.5	512 × 217	3.7	16	54,129	24
KSC	Aerial: AVIRIS	224 (176)	0.4–2.5	512 × 614	18.0	13	4756	59
Washington DC Mall	Aerial: HYDICE	210 (191)	0.4–2.4	1208 × 307	1.5 ~ 3.0	7	26,332	399
DFC 2013	Aerial: CASI-1500	144 (144)	0.38–1.05	1095 × 349	2.5	15	15,029	547
DFC 2018	Aerial: CASI-1500	50 (48)	0.38–1.05	4786 × 1202	1.0	20	573 k	2255
Pavia U	Aerial: ROSIS	115 (103)	0.43–0.85	610 × 340	1.3	9	42,776	265
Pavia C	Aerial: ROSIS	115 (102)	0.43–0.85	1096 × 715	1.3	9	148,152	663
AeroRIT (Rangnekar et al., 2020)	Aerial: Headwall Micro E	372	0.397–1.003	1973 × 3975	0.4	5	6,306,697	2534
WHU-Hi-HanChuan	UAV aerial: Headwall Nano-HS imaging	270	0.40–1.00	1217 × 303	0.109	16	368,751	155
WHU-Hi-HongHu	UAV aerial: Headwall Nano-HS imaging	270	0.40–1.00	940 × 475	0.043	22	446,500	44
WHU-Hi-LongKou	UAV aerial: Headwall Nano-HS imaging	270	0.40–1.00	550 × 400	0.463	9	220,000	22
Matiwan Village (Cen et al., 2020)	UAV aerial: V-NIR imaging spectrometer	250	0.40–1.00	3750 × 1580	0.5	20	5,925,000	135
Botswana	Satellite: Hyperion	242 (145)	0.4–2.5	256 × 1476	30.0	14	3248	86
WHU-OHS	Satellite: Orbita hyperspectral micro-nano satellites	32	0.40–1.00	7795 × 512 × 512	10	24	90 million	117,286

* A parcel represents a collection of identically labeled, spatially contiguous pixels.

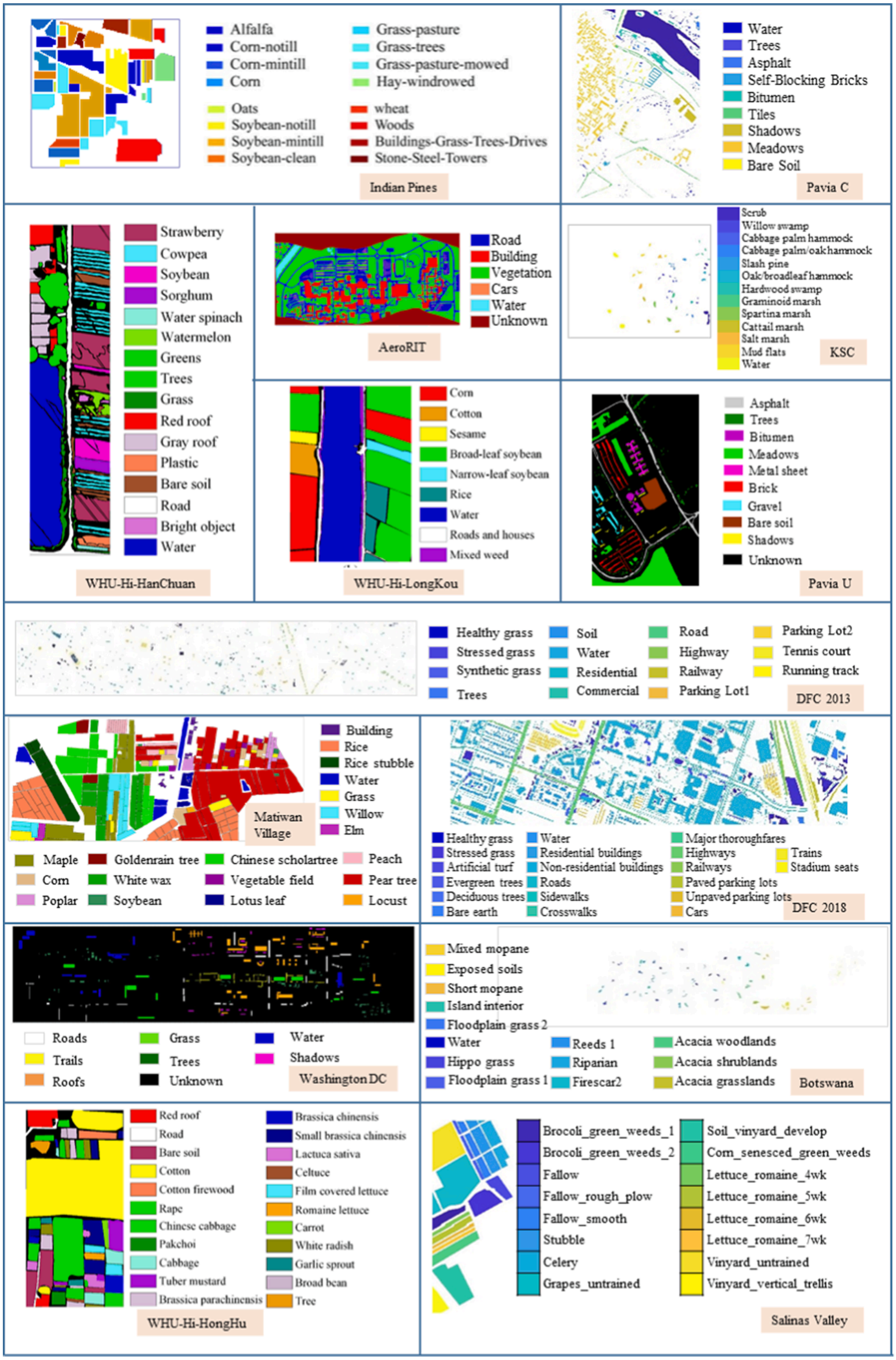


Fig. 1. Snapshots of each public dataset for HSI classification.

Table 3
The central wavelength of each spectral band of the OHS imagery.

No. of bands	CWB (nm)	No. of bands	CWB (nm)	No. of bands	CWB (nm)	No. of bands	CWB (nm)
1	466	9	569	17	716	25	836
2	480	10	610	18	730	26	850
3	500	11	626	19	746	27	866
4	520	12	640	20	760	28	880
5	536	13	656	21	776	29	896
6	550	14	670	22	790	30	910
7	566	15	686	23	806	31	926
8	580	16	700	24	820	32	940

CWB = central wavelength of each band.

acquired from >40 different locations in China (Fig. 2). Each OHS image is named after its location, and is listed and abbreviated as described in the readme file of the dataset (e.g., "Jinzhong" in Shanxi province is abbreviated as O1). The extensive geographical distribution, the large spatial coverage, and the widely used classification system make the WHU-OHS dataset a challenging benchmark database for HSI classification.

Each OHS image with around 2.3 million manually labeled pixels has been seamlessly cropped into 16 (i.e., 4 × 4) tiles with equal size, where 12 and 4 tiles are employed as the training and test samples, respectively. We cropped each tile into 512 × 512 pixels with a stride of 32. All the sub-images in the test tiles were used for the inference, and the sub-images in the training tiles were randomly divided into training and validation sets, with a ratio of 9:1.

The dataset was organized in the format shown in Fig. 3. The train (tr), validation (val), and test (ts) folders and a readme text file are in the first-level subdirectory, where each folder contains two subfolders, called image and label, respectively. Each file in the third-level subdirectory utilizes the abbreviation of the OHS imagery and the number of the unordered sub-images. For instance, T1_0002.img means the second sub-image of the first target OHS image, and T1_0002.jpg records the associated pixel-wise label. For the transferability testing, we chose eight pairs of OHS images, in which each pair contained one source image and one target image (denoted as S1 and T1), which were spatially adjacent and had similar land-cover landscapes. The readme file records the name and abbreviation of every OHS image (e.g., S1 is short for Changchun). This dataset will be made publicly available for all research needs. There are 4822, 513, and 2460 sub-images in the training, validation, and test sets, respectively. Several sub-image examples are presented in Fig. 4.

3. Representative deep learning methods for HSI classification

Referring to the recent reviews (Audebert et al., 2019; Li et al., 2019b; Wambugu et al., 2021), the existing deep learning methods for HSI classification can be categorized as spectral, spatial, and spectral-spatial methods, in terms of the feature types. As shown in Table 1, the 96 CNN-related articles indicate that convolutional filters (abbreviated as conv) are the layers that are most often used to model the spectral and spectral-spatial features. Generally speaking, there are two approaches to exploiting spectral-spatial information by 2D

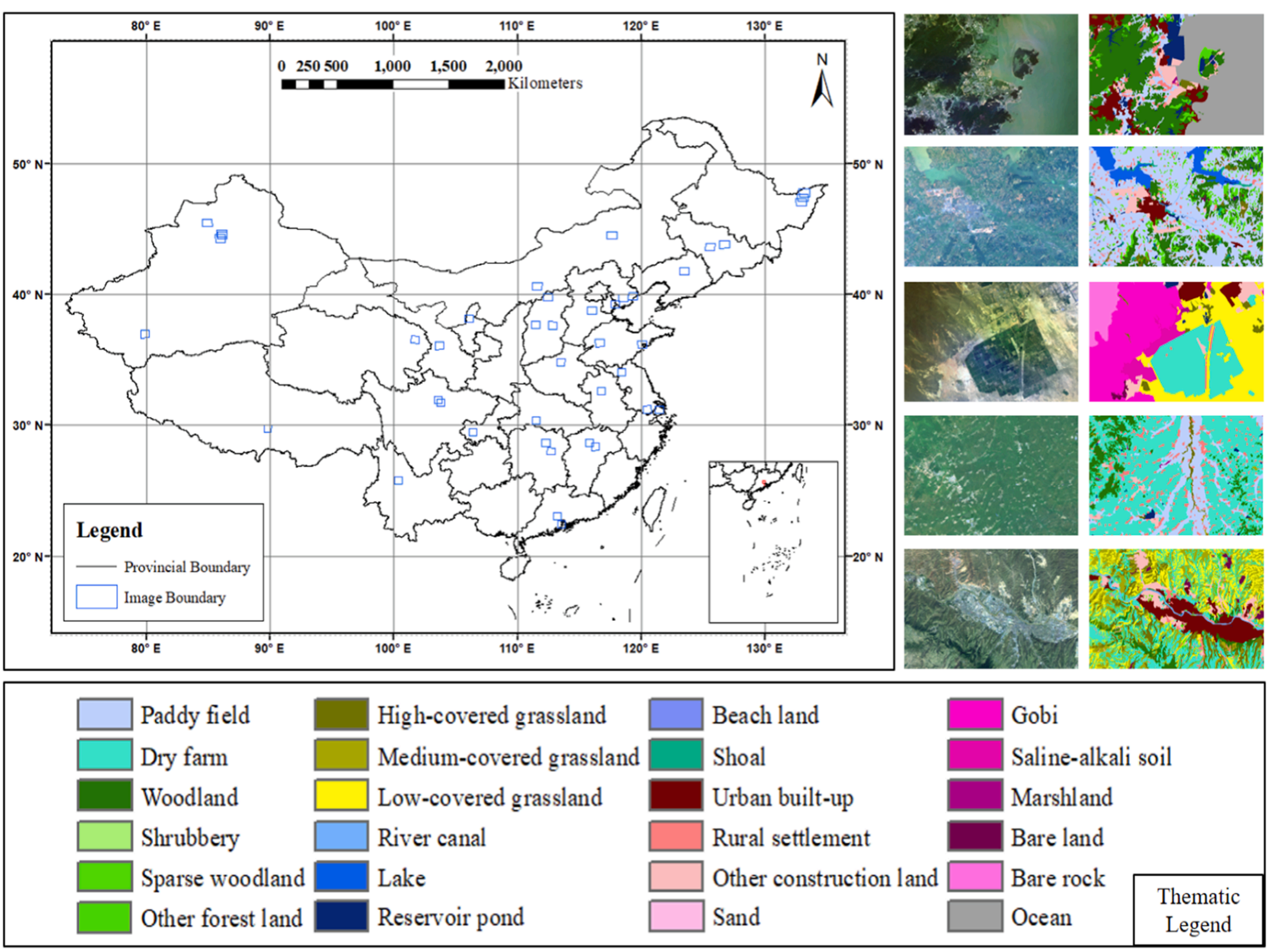


Fig. 2. Left: The geographical locations of the 42 images in the WHU-OHS dataset. Right: Examples of local OHS parcels (true-color compositions with R: 670 nm; G: 566 nm; B: 480 nm) and their corresponding reference labels.

Table 4
The hierarchical classification system and the intra-class diversity of the WHU-OHS dataset.

Major class	Class	Description	ICD
Farmland		Land where crops are grown, including beaches and shore land that has been cultivated for more than three years.	
	Paddy field	Arable land with guaranteed water source and irrigation facilities, including arable land where rice and dryland crops are rotated.	0.6182
	Dry farm	Cultivated land without man-made water sources, including irrigated dry planting, vegetable fields, and fallow fields.	0.4555
Forest		Forestry land, including growing trees, shrubs, trees, bamboo, and coastal mangrove land.	
	Woodland	Forest land with crown density > 30 %.	0.4452
	Shrubbery	Coppice land and shrubland with crown density > 30 % and canopy < 2 m.	0.5865
	Sparse woodland	Forest land with crown density of 10–30 %.	0.4819
	Other forest land	Slash, nurseries, and garden plots, such as orchards, mulberry gardens, and tea plantations.	0.3350
Grassland		Dominated by herbaceous annuals.	
	High-covered grassland	Herbaceous area with sufficient soil moisture. Coverage > 50 %.	0.3706
	Medium- covered grassland	Herbaceous area with insufficient moisture. Coverage: 20–50 %.	0.4844
	Low-covered grassland	Sparse natural herbaceous area with insufficient moisture. Coverage: 5–20 %.	0.5779
Water body		Natural inland water bodies.	
	River/canal	Natural rivers or artificially excavated canals (including embankments).	0.5453
	Lake	Natural waterlogged areas below the perennial water level.	0.9947
	Reservoir/pond	Land below the perennial water level in artificially constructed water storage areas.	0.5042
	Beach land	The tidal immersion zone between the high tide level and the low tide level of the coastal spring tide.	1.0326
	Shoal	natural submerged ridge, bank, or bar that consists of, or is covered by, sand or other unconsolidated material and rises from the bed of a body of water to near the surface	0.4102
Built-up land		Human settlement land, industrial and mining land, and transportation land.	
	Urban built-up	Built-up in urban areas.	0.4005
	Rural settlement	Human settlements in rural areas.	0.2769
	Other construction land	Factories and mines, large-scale industrial districts, quarries, oil-fields, and transportation facilities.	0.3489
Unused land		Unused land, and land that is difficult to use.	
	Sand	Desert and other sandy land. Vegetation coverage < 5 %.	0.5730
	Gobi	Surface mainly composed of crushed gravel. Vegetation coverage < 5 %.	1.3903
	Saline/alkali soil	The accumulation of saline/alkali soil, with sparse saline-tolerant plants.	0.5480
	Marshland	District characterized by marshes, swamps, bogs, or the like.	0.6444
	Bare land	Dominated by soil, vegetation coverage < 5 %.	0.5023
	Bare rock	Rock or gravel cover > 5 %.	0.6818
Ocean		Marine area.	1.5408
	Class-wise average ICD		0.6145

convolution. The first approach is the same as that used in natural imagery, i.e., the spatially convolved features extracted from each spectral band are equally summed (Zhu et al., 2019). The second approach is designed especially for remote sensing images, i.e., an HSI 3D patch centered on a test sample is flattened to the format of (height × width) × bands, and 2D convolution is then performed on it (Luo et al., 2018). Meanwhile, considering that an HSI patch is naturally a 3D tensor, 3D convolution seems more reasonable for HSI classification. In fact, 28 of the 96 articles are based on 3D CNN models. In addition, there are 10 articles based on recurrent neural networks (RNNs) and 12 on generative adversarial networks (GANs).

In terms of the processing unit, most methods use a patch (i.e., a set of pixels within a spatial neighborhood) centered on a test pixel to mine the spectral-spatial information. However, as each sample can be repeatedly used as the spatial or neighboring context, this patch-based approach leads to the utilization of redundant information. Since HSI classification involves assigning a label for each pixel, pixel-based spectral-spatial sematic segmentation has also been a research hotspot. In this study, seven representative deep learning based HSI classification methods were chosen for a series of comprehensive tests on the WHU-OHS dataset (Table 5 and Fig. 5):

Gated recurrent unit (GRU) (Mou et al., 2017): Considering that each pixel can be modeled as an orderly and continuous spectral vector, a GRU is employed to simultaneously characterize the local spectral correlation and global band-to-band variability.

One-dimensional convolutional neural network (1D-CNN) (Hu et al., 2015): This model uses two 1D convolutional layers to hierarchically mine the local spectral correlation of each HSI pixel.

Two-dimensional convolutional neural network (2D-CNN) (Luo et al., 2018): In this model, all the pixels in a patch are separately fed into a single 1D convolutional layer, stitched into a 2D matrix, and are then transformed via two 2D convolutional layers to mine the spectral-spatial features.

HSI-GAN (Zhu et al., 2018): This model is composed of a generator (a 2D-CNN with five convolutional layers) that tries to generate fake patches that are as real as possible, and a discriminator (a 2D-CNN with six convolutional layers) that tries to identify the real patches. The spectral-spatial features can be extracted from the discriminator when the network reaches the Nash equilibrium.

Capsule network (CapsNet) (Zhu et al., 2019): To deal with the problem of CNNs, i.e., the 2D convolutions in the same layer are irrelevant, this network uses capsule layers to mine the correlation between channels, and simultaneously learn the spatial-channel features.

Three-dimensional convolutional neural network (3D-CNN) (Chen et al., 2016): This model utilizes two 3D convolutional layers to hierarchically mine the local spectral-spatial correlation of each HSI patch.

Three-dimensional fully convolutional neural network (3D-FCN) (Zou et al., 2020): This semantic segmentation network concatenates four 3D (i.e., spectral-spatial) and five 1D (i.e., spectral) fully convolutional layers (with a spatial stride of 1) to explore the spectral-spatial features for each HSI pixel.

4. Experiments on single OHS images

4.1. Experimental settings

From the perspective of pattern recognition, a classifier would expect that the training data and the test data have the same feature distribution. In this section, we evaluate the representative deep learning methods under this condition. To meet this condition, the experiments were independently conducted on each OHS image.

All the experiments were implemented on a hardware environment with an Inter(R) Core(TM) i9-9900X processor and an NVIDIA GeForce RTX 2080Ti GPU with 12 GB memory. For each network, the learning rate was exponentially decayed when the validation loss did not reduce during the training stage, and the batch size was set to 100. The other

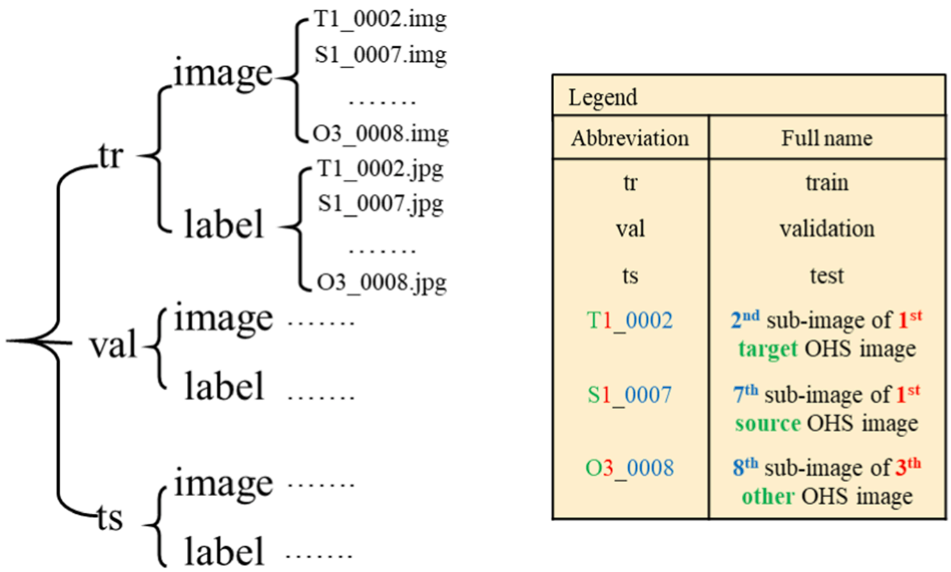


Fig. 3. Data organization of the WHU-OHS dataset.

hyper-parameters for each network are listed in Table 6.

4.2. Experimental results

The "Num. of images" in Table 7 indicates the numbers of OHS images that include the corresponding category. In Table 7, the class-wise performance is reported via the average of the F1-score (i.e., the geometric average of the producer's accuracy and user's accuracy), based on the 42 OHS images. In this table, the best and second-best results are highlighted with bold and underlined, respectively.

We first focus on the class-wise characteristics. On average, each image contains 13 classes, which is comparable to the existing datasets (Table 2). Except for a few categories (e.g., beach land, ocean, Gobi, and saline/alkali soil), most of the land-cover categories are widely distributed in the 42 OHS images. In view of the class-wise accuracies of the two spectral-only networks, the ocean, lake, and sand categories have desirable F1-scores, which is mainly due to their distinctive spectral characteristics. With the aid of spatial information, the performances for all the classes, except ocean, are improved, and the built-up and bare land classes can also achieve a desirable accuracy (i.e., all the F1-scores are >0.7).

We then compare the representative deep learning methods. The spectral-spatial networks are far superior to the spectral-only networks, and the 3D-FCN model achieves the best results. For the spectral-only networks, i.e., the 1D-CNN and GRU models, the former learns the local band-to-band variability with two 1D convolutional layers, and the latter simultaneously extracts the local spectral variability and the global spectral correlation through recurrent layers.

When analyzing the five spectral-spatial networks, the 2D-CNN model obtains the worst performance, even though it has the most parameters. The 3D-CNN model outperforms the 2D-CNN model by 5.4 % in the class average F1-score (CF1). This can be attributed to the fact that the 3D-CNN model considers the intrinsic tensor data structure of HSI data, while the spatial domain in the 2D-CNN model is flattened to one dimension. CapsNet uses a "capsule operator" to preserve the hierarchical relationships between the 2D convolutional filters, and achieves a CF1 increment of 4.8 %. The HSI-GAN model achieves a CF1 increment of 3.9 %, courtesy of the generator used to simulate the HSI signals. The HSI-GAN and CapsNet models show comparable performances, whereas the 3D-FCN and 3D-CNN models achieve superior accuracies. The 3D-FCN model obtains the best results, overall, due to its deeper network layers and the concatenation of the spectral-spatial features extracted by the 3D convolutional layers and the spectral features extracted by the 1D convolutional layers.

In the following, we compare the existing hyperspectral datasets and the newly developed WHU-OHS dataset, based on these representative methods. As reported in the previous articles (Hu et al., 2015; Chen et al., 2016; Mou et al., 2017; Luo et al., 2018; Zhu et al., 2018; Zhu et al., 2019; Zou et al., 2020), all the representative methods applied on the existing datasets can reach performance levels of OA $>$ 98 %, which are much higher levels than those achieved on the WHU-OHS dataset (Table 7). The reasons for this are summarized as follows.

Firstly, owing to the small spatial coverage of the existing HSI datasets, although the training and test samples are spatially disjoint, it is still difficult to meet the assumption of the training and test data being independent. Although the simulated HSI signals of the HSI-GAN model and the pixel-wise procedure of the 3D-FCN model can be helpful for the classification task with limited training samples, the parameters of these two networks (i.e., HSI-GAN: 1.70 mil.; 3D-FCN: 2.04 mil.) are still huge, leading to a risk of model overfitting. Thus, the accuracy assessed by the use of the test samples in the existing datasets might be overestimated, as demonstrated in Zou et al. (2020).

Secondly, all the existing datasets (except for the Botswana dataset) were obtained from airborne remote sensing platforms, and the imaging conditions of the airborne platforms are more stable than those of satellite observations. That is, although both platforms observe real scenarios, samples in the airborne datasets are purer, due to their higher spatial resolution and lower signal-to-noise ratio. In contrast the samples in the new WHU-OHS satellite dataset are more diverse. Thus, the WHU-OHS dataset is much more challenging than the existing datasets, and the large number of samples is more appropriate for evaluating the advanced and complex networks in a more fair and reasonable manner.

5. Transfer learning experiments on OHS images

5.1. Experimental settings

As a real-world HSI classification task often provides a limited number of training samples, transfer learning, which involves applying information extracted from a source HSI scene to the target scene, is a practical technique. Considering the large sample size and the wide

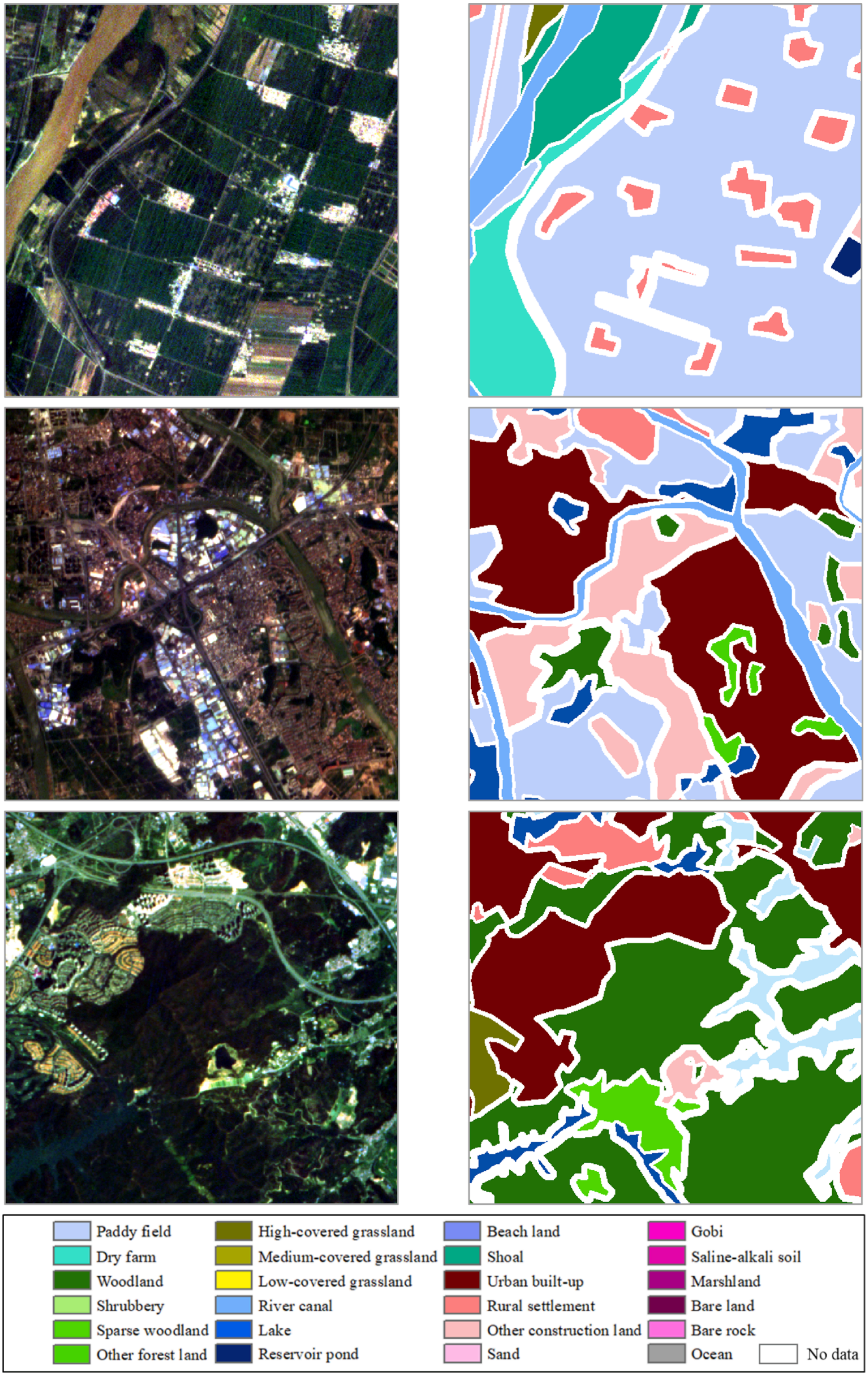


Fig. 4. Samples from the WHU-OHS dataset. The top sub-image is dominated by cropland (including both paddy field and dry farmland), and several rural settlements are scattered in the scene. The sub-images in the middle and bottom are all located in the suburbs and contain human-dominated landscapes and natural objects. The middle sub-image is in a plain with a dense network of rivers, and the bottom sub-image is of a mountainous city with various woodland types.

Table 5
Characteristics of the seven representative deep learning based HSI classification methods.

	Feature type	Model	Numbers of parameters*	Processing unit
GRU	Spectral	RNN	53.53 k	Pixel
1D-CNN	Spectral	CNN with 1D conv	62.60 k	Pixel
2D-CNN	Spectral-spatial	CNN with 1D and 2D conv	30.52 mil	Patch
HSI-GAN	Spectral-spatial	CNN with 2D conv and GAN	1.70 mil	Patch
CapsNet	Spectral-spatial	Capsule layer and CNN with 2D conv	104.54 k	Patch
3D-CNN	Spectral-spatial	CNN with 3D conv	233.59 k	Patch
3D-FCN	Spectral-spatial	CNN with 3D conv	2.041 mil	Pixel

*In this table, k is the abbreviation for thousand, and mil. is the abbreviation for million.

geographical distribution, it is of interest to investigate the potential of the WHU-OHS dataset as source data for transfer learning.

Specifically, we employed two types of source domain settings. In the first case (Fig. 6a), we chose eight pairs of OHS images, in which each pair contained one source image and one target image, which were spatially adjacent and had similar land-cover landscapes. In fact, the class-wise average ICD of each pair (see Supplementary Material II for more details) suggested that domain shift existed. To comprehensively test the transferability, the following three scenarios were adopted:

S-I: source to target direct prediction is to train the target network according to the training samples from the source domain only.

S-II: target training from scratch indicates that the network is randomly initialized and then trained using the samples from the target image.

S-III: source pre-training and target fine-tuning indicates that the network is first trained by the training samples from the source image, and is then fine-tuned with the training samples from the paired target image.

As illustrated in Fig. 6b, the second experimental setting was on the basis of all 42 OHS images. These images were divided into two parts: the eight target images used in the first experiment were still considered to be the targets, and the other 34 images were used for the pre-training. Please note that, in order to fairly compare the results of the first and second experiments, the network was pre-trained with all 34 source images and tested on each of the eight target images. This experimental setting was aimed at testing whether large-scale pre-training can boost the classification accuracy for the target domain. The following three test scenarios were designed:

M-I: represents the direct prediction of the target image by the network trained with only the samples of the 34 source images.

M-II: the same as S-II.

M-III: indicates that the network was pre-trained by the samples from the 34 source images and then further fine-tuned by the samples from the target image.

Considering the performance and the complexity of the seven deep networks mentioned above, we chose the 3D-CNN and 3D-FCN models in these experiments. During the pre-training and fine-tuning phase of each network, the learning rate was exponentially decayed when the validation loss did not reduce during the training stage, and the batch size was set to 100. The hyper-parameters for each network are listed in Table 8. As scenario II (including both S-II and M-II) is actually the initialization of scenario III, the hyper-parameters for the pre-training of scenario III and the training of scenario II were the same. The training and test data for these images were the same as those in Section 4.

5.2. Experimental results

Generally speaking, as seen in Table 9, the CF1 values of all the direct predictions are lower than those of training from scratch, even though the source and target images were collected from the same HSI sensor (i.e., OHS). This indicates that the imaging differences severely limit the generalization ability of the deep learning networks, and the knowledge from the target image is still essential for model transfer. In the case of the 3D-CNN model, with the pre-learned features, most of the fine-tuned networks achieve higher CF1 scores, and the increments increase with the number of source training samples. The 3D-FCN model is inferior to the 3D-CNN model in scenario I, indicating the better generalization ability of the latter. Meanwhile, the 3D-FCN model obtains a higher CF1 score in scenario II, suggesting that the 3D-FCN model depends more on the information from the local image scene, due to its extra 1D spectral convolutional branch. When we focus on scenario III, it is clear that the benefits of a large-scale dataset are more significant for the 3D-CNN model.

For the class-wise accuracy, in scenario I, the two farmland classes, low-covered grassland, and the three built-up classes show a desirable generalization ability, indicating the potential for the related thematic mapping tasks. However, by observing the performance difference between scenarios I and II for ocean, lake, sand, and bare land, the samples of these categories in the target domain are essential, and the imaging difference severely restricts the model generalization.

From the CF1 accuracy, it can also be seen that the results for scenario II are better than those for scenario I, in most cases. Considering the accuracy difference between scenarios II and III, the 3D-CNN and 3D-FCN models achieve 86 and 57 positive transfers (i.e., accuracy of III $>$ accuracy of II) in the 100 class transfers (Table 8), respectively. This infers that, in most cases, the information (i.e., samples) from the source images is beneficial for improving the classification of the target images. In particular, the results show that the 3D-CNN model has better transferability than the 3D-FCN model. Meanwhile, the negative transfer indicates that the pre-trained features may contain source domain biased information, which degrades the performance of the fine-tuning. In summary, the transfer learning experiments confirm the potential of pre-training features on the large-scale WHU-OHS dataset, and also indicate the requirement for designing more advanced transfer learning methods.

6. Transfer learning from the WHU-OHS dataset to the existing datasets

Considering the large differences in spectral range and resolution between the existing hyperspectral sensors (Table 2), whether and how knowledge can be transferred and applied across sensors is an important question. In this section, we describe the experiments conducted to investigate the transferability of the WHU-OHS dataset to the other existing and widely used hyperspectral datasets.

6.1. Experimental settings

Nine public and commonly used hyperspectral datasets, i.e., Indian Pines, Salinas, KSC, Botswana, Pavia U, Pavia C, DFC2013, DFC2018, and Washington DC Mall (see Table 2 for details), were tested in these experiments. The training and test sets of the Indian Pines and Pavia University datasets were downloaded from https://dase.grss-ieee.org, and the training and test sets of the DFC2013, DFC2018, and Washington DC Mall datasets were downloaded from https://hyperspectral.ee.uh.edu/ and https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html, respectively. For the other five datasets, we randomly divided the whole ground truth into two sets, ensuring that the training set and test set were spatially disjoint. Please note that the spatial resolution of the WHU-OHS dataset is 10 m, whereas the first four public datasets have a spatial resolution of 20–30 m, and the other five

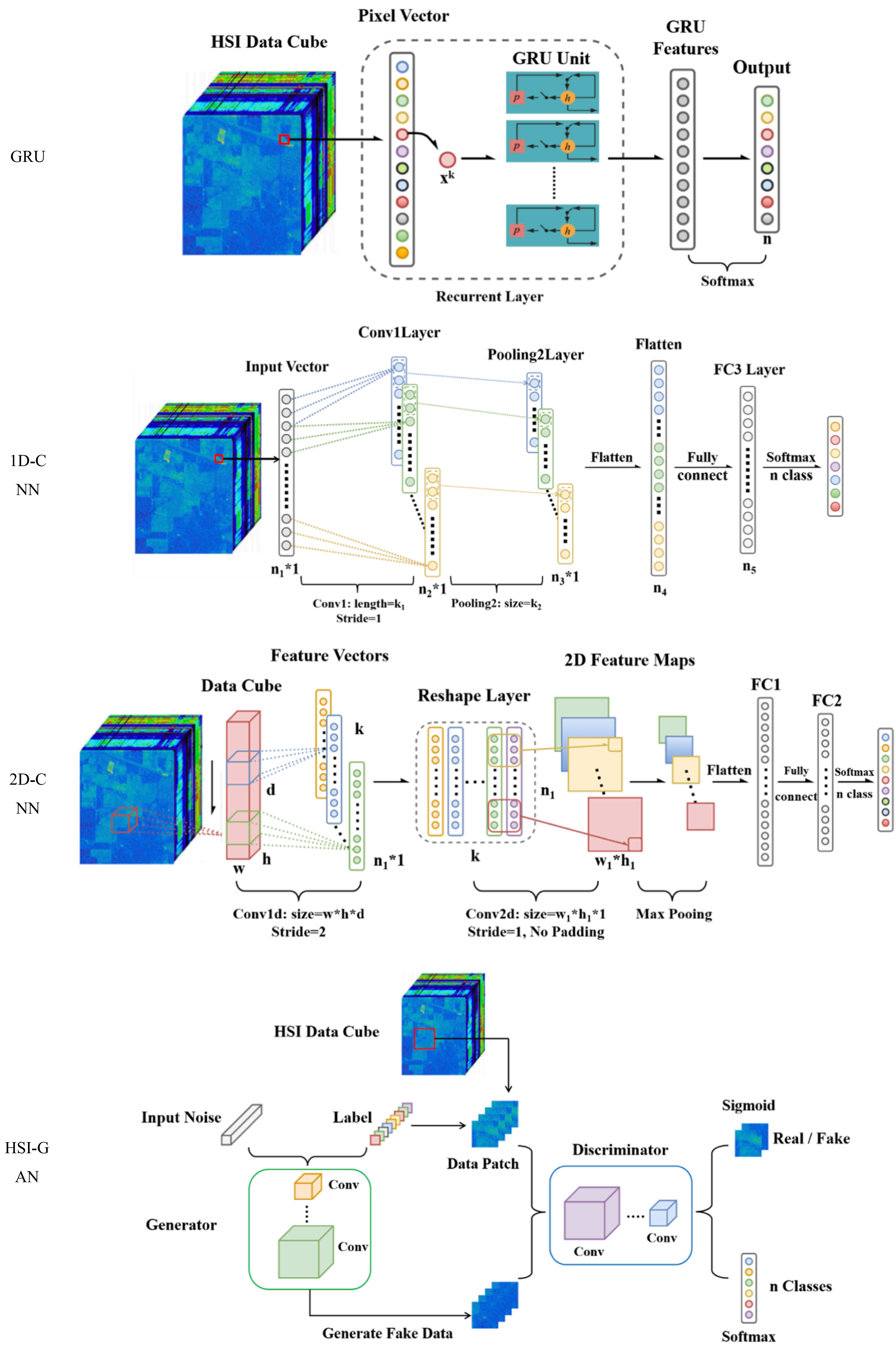


Fig. 5. Frameworks of the seven representative deep learning based HSI classification methods.

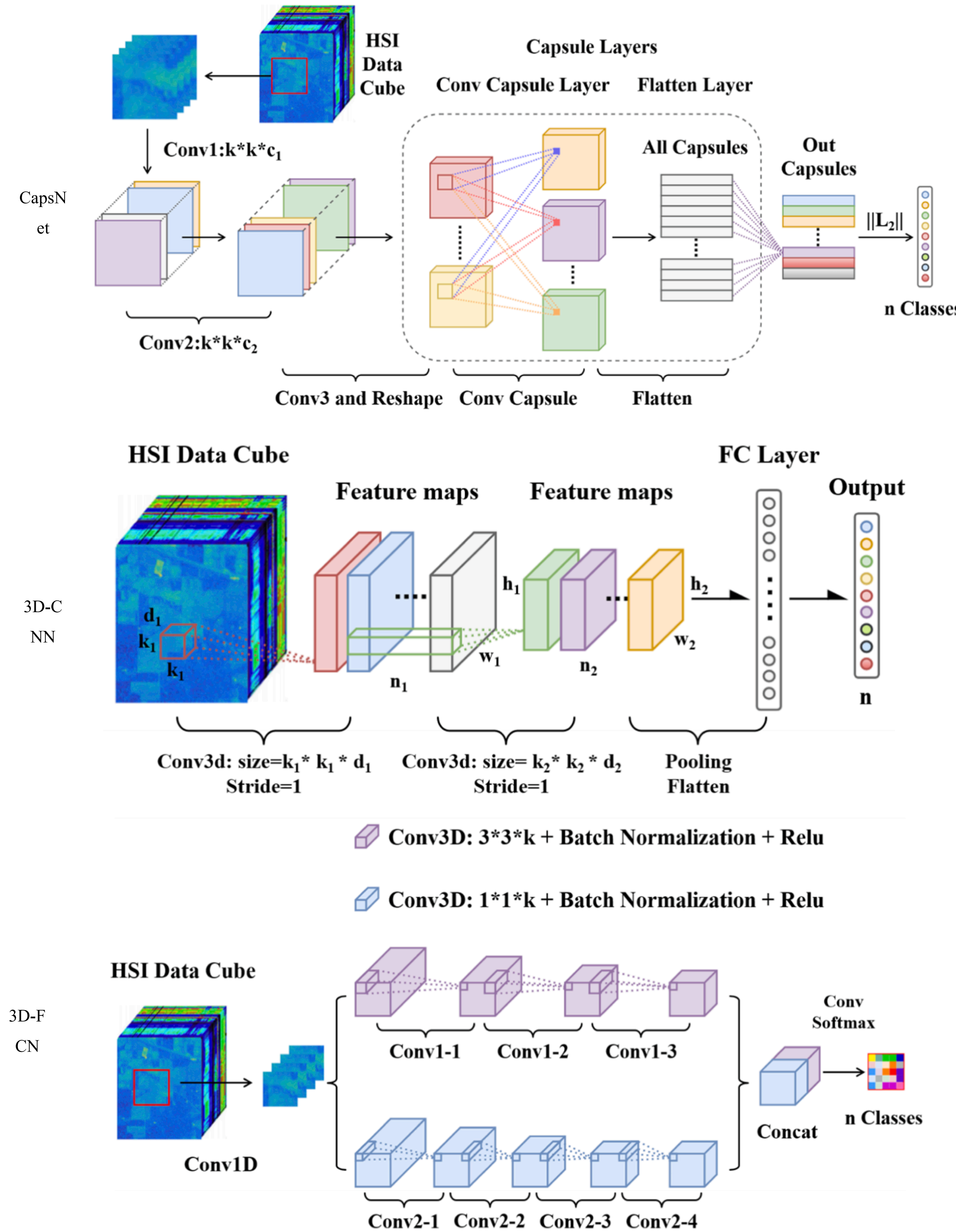


Fig. 5. (*continued*).

datasets have a very high spatial resolution of 0.5–2 m. The numbers of training and test samples for each dataset are presented in Table 10.

Meanwhile, considering that these public datasets have distinct spectral bands and intervals, the following four scenarios were conducted to test the transferability from the WHU-OHS dataset to the other existing public HSI datasets:

(I) Full-Random: As a baseline, the network was randomly initialized and trained from scratch. All the spectral bands were used in this approach, without additional information from the WHU-OHS dataset.

(II) Sub-Random: Considering the large difference in spectral range and spatial resolution between these hyperspectral datasets, 32 bands were manually selected for each public dataset, based on the central wavelength of each spectral band of the OHS images, and the network was trained from scratch. The 32 sub-channels of all the spectral bands were used in this approach, without information transferred from the WHU-OHS dataset.

(III) Sub-OHS: The network that was pre-trained on the WHU-OHS dataset was fine-tuned by the use of the training samples from the 32 selected bands.

(IV) Full-OHS: The network that was pre-trained on the WHU-OHS dataset was fine-tuned. All the spectral bands and additional information from the WHU-OHS dataset were used in this approach. Thus, the

Table 6
Hyper-parameters used to train each deep network on single OHS images.

Network	Initial learning rate	Optimizer	Num. of epochs
1D-CNN	0.001	Adam	20
GRU	0.001	Adam	20
2D-CNN	0.1	Stochastic gradient descent (SGD)	40
HIS-GAN	Generator: 0.00002 Discriminator: 0.0002	Generator: Adam Discriminator: SGD	100
CapsNet	0.001	Adam	35
3D-CNN	0.01	SGD	40
3D-FCN	0.001	Adam	45

initial networks for fine-tuning were composed of the parameters pre-trained by the WHU-OHS dataset and other random initialization parameters.

The 3D-CNN and 3D-FCN models were chosen in these experiments, with the OA, Kappa, and CF1 scores used for the evaluation. Considering the large training sample sets of the Pavia C and DFC2018 datasets, the epochs for these two datasets were set to 50, and for the other seven datasets, the epochs were set to 200 and 100 for the 3D-CNN and 3D-FCN models, respectively. For each public dataset, stochastic gradient descent (SGD) was employed as the optimizer for the 3D-CNN model, with a learning rate of 0.001 and a batch size of 100. The Adam optimizer was used with the 3D-FCN model, with a learning rate of 0. 0001 and a batch size of 100.

6.2. Experimental results

It is clear from Fig. 7 that, for the 3D-CNN model, the Full-OHS scenario is superior to the other three scenarios on eight datasets (except for Salinas), in terms of OA, Kappa, and particularly CF1. For the 3D-FCN model, Sub-OHS performs better than Sub-Random on eight datasets (except for Washington DC Mall), and Full-OHS outperforms

Table 7
Average accuracy of the seven representative methods for the 42 OHS images.

Class name	Num. of images	1DCNN	GRU	2D-CNN	HSI-GAN	3D-CNN	Capsule	3D-FCN
Paddy field	25	0.572	0.573	0.624	0.654	0.701	0.679	**0.713**
Dry farm	35	0.584	0.583	0.602	0.634	**0.680**	0.670	0.679
Woodland	35	0.635	0.633	0.659	0.703	0.718	0.713	**0.747**
Shrubbery	20	0.340	0.382	0.452	0.461	0.494	0.432	**0.551**
Sparse woodland	20	0.263	0.284	0.326	0.363	0.367	0.390	**0.419**
Other forest land	30	0.285	0.294	0.340	0.402	0.418	0.421	**0.466**
High-covered grassland	36	0.327	0.338	0.405	0.425	0.430	0.459	**0.561**
Medium-covered grassland	26	0.342	0.338	0.395	0.442	0.426	0.442	**0.513**
Low-covered grassland	19	0.442	0.447	0.490	0.530	0.574	0.566	**0.599**
River/canal	36	0.579	0.586	0.624	**0.702**	0.693	0.667	0.687
Lake	19	0.832	0.819	0.851	0.866	**0.894**	0.852	0.893
Reservoir/pond	35	0.672	0.672	0.694	0.728	0.739	0.748	**0.774**
Beach land	4	0.562	0.620	0.610	0.629	0.666	0.600	**0.678**
Shoal	28	0.398	0.398	0.476	0.534	0.550	0.505	**0.649**
Urban built-up	39	0.649	0.645	0.702	0.760	0.784	0.789	**0.792**
Rural settlement	40	0.453	0.455	0.484	0.545	**0.616**	0.593	0.594
Other construction land	40	0.462	0.467	0.493	0.548	**0.624**	0.587	0.600
Sand	5	0.715	0.719	0.759	0.759	0.803	0.794	**0.831**
Gobi	3	0.579	0.572	0.663	0.716	0.502	0.695	**0.732**
Saline/alkali soil	5	0.563	0.544	0.582	0.655	0.615	0.649	**0.662**
Marshland	16	0.542	0.551	0.630	0.684	0.677	0.688	**0.752**
Bare land	16	0.642	0.629	0.715	0.741	**0.787**	0.713	0.778
Bare rock	16	0.509	0.516	0.631	0.674	0.728	0.703	**0.779**
Ocean	4	0.960	0.959	0.949	0.949	**0.965**	0.962	0.948
Class average F1-score (CF1)		0.538	0.543	0.590	0.629	0.644	0.638	**0.683**

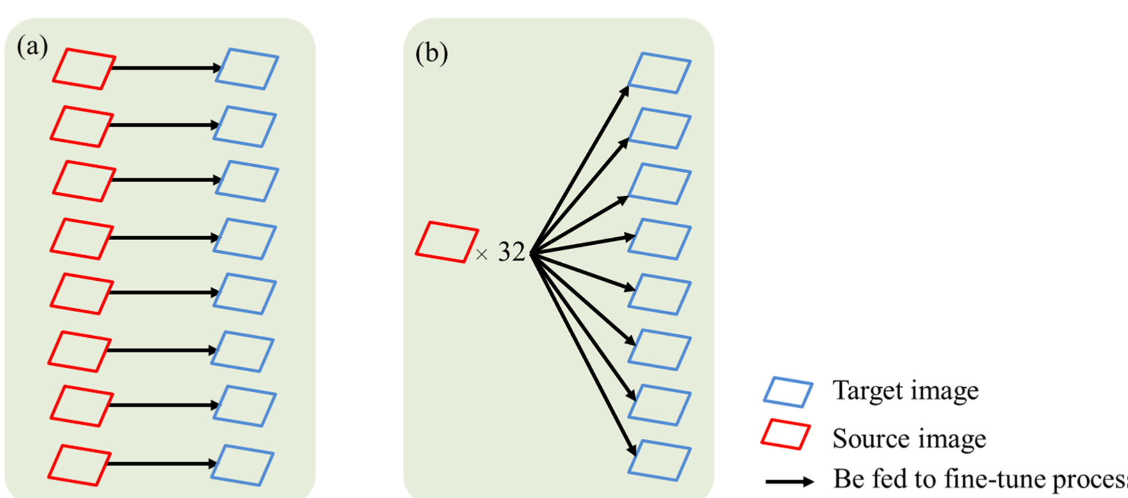


Fig. 6. The two types of source domain settings in the transfer learning experiments: (a) the first experimental setting (single source, involving S-I, S-II, and S-III); and (b) the second setting (multiple sources, involving M-I, M-II, and M-III).

Table 8
Hyper-parameters used in the experimental settings described in Section 5.

Deep network	Phase	Initial learning rate	Optimizer	Num. of epochs
3D-CNN	S-III: fine-tune	0.01	SGD	40
	M-III: pre-train	0.0001	SGD	40
	M-III: fine-tune	0.0001	SGD	40
3D-FCN	S-III: fine-tune	0.001	Adam	45
	M-III: pre-train	0.000001	Adam	20
	M-III: fine-tune	0.000001	Adam	45

Full-Random on five datasets (but not Pavia C, KSC, DFC 2013, and DFC 2018). Please note that, for the cases of the exceptions, the accuracy degradation is marginal, but in the cases of the positive examples, the accuracy gains obtained from the WHU-OHS pre-training are significant. In particular, although the spatial resolution of the WHU-OHS dataset is 10 m, it is able to bring an accuracy increment for both the middle-resolution datasets (i.e., 20–30 m) and the high-resolution datasets (i.e., meter/sub-meter).

Furthermore, the training loss and the training accuracy curves are presented in Fig. 8 to analyze the gains obtained with pre-training on the WHU-OHS dataset. In the case of the 3D-CNN model, with the assistance of the WHU-OHS dataset, the Sub-OHS scenario achieves both better and faster convergence. In the case of the 3D-FCN model, the curves for the Sub-OHS scenario are inferior, even though the classification accuracy for the Sub-OHS scenario is superior. This can possibly be attributed to the over-parameterization of the networks. Specifically, the training sample size for the Indian Pines, Pavia U, and DFC2013 datasets was 4457, 2381, and 2253, respectively, and the parameter size for the 3D-CNN and 3D-FCN models was 233.59 k and 2.041 mil., respectively. Thus, the 3D-FCN model may suffer more from model overfitting than the 3D-CNN model, while the overfitting of the 3D-FCN model can be effectively relieved courtesy of the pre-training on the WHU-OHS dataset. In summary, it can be concluded that the proposed WHU-OHS dataset can make a positive contribution when classifying the existing HSI datasets.

7. Discussion: The use of the WHU-OHS dataset in future work

A direct use of the WHU-OHS dataset will be as an HSI classification benchmark, for developing and testing new algorithms. Future HSI classification networks will be able to go deeper with more complicated but flexible modules (e.g., the 3D swin transformer (Liu et al., 2022) and non-local blocks (Zhu et al., 2021)) and architectures (Wang et al., 2020). With this large-scale hyperspectral dataset, researchers should be able to design a backbone specifically for HSI classification.

From the results of the transferability testing, it was confirmed that the WHU-OHS dataset can be used for pre-training, and the knowledge can be transferred to other hyperspectral datasets. Moreover, in this study, the band selection during the transfer can be regarded as a "prompt processing" (Ben-David et al., 2021), which reformulates and adapts the downstream task according to the upstream task. In this context, the downstream task is classification of a certain specific hyperspectral dataset, and the upstream task is pre-training on the WHU-OHS dataset. The new "prompt-based learning" paradigm, as well as self-supervised learning (Chen et al., 2020), is becoming a research

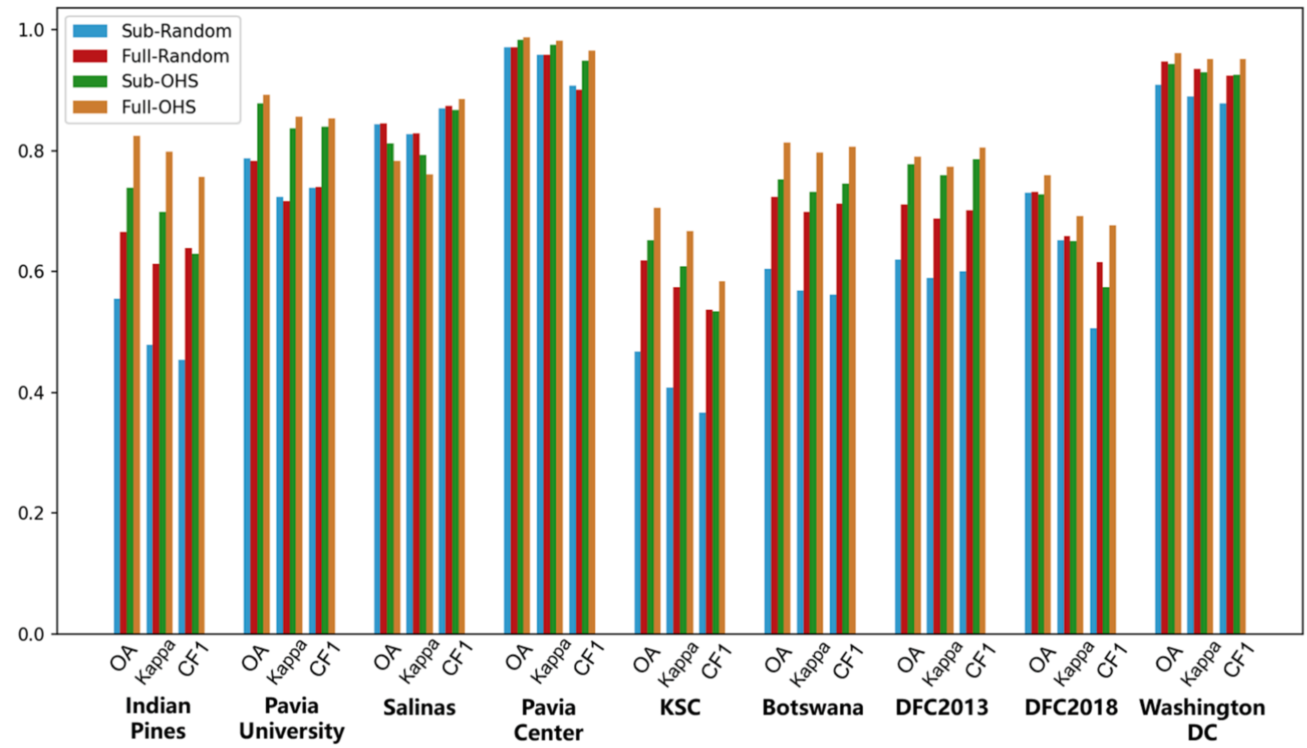
Table 10
Sample sizes for the nine public hyperspectral datasets.

Existing dataset	Number of training samples	Number of test samples
Indian Pines	4457	4384
Pavia U	2381	36,825
Salinas	13,916	35,711
Pavia C	46,243	93,028
KSC	1524	3299
Botswana	1248	1824
DFC2013	2533	12,074
DFC2018	219,397	248,407
Washington DC Mall	672	19,159

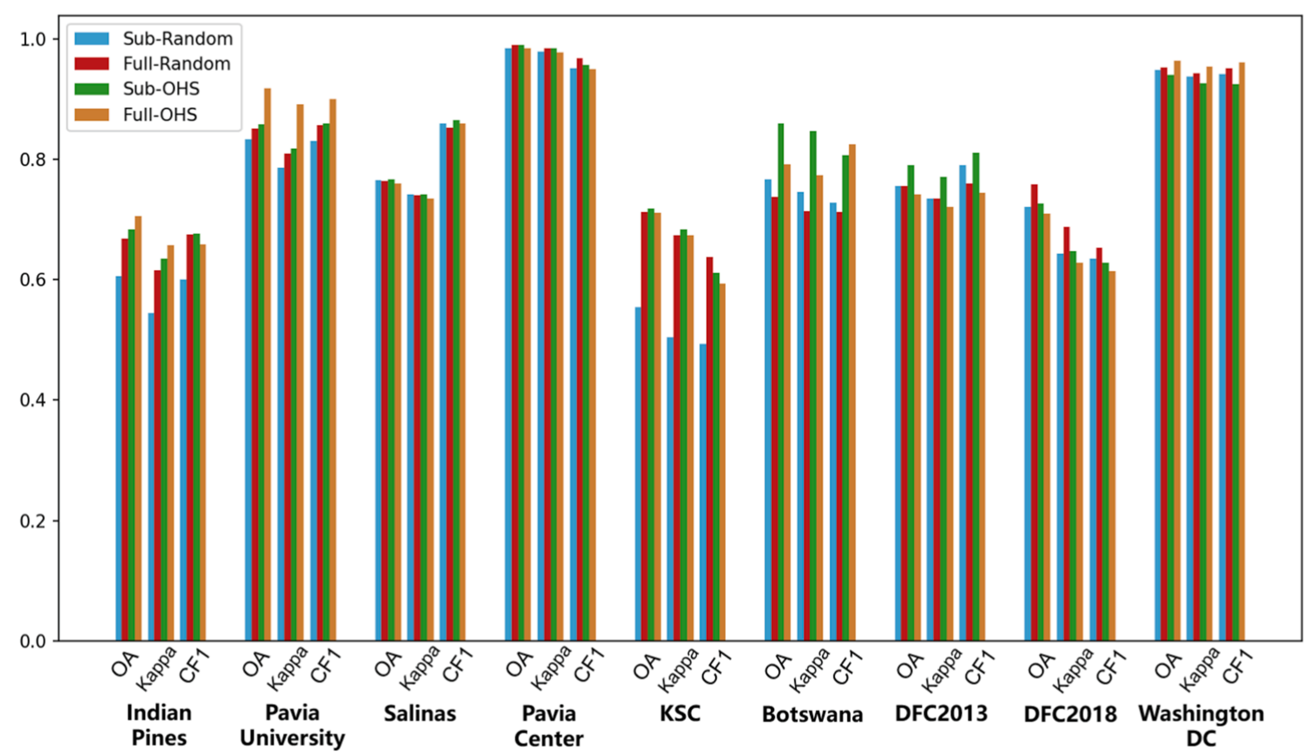
Table 9
Direct prediction (I), training from scratch (II), and fine-tuning (III) on eight target images by the use of single (S) and multiple (M) source OHS images.

Network		3D-CNN					3D-FCN				
Class name	Num. of target images	**S-I**	**M-I**	**S-II**	**S-III**	**M-III**	**S-I**	**M-I**	**S-II**	**S-III**	**M-III**
Paddy field	5	0.354	0.443	0.711	0.721	**0.726**	0.333	0.253	**0.690**	0.684	0.684
Dry farm	4	0.307	0.384	0.729	0.747	**0.750**	0.234	0.278	0.692	**0.715**	0.689
Woodland	7	0.382	0.516	0.666	**0.678**	0.685	0.277	0.547	0.675	**0.686**	0.680
Shrubbery	2	0.000	0.114	0.169	**0.187**	0.178	0.000	0.002	0.167	**0.182**	0.172
Sparse woodland	3	0.001	0.021	0.187	**0.224**	0.209	0.001	0.139	0.192	**0.205**	0.193
Other forest land	6	0.053	0.077	**0.240**	0.221	0.239	0.006	0.029	0.244	**0.251**	0.248
High-covered grassland	7	0.016	0.098	0.182	**0.269**	0.251	0.001	0.113	**0.330**	0.291	0.314
Medium-covered grassland	3	0.042	0.193	0.409	0.579	**0.591**	0.025	0.309	0.499	**0.519**	0.512
Low-covered grassland	2	0.294	0.055	**0.733**	0.665	0.667	0.236	0.082	**0.584**	0.555	0.579
River/canal	7	0.218	0.627	0.864	0.883	**0.896**	0.194	0.451	**0.827**	0.809	0.819
Lake	2	0.079	0.250	**0.967**	0.957	0.960	0.067	0.558	0.978	0.974	**0.979**
Reservoir/pond	8	0.134	0.360	0.585	0.609	**0.639**	0.093	0.283	0.694	0.702	**0.703**
Beach land	1	0.000	0.000	0.556	**0.605**	0.550	0.000	0.002	0.577	0.552	**0.641**
Shoal	5	0.053	0.009	0.589	**0.635**	0.628	0.009	0.009	0.667	0.667	**0.676**
Urban built-up	8	0.329	0.588	**0.836**	0.825	0.819	0.182	0.418	**0.800**	0.794	0.797
Rural settlement	7	0.243	0.334	0.574	0.625	**0.627**	0.201	0.247	**0.527**	0.524	0.508
Other construction land	8	0.336	0.449	**0.724**	0.664	0.659	0.180	0.370	**0.648**	0.645	0.639
Gobi	1	0.051	0.000	0.113	0.836	**0.846**	0.031	0.000	0.792	**0.835**	0.809
Saline/alkali soil	1	0.000	0.000	0.833	0.859	**0.883**	0.000	0.002	0.726	**0.800**	0.783
Marshland	6	0.000	0.000	0.675	0.733	**0.770**	0.000	0.000	0.771	0.743	**0.777**
Bare land	4	0.001	0.001	**0.729**	0.699	0.728	0.000	0.000	0.595	**0.657**	0.610
Bare rock	2	0.050	0.004	0.580	**0.701**	0.696	0.041	0.002	0.858	**0.865**	0.849
Ocean	1	0.000	0.024	0.953	0.955	**0.977**	0.000	0.228	0.921	**0.937**	0.932
Class average F1-score (CF1)		0.170	0.280	0.601	0.629	**0.635**	0.118	0.235	**0.619**	**0.619**	**0.619**

* ∑Num. of target images = 100. The CF1 accuracy in the table indicates the average for the 100 classes.
MII is the same as S-II.
For each accuracy metric, the best and second-best results are highlighted in bold and underlined, respectively.



(a)



(b)

Fig. 7. Accuracy assessment on the nine public hyperspectral datasets: (a) 3D-CNN, (b) 3D-FCN. Each set of columns illustrates the results obtained on one public hyperspectral dataset. The blue, red, green, and orange columns refer to the quantitative evaluation of the Sub-Random, Full-Random, Sub-OHS, and Full-OHS scenarios, respectively. The OA, Kappa, and CF1 values are presented from left to right.

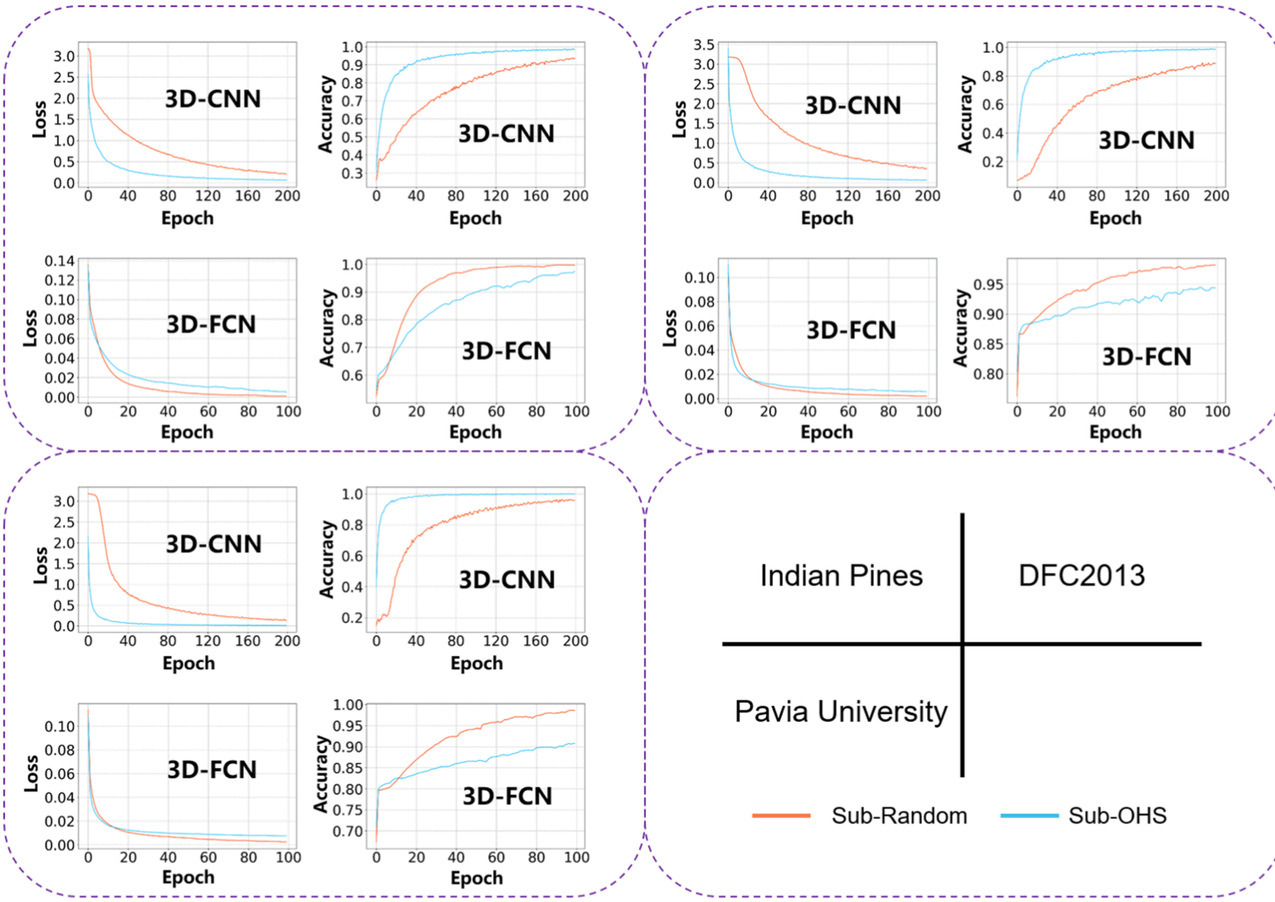


Fig. 8. Training loss and training overall accuracy for the 3D-CNN and 3D-FCN models on the Indian Pines (upper left), Pavia U (bottom left), and DFC2013 (upper right) datasets. The larger/smaller the area under the curve of the training loss/training overall accuracy curve, the better the performance.

hotspot in the field of natural image processing. The WHU-OHS dataset should be an appropriate dataset for testing these new learning paradigms in HSI processing.

Considering the wide spatial distribution, large spatial coverage, diverse geographical landscapes, and the widely used LULC classification system, the WHU-OHS represents an interesting and promising large-scale hyperspectral dataset for pre-training, training, validation, and testing.

8. Conclusion

With the wide spatial distribution and large spatial coverage, the WHU-OHS dataset represents the largest HSI classification dataset to date. According to the comprehensive experiments conducted in this study, it was confirmed that the WHU-OHS dataset is much more challenging than the existing hyperspectral datasets, due to its large data volume, wide geophysical distribution, and large sample size. In terms of the transferability testing, it was found that the large-scale dataset contains diverse imaging conditions and geographical landscapes, and hence overcomes the limitations of the imaging differences between various sensors and study areas, to some extent. Thus, the WHU-OHS dataset represents a challenging data benchmark for hyperspectral image classification, especially in the era of deep learning. Moreover, the open-access PyTorch toolbox with seven representative deep neural networks for large-scale HSI image classification will also be beneficial to the development of this field.

CRediT authorship contribution statement

Jiayi Li: Conceptualization, Methodology, Writing – original draft. **Xin Huang:** Supervision. **Lilin Tu:** Software, Validation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The research was supported by the National Natural Science Foundation of China (under Grants 42071311, 41971295 and 42271328), the Special Fund of Hubei Luojia Laboratory (under Grant 220100031), and Wuhan 2022 Shuguang Project (under Grant 2022010801020123).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2022.103022.

References

Altman, E.I., Marco, G., Varetto, F., 1994. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). J. Bank. Finance 18 (3), 505–529. https://doi.org/10.1016/0378-4266(94)90007-8.

Audebert, N., Saux, B.L., Lefevre, S., 2019. Deep Learning for Classification of Hyperspectral Data: A Comparative Review. IEEE Geosci. Remote Sens. Mag. 7, 159–173. https://doi.org/10.1109/MGRS.2019.2912563.

Berk, A., Anderson, G.P., Bernstein, L.S., Acharya, P.K., Dothe, H., Matthew, M.W., Adler-Goldena S.M., Chetwynd J.H., Jr.b, Richtsmeiera, S.C., Pukallb, B., Allredb, C.L. , Jeongb , L.S., Hoke, M.L., 1999. MODTRAN4 radiative transfer modeling for atmospheric correction. In: Optical spectroscopic techniques and instrumentation for atmospheric and space research III, 3756, 348-353. 10.1117/12.366388.

Ben-David, E., Oved, N., Reichart, R., 2021. PADA: A Prompt-based Autoregressive Approach for Adaptation to Unseen Domains. https://arxiv.org/abs/2102.12206.

Cen, Y., Zhang, L., Zhang, X., Wang, Y., Qi, W., Tang, S., Zhang, P., 2020. Aerial hyperspectral remote sensing classification dataset of Xiongan New Area (Matiwan Village). J. Rem. Sens. (Chinese) 24 (11), 1299–1306. https://doi.org/10.11834/jrs.20209065.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A Simple Framework for Contrastive Learning of Visual Representations. PMLR 1597–1607.

Chen, Y., Bruzzone, L., 2022. Self-supervised Change Detection in Multi-view Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P., 2016. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. IEEE Trans. Geosci. Remote Sens. 54, 6232–6251. https://doi.org/10.1109/TGRS.2016.2584107.

Danielson J.J., Gesch, D.B., 2011 Global multi-resolution terrain elevation data 2010 (GMTED2010), Washington, DC, USA: US Department of the Interior, US Geological Survey.

Datta, D., Mallick, P.K., Bhoi, A.K., Ijaz, M.F., Shafi, J., Choi, J., Ding, B., 2022. Hyperspectral Image Classification: Potentials, Challenges, and Future Directions. Comput. Intell. Neurosci. 2022, 1–36.

Gao, H., Chen, Z., Xu, F., 2022. Adaptive spectral-spatial feature fusion network for hyperspectral image classification using limited training samples. Int. J. Appl. Earth Obs. 107, 102687. https://doi.org/10.1016/j.jag.2022.102687.

Hu, T., Huang, X., Li, J., Zhang, L., 2018. A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery. Remote Sens. Environ. 217, 144–157. https://doi.org/10.1016/j.rse.2018.08.017.

Hu, W., Huang, Y., Wei, L.i., Zhang, F., Li, H., 2015. Deep Convolutional Neural Networks for Hyperspectral Image Classification. J. Sens. 2015, 1–12.

Hu, X., Chu, L., Pei, J., Liu, W., Bian, J., 2021. Model Complexity of Deep Learning: A Survey. Knowl. Inf. Syst. 63, 2585–2619. https://doi.org/10.1007/s10115-021-01605-0.

Jaiswal, G., Arun, S., Sumit, K.Y., 2021. Critical insights into modern hyperspectral image applications through deep learning. Wiley Interdiscipl. Rev.: Data Min. Knowledge Discov. 11 (6), e1426.

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., Tao, C., 2021. Remote Sensing Images Semantic Segmentation with General Remote Sensing Vision Model via a Self-Supervised Contrastive Learning Method. https://arxiv.org/abs/2106.10605.

Li, J., Huang, X., Gong, J., 2019a. Deep neural network for remote-sensing image interpretation: Status and perspectives. Natl. Sci. Rev. 6, 1082–1086. https://doi.org/10.1093/nsr/nwz058.

Li, J., Huang, X., Chang, X., 2020. A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis. ISPRS J. Photogramm. Remote Sens. 163, 1–17. https://doi.org/10.1016/j.isprsjprs.2020.02.022.

Li, J., Zhang, H., Huang, Y., Zhang, L., 2014. Hyperspectral Image Classification by Nonlocal Joint Collaborative Representation with a Locally Adaptive Dictionary. IEEE Trans. Geosci. Remote Sens. 52 (6), 3707–3719. https://doi.org/10.1109/TGRS.2013.2274875.

Li, J., Zhang, H., Zhang, L., 2015a. A Nonlinear Multiple Feature Learning Classifier for Hyperspectral Images With Limited Training Samples. IEEE. J. Sel. Topics Appl. Earth Observ. 8, 2728–2738. https://doi.org/10.1109/JSTARS.2015.2400634.

Li, J., Zhang, H., Zhang, L., 2015b. Efficient Superpixel-Level Multitask Joint Sparse Representation for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 53 (10), 5338–5351. https://doi.org/10.1109/TGRS.2015.2421638.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019b. Deep Learning for Hyperspectral Image Classification: An Overview. IEEE Trans. Geosci. Remote Sens 57, 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932.

Liang, X., Zhang, Y., Zhang, J., 2021. Attention Multisource Fusion-Based Deep Few-Shot Learning for Hyperspectral Image Classification. IEEE. J. Sel. Topics Appl. Earth Observ. 14, 8773–8788. https://doi.org/10.1109/JSTARS.2021.3109012.

Liu, C., Chen, L.-C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Li, F., 2019. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. CVPR 82–92. https://doi.org/10.1109/CVPR.2019.00017.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video Swin Transformer. CVPR 3202–3211.

Luo, Y., Zou, J., Yao, C., Zhao, X., Li, T., Bai, G., 2018. HSI-CNN: A Novel Convolution Neural Network for Hyperspectral Image. ICALIP 464–469.

Manolakis, D.G., Lockwood, R.B., Cooley, T.W., 2016. Hyperspectral imaging remote sensing: physics, sensors, and algorithms. Cambridge University Press.

Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep Recurrent Neural Networks for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 55, 3639–3655. https://doi.org/10.1109/TGRS.2016.2636241.

Rangnekar, A., Mokashi, N., Ientilucci, E.J., Kanan, C., Hoffman, M.J., 2020. AeroRIT: A New Scene for Hyperspectral Image Analysis. IEEE Trans. Geosci. Remote Sens. 58, 8116–8124. https://doi.org/10.1109/TGRS.2020.2987199.

Rao, W., Qu, Y., Gao, L., Sun, X., Wu, Y., Zhang, B., 2022. Transferable network with Siamese architecture for anomaly detection in hyperspectral images. Int. J. Appl. Earth Obs. 106, 102669. https://doi.org/10.1016/j.jag.2021.102669.

Rinker, J.N., 1990. Hyperspectral imagery - What is it? What can it do? Usace Seventh Remote Sensing Symposium, Portland, OR.

Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L., 2020. SOLO: Segmenting Objects by Locations. ECCV 649–665.

Wambugu, N., Chen, Y., Xiao, Z., Tan, K., Wei, M., Liu, X., Li, J., 2021. Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review. Int. J. Appl. Earth Obs. 105, 102603. https://doi.org/10.1016/j.jag.2021.102603.

Windrim, L., Melkumyan, A., Murphy, R.J., Chlingaryan, A., Ramakrishnan, R., 2018. Pretraining for Hyperspectral Convolutional Neural Network Classification. IEEE Trans. Geosci. Remote Sens. 56, 2798–2810. https://doi.org/10.1109/TGRS.2017.2783886.

Xu, Y., Gong, J., Huang, X., Hu, X., Li, J., Li, Q., Peng, M., 2022. Luojia-HSSR: A high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet. Geo-spatial Inform. Sci. https://doi.org/10.1080/10095020.2022.2070555.

Xue, Z., Zhang, M., Liu, Y., Du, P., 2021. Attention-Based Second-Order Pooling Network for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 59 (11), 9600–9615. https://doi.org/10.1109/TGRS.2020.3048128.

Zhang, X., Wang, Y., Zhang, N., Xu, D., Luo, H., Chen, B., Ben, G., 2021. Spectral-Spatial Fractal Residual Convolutional Neural Network With Data Balance Augmentation for Hyperspectral Classification. IEEE Trans. Geosci. Rem. Sens. 59 (12), 10473–10487. https://doi.org/10.1109/TGRS.2020.3046840.

Zhao, Z., Luo, Z., Li, J., Chen, C., Piao, Y., 2020. When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework. Remote Sens. 12, 3276. https://doi.org/10.3390/rs12203276.

Zhu, K., Chen, Y., Ghamisi, P., Jia, X., Benediktsson, J., 2019. Deep Convolutional Capsule Network for Hyperspectral Image Spectral and Spectral-Spatial Classification. Remote Sens. 11, 223. https://doi.org/10.3390/rs11030223.

Zhu, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2018. Generative Adversarial Networks for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 56, 5046–5063. https://doi.org/10.1109/TGRS.2018.2805286.

Zhu, L., She, Q., Li, D., Lu, Y., Kang, X., Hu, J., Wang, C., 2021. Unifying Nonlocal Blocks for Neural Networks. ICCV 12292–12301. https://doi.org/10.1109/ICCV48922.2021.01207.

Zou, L., Zhu, X., Wu, C., Liu, Y., Qu, L., 2020. Spectral-Spatial Exploration for Hyperspectral Image Classification via the Fusion of Fully Convolutional Networks. IEEE. J. Sel. Topics Appl. Earth Observ. 13, 659–674. https://doi.org/10.1109/JSTARS.2020.2968179.