*remote sensing*

MDPI

*Article*

# Learning General-Purpose Representations for Cross-Domain Hyperspectral Images Classification with Small Samples

**Kuiliang Gao, Anzhu Yu, Xiong You *, Chunping Qiu, Bing Liu and Wenyue Guo**

Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450046, China

* Correspondence: youarexiong@163.com

**Abstract:** Cross-domain classification with small samples is a more challenging and realistic experimental setup. Until now, few studies have focused on the problem of small-sample cross-domain classification between completely different hyperspectral images (HSIs) since they possess different land cover types and statistical characteristics. To this end, this paper proposes a general-purpose representation learning method for cross-domain HSI classification, aiming to enable the model to learn more general-purpose deep representations that can quickly adapt to different target domains with small samples. The core of this method is to propose a novel three-level distillation strategy to transfer knowledge from multiple models well-trained on source HSIs into a single distilled model at the channel-, feature- and logit-level simultaneously. The learned representations can be further fine-tuned with small samples and quickly adapt to new target HSIs and previously unseen classes. Specifically, to transfer and fuse knowledge from multiple-source domains into a single model simultaneously and solve the inconsistency of the number of bands in different HSIs, an extensible multi-task model, including the channel transformation module, the feature extraction module and the linear classification module, is designed. Only the feature extraction module is shared across different HSIs, while the other two modules are domain-specific. Furthermore, the typical episode-based learning strategy of the metric-based meta-learning is adopted in the whole learning process to further improve the generalization ability and data efficiency. Extensive experiments are conducted on six source HSIs and four target HSIs, and the results demonstrate that the proposed method outperforms the existing advanced methods in cross-domain HSI classification with small samples.

**Keywords:** cross-domain hyperspectral image classification; small samples; general-purpose representations; knowledge distillation; multi-task learning; meta-learning



**Citation:** Gao, K.; Yu, A.; You, X.; Qiu, C.; Liu, B.; Guo, W. Learning General-Purpose Representations for Cross-Domain Hyperspectral Images Classification with Small Samples. *Remote Sens.* **2023**, *15*, 1080. https://doi.org/10.3390/rs15041080

Academic Editors: Chein-I Chang, Shengwei Zhong and Shuhan Chen

Received: 15 January 2023
Revised: 10 February 2023
Accepted: 12 February 2023
Published: 16 February 2023



## 1. Introduction

Hyperspectral imaging, one of the major advances in remote sensing technologies, can simultaneously obtain rich spectral and spatial information and express them in a unified three-dimensional data cube [1]. Hyperspectral image (HSI) classification, converting three-dimensional cubes into simple classification maps, has attracted extensive attention, and its direct products have been widely applied in many fields, such as mineral recognition [2], target detection [3] and fine agriculture [4].

In recent years, with the wide application of deep learning in remote sensing, deep model-based classification methods have gradually become a research hotspot worldwide [5,6], which can automatically learn deep features beneficial to target tasks in a data-driven manner, thus obtaining better and more stable results. Among them, a convolutional neural network (CNN) is one of the most mainstream models, and its unique convolution operation can directly process grid data and effectively utilize high-level features in hyperspectral cubes [7]. In addition, deep models such as capsule network (CN) [8], recurrent neural network (RNN) [9] and stack autoencoder (SAE) [10] have also been applied to HSI classification. Undoubtedly, with the introduction of various deep models

and the adoption of advanced learning methods, the accuracy of HSI classification with sufficient samples has been constantly updated [11–13]. However, the existing classification methods based on deep models still have the following two main shortcomings:

(1) The whole classification process of the deep models is confined to a single hyperspectral domain, and the models cannot utilize the valuable information and knowledge contained in related HSIs. Therefore, the utilization rate of the models is low, and the generalization ability between different HSIs is poor.
(2) The performance of deep models deteriorates rapidly with the decrease in the number of labeled samples. Under the condition of small samples where only a few labeled samples can be used for training (e.g., 5 samples per class), the models cannot obtain satisfactory classification results.

To this end, cross-domain HSI classification with small samples has begun to attract the attention of many researchers. It aims to make the deep models learn more abundant and relevant features through pre-training on a large number of source HSIs so that the models can better generalize to new tasks and obtain satisfactory classification results with few labeled samples when meeting target HSIs [14,15]. The source HSIs used for pre-training are completely different from the target HSIs in terms of land cover types, spectral ranges, resolutions and so on, so they belong to different hyperspectral domains [16]. Recently, there have been studies on cross-data and cross-scene classification, which mainly construct cross-classification scenarios by recombining bands of original data [17–19] or selecting pairs of HSIs with the same classes obtained by the same sensor [20–22]. Therefore, they do not strictly fall under small-sample cross-domain classification, which is a more challenging and realistic experimental setting.

To our knowledge, few studies so far have focused on the problem of small-sample cross-domain classification between completely different HSIs. In the few related studies, Liu et al. [14] and Gao et al. [15] preliminarily explore this problem based on the prototype network and relation network, respectively. Lee et al. also analyze the performance of the pre-trained CNN on HSI cross-domain classification [16]. However, there are some non-negligible shortcomings in the above methods, such as the insufficient use of knowledge and information in source domains, the loss of spatial–spectral features caused by dimensionality reduction and the dissatisfactory classification results that still need to be improved. To this end, a general-purpose representation learning method called GPRL is proposed in this paper, aiming to further improve the performance of cross-domain HSI classification with small samples. The core of the proposed method is to learn more general-purpose deep representations from multiple hyperspectral domains by distilling knowledge from multiple single-task models well-trained on source HSIs into a single distilled model at three different levels simultaneously. The learned general-purpose representations can be further fine-tuned with small samples for new target HSIs and previously unseen classes, achieving better classification performance. In addition, an extensible multi-task model that can adapt to any number of spectral bands is designed and the episode-based learning strategy is introduced to further improve the data efficiency for small samples while making full use of the spatial–spectral information in HSIs. The main contributions can be summarized as follows.

(1) A general-purpose representation learning method is proposed for cross-domain HSI classification with small samples, and extensive experiments demonstrate that the proposed method outperforms the existing advanced methods.
(2) A novel three-level distillation strategy is proposed to improve the effectiveness of knowledge transfer from multiple-source domains to a single distilled model through simultaneous distillation at the channel-, feature- and logit-level. To the best of our knowledge, this is the first application of knowledge distillation in cross-domain HSI classification.
(3) To distill knowledge from multiple-source domains into a single model simultaneously, a multi-task model, including the channel transformation module, the feature extraction module and the linear classification module, is designed. The channel

transformation module can enable HSIs with different bands to participate in cross-domain knowledge learning, effectively improving the expansibility of the model and avoiding the loss of spatial–spectral features caused by dimensionality reduction.

(4) The episode-based learning strategy is adopted, and the designed model is trained and fine-tuned, referring to the typical metric-based meta-learning process to further improve its generalization ability between different HSIs and data efficiency for small samples.

## 2. Related Work

### *2.1. Hyperspectral Images Classification*

Since Chen et al. applied the SAE model to HSI classification [23], various classification methods based on deep models have mushroomed and continuously improved classification performance [24–26]. Considering the characteristics of hyperspectral data, the existing advanced methods seek to obtain more accurate classification results by exploiting the deep spatial–spectral features of HSIs. For example, Liu et al. designed a supervised deep feature extraction method based on metric learning, which can effectively improve the separability between heterogeneous samples [27]. Gao et al. attempted to extract class-level features by embedding the dynamic routing mechanism into a deep residual network, effectively improving classification performance [28]. Xue et al. designed a multiscale deep-learning network with self-calibrated convolutions and self-attention modules to jointly utilize abstract features at different scales [29]. Recently, the transformer model has been introduced due to its excellent performance in many computer vision tasks. Tan et al. propose to model the patch- and pixel-level features by constructing a deep transformer network, obtaining better results than conventional CNN models [30]. In addition, the performance of the capsule network on spatial–spectral feature extraction has been widely explored [8,31].

Meanwhile, advanced learning methods such as semi-supervised learning, transfer learning and contrastive learning are widely applied, successfully reducing the excessive dependence on a large number of training samples. Graph convolutional network (GCN), one of the representative semi-supervised models, can effectively utilize the potential features in unlabeled samples through graph construction [32,33]. Transfer learning initializes deep models with transferable parameters learned from relevant tasks, making it easier for them to find the optimal solutions during the training process on target HSIs [34,35]. Contrastive learning can construct self-supervised learning tasks through different data augmentations on unlabeled samples, which can extract more discriminative deep features [36,37]. Many studies have shown that the above learning methods can effectively improve HSI classification accuracy.

However, most of the existing studies limit the whole classification process to a single hyperspectral domain and fail to achieve satisfactory performance when only a few labeled samples are available. Few studies have focused on cross-domain HSI classification with small samples, which is a more challenging and realistic experimental setting.

### *2.2. Knowledge Distillation*

The purpose of knowledge distillation is to transfer the knowledge learned by a cumbersome teacher network to a compact student network [38,39]. Consequently, a compact network with stronger feature learning ability can be obtained. The initial knowledge distillation method proposes to add constraints on the logit outputs according to the soft targets and trains a student network in conjunction with hard labels [40]. Subsequently, a series of improved methods distill knowledge by selecting different transfer mediums, such as intermediate features [41], attention maps [42] or the flow of solution procedure matrix [43], effectively improving the performance of the distilled network. In multi-task learning, Li et al. distill the knowledge from multiple single-task networks to a single multi-task network through task-specific adapters [44], and our method borrows this idea.

There have been studies on applying knowledge distillation to HSI classification. For example, Shi et al. design an explainable scale distillation method, integrating spatial features within multiple scales into a compact network [45]. Yue et al. propose a self-supervised learning method with adaptive distillation to train a deep neural network with extensive unlabeled samples [46]. However, the existing methods combined with knowledge distillation all conduct training and prediction with sufficient labeled samples from a single hyperspectral domain without the exploration of cross-domain HSI classification with small samples. To the best of our knowledge, this paper is the first study applying knowledge distillation to cross-domain HSI classification with small samples.

### *2.3. Meta-Learning*

Meta-learning, known as learning how to learn, is a potential paradigm that learns features from vast tasks and generalizes them to new unseen tasks with few labeled samples. Generally, optimization-based and metric-based meta-learning methods are two mainstream methods. The former aims to learn a deep model that can adapt to new tasks with a small number of iterations and labeled samples. Model-agnostic meta-learning (MAML) [47] is one of the representative algorithms. The core idea of the latter is to map the raw data into a deep feature space, cluster together the samples belonging to the same class and separate the samples belonging to different classes. Typical methods include a prototypical network [48], relation network [49] and induction network [50].

In the field of remote sensing, the two meta-learning methods above have been introduced into HSI classification. Gao et al. preliminarily analyze the generalization performance of MAML on different classes in the same scene [51]. Liu et al. [14] and Ma et al. [52] explore the effectiveness of metric-based meta-learning methods in HSI classification based on a prototypical network and relation network, respectively. However, the existing meta-learning methods mechanically train a single network according to the "pre-training + fine-tuning" mode, failing to fully transfer and fuse the information and knowledge from source HSIs. Different from these methods, our method distills knowledge from multiple single-task models to a single multi-task model through the proposed three-level distillation strategy, which can better integrate knowledge and features from different source HSIs and learn more general-purpose representations.

### *2.4. Cross-Domain Classification with Small Samples*

Recently, in the field of remote sensing, more and more attention has been paid to the problem of cross-domain classification with small samples, which aims to utilize the knowledge learned from source domains to guide the small sample classification in target domains [53]. For example, Bi et al. design a contrastive domain adaptation-based sparse SAR target classification method to solve the problem of insufficient samples of target domains [54]. Rostami et al. present a novel transfer learning framework, which can learn a shared invariant embedding space for small sample classifications [55]. Lasloum et al. and Shi et al., respectively, explore the performance of semi-supervised domain adaptation in HSI target detection and remote sensing scene classification, achieving better results in target domains [56,57]. However, most of the methods conduct cross-domain learning through adversarial-based domain adaptation training, ignoring the integration and utilization of different knowledge and information from multiple-source domains. In contrast, based on the proposed three-level distillation strategy and the designed multi-task model, the proposed method can effectively distill knowledge from different source domains into a deep model so that it can learn the more general-purpose representations, which can quickly adapt to new classification tasks.

## 3. Methodologies

### *3.1. Workflow*

The proposed GPRL method aims to make the model learn more general-purpose deep representations from multiple hyperspectral domains and obtain better results in cross-domain HSI classification with small samples. The workflow can be divided into the following three steps:

(1) Pre-training on source HSIs (illustrated in Figure 1): Multiple different source HSIs are collected in advance, and multiple single-task models are fully trained on different source HSIs, respectively. Consequently, each model can acquire important information and knowledge from the corresponding hyperspectral domain.

(2) Learning general-purpose representations (illustrated in Figure 2): All the parameters of the multiple single-task models well-trained on source HSIs are frozen, and the randomly initialized multi-task model is fully trained with the three-level distillation strategy and episode-based learning strategy to learn general-purpose representations from multiple different hyperspectral domains. In the designed multi-task model, the single feature extraction module is shared across different HSIs, while multiple channel transformation modules and linear classification modules are domain-specific (illustrated in the bottom half of Figure 2).

(3) Fast adaption to target HSIs (illustrated in Figure 3): For each target HSI, only the feature extraction module well-trained in the previous step is inherited, while the channel transform module and the linear classification module are randomly initialized. Then, the whole model is fine-tuned using a few labeled samples to quickly adapt to the new hyperspectral domain.

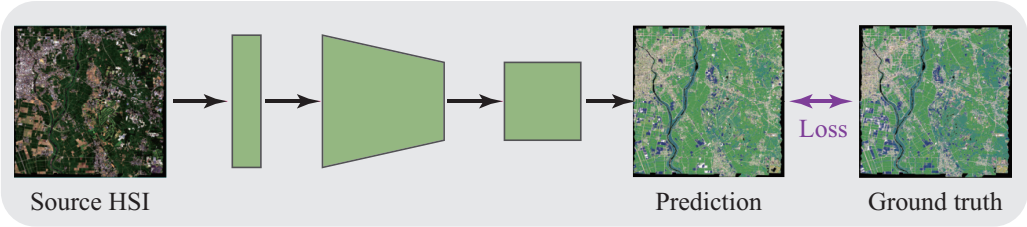



**Figure 1.** Schematic of pre-training on source HSIs (the Chikusei dataset is presented as an example).

In the first step, the number of source HSIs determines the number of pre-trained single-task models. Obviously, the second step is the core of the proposed method, and the characteristics of the learned deep representations directly determine the performance of cross-domain small-sample classification on target HSIs. In the third step, only a few labeled samples (for example, 5 labeled samples per class) from the target HSI are used to fine-tune the whole model. In the remainder of this section, we will focus on the second step as well as the designed loss function, the constructed multi-task model and the adopted learning strategy. Before we begin, the mathematical notations for several important parts in Figure 2 are given: the channel transformation module, the feature extraction module and the linear classification module are denoted as $t_\varphi$, $f_\phi$ and $h_\psi$, respectively. The adopter is denoted as $A_\theta$, and the input data are denoted as $x$.

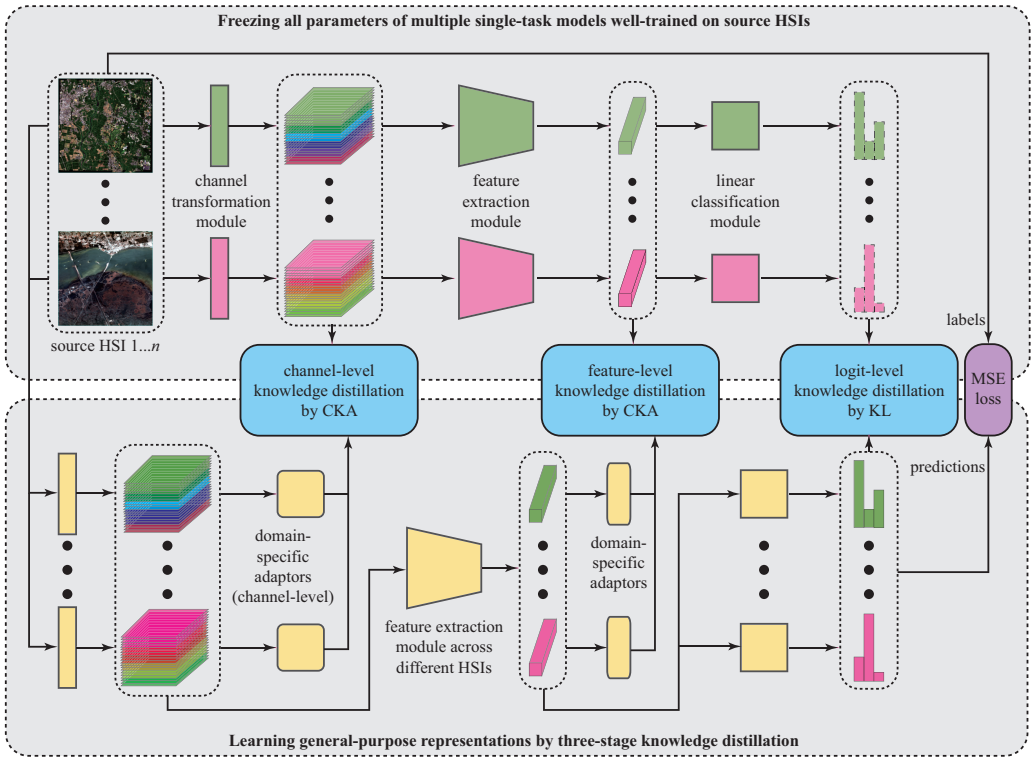



**Figure 2.** Schematic of the proposed general-purpose representation learning method. We first freeze all parameters of multiple single-task models well-trained on source HSIs and then attempt to make the designed multi-task model learn general-purpose representations from multiple different hyperspectral domains through the proposed three-level knowledge distillation strategy. During knowledge distillation, in addition to the three designed modules, two types of domain-specific adaptors are inserted to align the features generated by the single-task models and the multi-task model. In the channel- and feature-level knowledge distillation, a centered kernel alignment (CKA) similarity index is adopted, and in the logit-level knowledge distillation, the Kullback–Leibler (KL) divergence is adopted. For the predictions and true labels, the mean square error (MSE) function is adopted for calculating loss.

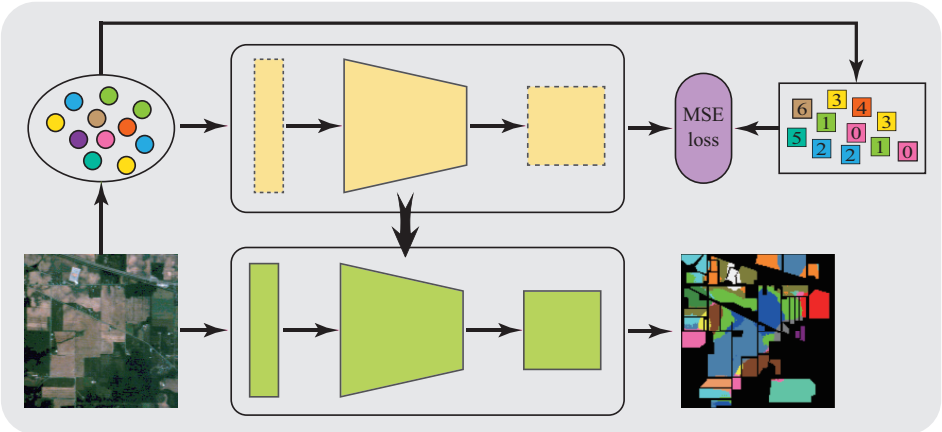



**Figure 3.** Schematic of fast adaption to target HSIs. The whole model containing the inherited feature extraction module and two other initialized modules is first trained on a few labeled samples to quickly adapt to the new hyperspectral domain. Then, the trained model performs label prediction on the target HSI.

### *3.2. Three-Level Knowledge Distillation*

We propose learning general-purpose deep representations from multiple hyperspectral domains. To this end, a three-level knowledge distillation method is proposed to distill

a single multi-task model from multiple single-task models. Consequently, the learned deep representations could automatically contain the required information from several relevant domains and are more general-purpose for the downstream target tasks.

As shown in Figure 2, based on the three designed modules, the three levels of channel, feature and logit can be divided from the whole process. Knowledge distillation is performed simultaneously at the three different levels by minimizing the distance between the logit outputs (logit-level), the distance between the intermediate features after channel transformation (channel-level), and also the distance between the learned deep representations (feature-level). In most existing research, the Kullback-Leibler (KL) divergence is widely used to calculate the distance between the logit outputs of student networks and teacher networks due to its excellent performance and efficient computation process. Therefore, in our method, the KL divergence is also adopted to calculate the logit-level distance between the probability outputs of the multi-task model $p^m$ and corresponding single-task models $p^s$, which can be formalized as follows:

$$KL(p^m, p^s) = \Sigma p^m \cdot log(\frac{p^m}{p^s}), \tag{1}$$

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}, \tag{2}$$

where $z$ denotes the logit outputs, and $T$ denotes the distillation temperature.

Different from logit outputs, which reflect probability distributions directly and simply, the intermediate features after channel transformation and learned deep representations possess higher dimensions and are more complicated, so more procedures are required to align them. On the one hand, referring to [44], domain-specific adaptors are inserted to map the features generated by the multi-task model into domain-specific vectors and optimized jointly, along with the parameters of the designed model. On the other hand, the centered kernel alignment (CKA) similarity index with radial basis function as the kernel is adopted to calculate the distance in high-dimensional spaces since it is capable of meaningful non-linear similarities.

(1) Domain-specific adaptors: The large difference between source HSIs means that the outputs of multiple single-task models can vary significantly, and the outputs of the multi-task model cannot match all of them simultaneously, whether at the channel level or feature level. Therefore, domain-specific adaptors are inserted (Figure 2) to map the outputs of the multi-task model into domain-specific vectors, which can be expressed as $A_\theta : \boldsymbol{R}^{C\times H\times W} \rightarrow \hat{\boldsymbol{R}}^{C\times H\times W}$, where $C$, $H$ and $W$ are the number of channels, height and width, and $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$ denote the inputs and outputs, respectively. In practice, the adaptor is instantiated with a $1\times1$ convolution layer with $C$ kernels, and the difference between adaptors at the channel level and feature- evel is the value of $C$.

(2) The CKA similarity index: Aligning features learned from substantially diverse domains requires a better and more complex distance function to model non-linear correlations between them. The calculation of CKA similarity can be divided into two steps. For the features $\boldsymbol{M}$ and $\boldsymbol{S}$ generated by the adopters of multi-task and single-task models, respectively, the radial basis function matrices $\boldsymbol{P}$ and $\boldsymbol{T}$ are first computed. Then, the similarity between them is measured as:

$$CKA(\boldsymbol{P}, \boldsymbol{T}) = tr(\boldsymbol{PHTH})/\sqrt{tr(\boldsymbol{PHPH})tr(\boldsymbol{THTH})}, \tag{3}$$

where $tr$ denotes the trace of a matrix, and $\boldsymbol{H}$ denotes a centering matrix. In the training process, the alignment between high-dimensional features is achieved by minimizing $1 - CKA(\boldsymbol{P}, \boldsymbol{T})$.

In each training iteration, distillations at three different levels are performed simultaneously. Through the simultaneous transfer of information and knowledge from multiple hyperspectral domains at different levels, the distilled model can learn more general-purpose knowledge and thus acquires better generalization ability for downstream target tasks.

### *3.3. Loss Function*

As shown in Figure 2, the loss function of the whole model consists of two main parts: distillation loss and prediction loss. The distillation loss can be further divided into three losses at different levels:

$$L_\tau^c = 1 - CKA(A_{\theta_\tau}^c \circ t_{\varphi_\tau}(x), t_{\varphi_\tau^*}(x)), \tag{4}$$

$$L_\tau^f = 1 - CKA(A_{\theta_\tau} \circ f_\phi \circ t_{\varphi_\tau}(x), f_{\phi_\tau^*} \circ t_{\varphi_\tau^*}(x)), \tag{5}$$

$$L_\tau^l = KL(h_{\psi_\tau} \circ f_\phi \circ t_{\varphi_\tau}(x), h_{\psi_\tau^*} \circ f_{\phi_\tau^*} \circ t_{\varphi_\tau^*}(x)), \tag{6}$$

where $L^c$, $L^f$ and $L^l$ represent the distillation losses at the channel-, feature-, and logit-level, respectively, and $\tau$ represents index source HSIs. The symbol $*$ is the identity of single-task models, and $A_\theta^c$ and $A_\theta$ denote the adopters at the channel- and feature-level, respectively. Note that $f_\phi$ is shared in the multi-task model, while $t_{\varphi_\tau}$ and $h_{\psi_\tau}$ depend on the particular source HSI.

In addition to the above distillation losses, the loss between the predictions and the true labels is calculated according to the mean square error (MSE) function:

$$L_\tau^{MSE} = \frac{\sum_{i=1}^n (h_{\psi_\tau} \circ f_\phi \circ t_{\varphi_\tau}(x) - y)^2}{n}, \tag{7}$$

where $y$ denotes the true labels, and $n$ is the dimension of class vectors and actually equals the number of classes in HSI. Now, the total loss can be given as follows:

$$L = \sum_{\tau=1}^{K} (\lambda(L_\tau^c + L_\tau^f + L_\tau^l) + L_\tau^{MSE}), \tag{8}$$

where $K$ is the number of source HSIs, and $\lambda$ is used to adjust the weight of distillation losses. In the training process, the multi-task model is optimized based on both the distillation loss and the prediction loss. The total loss function prompts the model to align with the outputs of multiple single-task models at different levels while enabling the model to make correct predictions, which is conducive to the fusion of knowledge from different hyperspectral domains and learning more general-purpose representations.

### *3.4. Extensible Multi-Task Model*

To transfer and fuse knowledge from multiple single-task models simultaneously into a single distilled model and better adapt to the characteristics of different HSIs, a novel multi-task model is designed. Figure 4 is actually a detailed expansion of the bottom half of Figure 2. As we can see, the designed model mainly consists of a shared feature extraction module, multiple domain-specific channel transformation modules and linear classification modules. The designed model can receive multiple-source HSIs at the same time and perform knowledge distillation and classification prediction on multiple domains, so it conforms to the paradigm of multi-task learning.

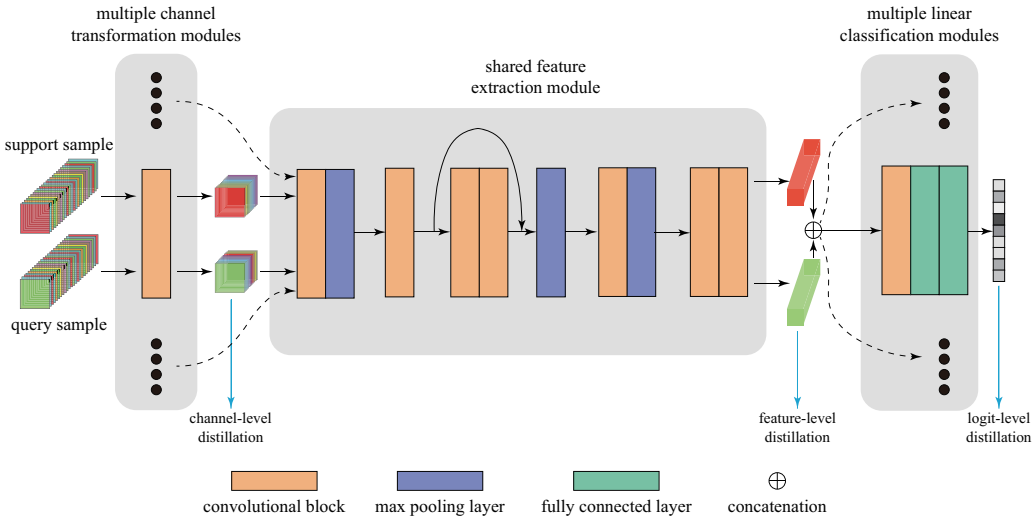



**Figure 4.** Schematic of the designed multi-task model for general-purpose representation learning (actually a detailed expansion of the bottom half of Figure 2). The designed model can receive multiple-source HSIs simultaneously and perform knowledge distillation and label prediction on multiple-source domains; therefore, it conforms to the paradigm of multi-task learning.

The channel transformation module is responsible for transforming input HSIs with any number of bands into cube data with $N_c$ bands without changing the space size. For multiple-source HSIs, the value of $N_c$ is artificially fixed and consistent. This module skillfully deals with the inconsistency of the number of spectral bands in different HSIs, so that HSIs with any number of bands can be used as source data sets, effectively improving the expansibility of the model and avoiding the loss of spatial–spectral features caused by dimensionality reduction. The feature extraction module is equivalent to a complex non-linear function, mapping input cubes into feature vectors containing rich spatial–spectral information. Undoubtedly, the performance of the feature extraction module determines the quality of the learned deep representations. The linear classification module, essentially a simple linear classifier, is used to assign a unique class label to each input vector. Besides the shared feature extraction module, the number of the other two modules is determined according to the number of source HSIs, so the scale and structure of the model are extensible.

Next, a branch of the multi-task model is taken as an example to describe the network structure in detail since the structure of channel transformation modules and linear classification modules corresponding to different HSIs are exactly the same. The cubes around the center pixels are used as the input support or query samples to make full use of the spatial–spectral information in HSIs. The channel transformation module, actually a convolutional block containing a two-dimensional convolution layer, a batch normalization layer and a ReLU activation function, first compresses the channel dimension of input samples to the fixed value $N_c$. Specifically, the size of convolutional kernels is $1 \times 1$. The feature extraction module is composed of convolutional blocks, pooling layers and residual connections. The pooling layers are embedded between convolutional blocks to gradually reduce the space size of input cubes while extracting deep features, and the residual connection promotes the joint utilization of features at different layers. Specifically, the size of convolution kernels is set as $3 \times 3$, the number of convolution kernels increases layer by layer according to $[128, 256, 512, 1024]$, and the size of kernels in max pooling layers is set as $2 \times 2$. The linear classification module is composed of a convolutional block and two fully connected layers and maps the concatenation vectors of support samples and query samples into class vectors. The size of the convolution kernel is $1 \times 1$, and the dimensions of the two fully connected layers are set as 512 and 128. In addition, the dropout operation and the sigmoid activation are added to the outputs of the two fully connected layers, respectively.

### *3.5. Episode-Based Learning Strategy*

Many studies have shown that, compared with the conventional batch-based training strategy, the episode-based learning strategy can effectively improve the data efficiency for small samples, that is, when generalized to new data sets, deep models can learn quickly and efficiently with few labeled samples [14,47,49]. Therefore, the episode-based learning strategy of the typical metric-based meta-learning is employed in the whole learning process.

As shown in Figure 5, an episode consists of a support set and a query set. Given a training data set, a large number of different episodes are generated by random sampling, so different episodes usually contain different classes. Whereas in one episode, the support set and the query set have exactly the same classes. However, it is noted that each sample in one episode is different from the others. In the training process, the support samples are used as supervised information to optimize the prediction results of deep models on the query samples. Specifically, in the metric-based meta-learning process, the class of the query sample is determined by measuring the similarity distances between the query sample and the support samples. Furthermore, in one episode, the number of support samples is often fewer than the number of query samples to simulate the small-sample setting. Formally, an episode can be described by three keywords: way, shot and query. The first keyword way represents the number of classes in an episode, and the keywords shot and query represent the number of samples per class in the support set and query set of the episode, respectively. For example, each episode in Figure 5 can be denoted as (4-way, 1-shot, 2-query). In the experiments, the value of way in episodes is set equal to the number of classes in the given HSIs so that the multi-task model can learn rich knowledge from the multiple hyperspectral domain, simultaneously. As for the two hyperparameters shot and query, Section 4.5.1 gives a detailed analysis.

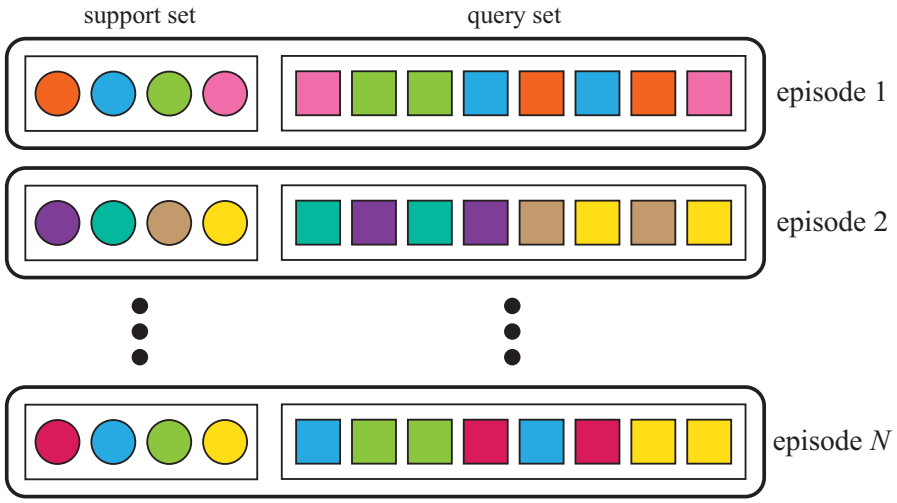



**Figure 5.** Schematic of episodes (4-way 1-shot 2-query) in the typical meta-learning process. Different colors represent different classes.

## 4. Experimental Results and Analysis

### *4.1. Data Sets*

To study the problem of small-sample cross-domain classification between completely different HSIs, 10 widely used HSIs were divided into source HSIs and target HSIs referring to existing researches [14,15,25]. Specifically, the six source HSIs include HanChuan (HC), Cuprite (CU), Houston 2013 (HS13), Botswana (BO), Kennedy Space Center (KSC) and Chikusei (CH), while the four target HSIs are University of Pavia (UP), Pavia Center (PC), Salinas (SA) and Indian Pines (IP), respectively. The details of these HSIs are listed in Table 1.

**Table 1.** Details of source HSIs and target HSIs. HanChuan (HC), Cuprite (CU), Houston 2013 (HS13), Botswana (BO), Kennedy Space Center (KSC), Chikusei (CH), University of Pavia (UP), Pavia Center (PC), Salinas (SA), Indian Pines (IP), ground sample distance (GSD)(m), spatial size (pixel), Spectral range (nm), airborne visible infrared imaging spectrometer (AVIRIS), reflective optics system imaging spectrometer (ROSIS).

| | **Source HSIs** | | | | | | **Target HSIs** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **HC** | **CU** | **HS13** | **BO** | **KSC** | **CH** | **UP** | **PC** | **SA** | **IP** |
| Spatial size | 1217 × 303 | 614 × 512 | 349 × 1905 | 1476 × 256 | 512 × 614 | 2517 × 2335 | 610 × 340 | 1096 × 715 | 512 × 217 | 145 × 145 |
| Spectral range | 400–1000 | 370–2480 | 380–1050 | 400–2500 | 400–2500 | 363–1018 | 430–860 | 430–860 | 400–2500 | 400–2500 |
| No. of bands | 274 | 190 | 144 | 145 | 176 | 128 | 103 | 102 | 204 | 200 |
| GSD | 0.109 | 20 | 2.5 | 30 | 18 | 2.5 | 1.3 | 1.3 | 3.7 | 20 |
| Sensor type | Headwall Nano-Hyperspec | AVIRIS | ITRES-CASI 1500 | EO-1 | AVIRIS | Hyperspec -VNIR-C | ROSIS | ROSIS | AVIRIS | AVIRIS |
| Areas | HanChuan | Cuprite | Houston | Botswana | Florida | Chikusei | Pavia | Pavia | California | Indiana |
| No. of classes | 16 | 8 | 15 | 14 | 13 | 19 | 9 | 9 | 16 | 16 |
| Total labeled samples | 257,530 | 3837 | 15,029 | 3248 | 5211 | 77,592 | 42,776 | 148,152 | 54,129 | 10,249 |
| Labeled samples for training | 200 per class | | | | | | 5 per class | | | |

#### 4.1.1. Source HSIs

The six source HSIs are captured by different sensors, respectively, and have completely different land cover types, ground sample distances, spectral ranges and band amounts. On the one hand, diverse source HSIs can effectively improve the richness and diversity of samples for representation learning, and on the other hand, inevitably increase the challenge for knowledge transfer and distillation. To keep the balance between the training difficulty and learning effect, 200 samples per class are randomly selected from each source HSI as representative data in this domain. Meanwhile, the 28 × 28 cubes in the neighborhood of pixels are selected as input samples to make full use of the spatial–spectral information in HSIs. Therefore, for the source HSI with $N_b$ bands and $M$ classes, the actual size of data involved in the training process is ($M$, 200, $N_b$, 28, 28).

#### 4.1.2. Target HSIs

Compared with the six source HSIs, the four target HSIs also have completely different classes, ground sample distances, spectral ranges and band amounts. Therefore, using the model distilled from multiple-source HSIs to classify target HSIs is the typical process of cross-domain HSI classification. For each target HSI, only five labeled samples per class are randomly selected for model training, and the remaining samples are used to evaluate the performance of cross-domain HSI classification. The dimensions of each sample are also ($N_b$, 28, 28). In addition, it should be noted that the four different target HSIs also provide a variety of scenes to fully validate the performance of the proposed method.

### *4.2. Environment and Settings*

All the results were generated on a computer equipped with an Nvidia A100 GPU and an Intel(R) Xeon(R) Gold 6152 CPU. All the algorithms and programs were developed by Python 3.9 and machine learning libraries, such as Pytorch, sklearn and numpy.

During pre-training on source HSIs, the learning rate and iteration times were set to 0.0001 and 1000, respectively, for each single-task model. The main structure of the single-task model is exactly the same as that of the multi-task model, and the difference between the two models lies in the number of modules and the followed paradigms. During knowledge distillation for learning general-purpose representations, the temperature $T$ is set to 4, and the value of $N_c$ is set to 100. In each episode, the number of classes depends on the corresponding HSIs, and the number of support samples and query samples are set to 5 and 15. The learning rate and iteration times are set to 0.0001 and 1000, respectively, and the Adam optimization algorithm is used for parameter updating. In addition, the probability value of $dropout$ is set to 50%. During fast adaption to target HSIs, the learning rate and

iteration times are set to 0.0001 and 500, respectively, and the $M_t$-way 2-shot 3-query tasks are constructed ($M_t$ is 9 for UP and PC, $M_t$ is 16 for SA and IP). During the whole training process, the hyperparameter $\lambda$ is set to 0.8.

Consistent with other existing studies, the overall accuracy (OA), average accuracy (AA) and kappa coefficient were used to evaluate the classification results. Furthermore, all of the algorithms were run 10 times with the same random seeds, and the results are expressed as mean value and standard deviation to further improve the credibility and persuasiveness of experimental results.

### *4.3. Classification Results and Analysis*

To evaluate the classification performance of the proposed method, seven advanced methods were selected for comparison, including three cross-domain classification methods (DFSL+SVM [14], RN-FSC [15] and UM$^2$L [25]), two semi-supervised methods (EMP+TSVM [58] and EMP+GCN [32]), an advanced GAN-based method (3D-HyperGAMO [59]) and a classic contrastive learning method (Barlow Twins (BT) [60]). For the three cross-domain methods and the proposed method, according to existing research, the four data sets, including HS13, BO, KSC and CH, were selected as source HSIs for a fair comparison [14,15,25]. The classification results of different methods are listed in Table 2, from which several observations can be obtained.

(1) The traditional method (EMP+SVM) performs semi-supervised classification based on the extracted shallow EMP features and cannot make full use of the deep abstract features in HSIs, so its classification performance is significantly worse than that of other deep models.
(2) The accuracy and robustness of the classification results of EMP+GCN, 3D-HyperGAMO and BT are obviously better than that of EMP+SVM. According to the OA, EMP+GCN has better performance on the UP and SA data sets, while 3D-HyperGAMO and BT can achieve better classification results on the PC and IP data sets, respectively. EMP+GCN and 3D-HyperGAMO can utilize unlabeled samples and synthetic samples, respectively, to assist model training on the target domain, and BT can utilize the more discriminative features in target HSIs, thus effectively improving the classification results.
(3) The three cross-domain methods, DFSL+SVM, RN-FSC and UM$^2$L, can further improve the performance of cross-domain HSI classification with samples. By using a large number of samples in the source HSIs to pre-train the deep models, the models can obtain a better initialization state compared with training from scratch so as to obtain higher classification accuracy in the target domains with small samples.
(4) Obviously, the proposed method achieves the best classification results. For the four target HSIs, the OA of the proposed method is 3.52%, 0.56%, 2.47% and 0.62% higher than that of the second place, respectively, and the kappa coefficient of the proposed method is 4.78%, 0.80%, 2.73% and 0.66% higher than that of the second place, respectively. Compared with the other three cross-domain methods, on the one hand, the proposed method can learn more general-purpose representations from multiple-source domains with the three-level distillation strategy, and on the other hand, the proposed method trains the deep model based on the multi-task learning paradigm, which can better adapt to the characteristics of different target HSIs and make full use of the spatial–spectral information when only a few labeled samples are available.

Classification maps are often used as a qualitative measure to compare the classification results of different methods from the perspective of visualization. The classification maps of different methods on the four target HSIs are given in Figures 6–9. As we can see, compared with other methods, the proposed method can produce the classification maps closest to the real ground truths, where the noise and misclassification are effectively reduced. Taking the UP data set as an example (Figure 6), in the corresponding regions of classes Bare Soil and Bricks, the noise is significantly reduced, and the regional coherence

is significantly improved. In short, the comparison of classification maps again proves the effectiveness of the proposed method in cross-domain HSI classification with small samples.

**Table 2.** The classification results from different methods. SD denotes the standard deviation of 10 experimental results.

| HSI | Criteria | EMP+TSVM *Mean* ± *SD* | EMP+GCN *Mean* ± *SD* | 3D-HyperGAMO *Mean* ± *SD* | BT *Mean* ± *SD* | DFSL+SVM *Mean* ± *SD* | RN-FSC *Mean* ± *SD* | UM$^2$L *Mean* ± *SD* | GPRL *Mean* ± *SD* |
|---|---|---|---|---|---|---|---|---|---|
| UP | OA | 69.58 ± 7.55 | 75.23 ± 3.96 | 70.42 ± 4.91 | 70.03 ± 3.09 | 71.64 ± 6.07 | 77.16 ± 5.82 | 80.49 ± 3.64 | **84.01 ± 4.31** |
| | AA | 72.65 ± 3.70 | 75.12 ± 3.27 | 69.97 ± 3.60 | 69.63 ± 2.84 | 74.12 ± 4.65 | 72.60 ± 4.38 | 77.79 ± 3.05 | **82.32 ± 2.99** |
| | kappa | 61.83 ± 7.67 | 68.02 ± 4.19 | 62.59 ± 5.45 | 62.40 ± 3.51 | 64.41 ± 6.53 | 70.86 ± 6.66 | 74.70 ± 4.05 | **79.48 ± 5.14** |
| PC | OA | 92.83 ± 1.90 | 95.01 ± 0.83 | 95.27 ± 2.23 | 95.19 ± 2.37 | 95.54 ± 1.52 | 95.61 ± 1.08 | 94.43 ± 1.12 | **96.17 ± 0.74** |
| | AA | 83.01 ± 2.77 | 85.43 ± 1.68 | 86.61 ± 4.62 | 86.53 ± 4.85 | 87.27 ± 2.59 | 88.14 ± 1.82 | 84.89 ± 2.61 | **89.00 ± 2.30** |
| | kappa | 89.96 ± 2.57 | 92.98 ± 1.16 | 93.35 ± 3.08 | 93.20 ± 3.22 | 93.75 ± 2.04 | 93.79 ± 1.52 | 92.17 ± 1.56 | **94.59 ± 1.04** |
| SA | OA | 83.58 ± 1.65 | 85.93 ± 0.99 | 83.70 ± 4.40 | 83.70 ± 4.17 | 84.52 ± 3.32 | 86.37 ± 3.32 | 89.82 ± 4.18 | **92.29 ± 3.55** |
| | AA | 87.54 ± 0.79 | 89.87 ± 0.98 | 88.76 ± 2.59 | 86.62 ± 2.49 | 91.67 ± 0.84 | 89.75 ± 1.61 | 92.99 ± 2.23 | **95.15 ± 1.76** |
| | kappa | 81.78 ± 1.81 | 84.33 ± 1.10 | 81.97 ± 4.84 | 81.98 ± 4.59 | 82.90 ± 3.60 | 84.91 ± 3.65 | 88.72 ± 4.61 | **91.45 ± 3.92** |
| IP | OA | 55.09 ± 3.51 | 55.98 ± 2.93 | 57.02 ± 3.00 | 58.36 ± 2.39 | 60.18 ± 3.53 | 60.84 ± 3.15 | 71.65 ± 2.17 | **72.27 ± 2.61** |
| | AA | 55.19 ± 1.96 | 54.77 ± 1.89 | 55.48 ± 2.47 | 52.86 ± 1.96 | 59.41 ± 1.73 | 56.66 ± 4.65 | 64.60 ± 2.74 | **65.26 ± 2.56** |
| | kappa | 49.62 ± 3.80 | 50.65 ± 3.00 | 52.17 ± 3.30 | 54.37 ± 2.50 | 55.66 ± 3.72 | 56.52 ± 3.41 | 68.29 ± 2.35 | **68.95 ± 2.88** |

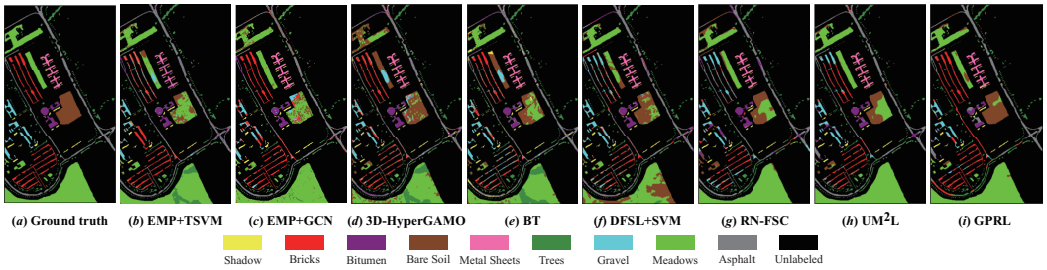



**Figure 6.** The classification maps of different methods on the UP data set.

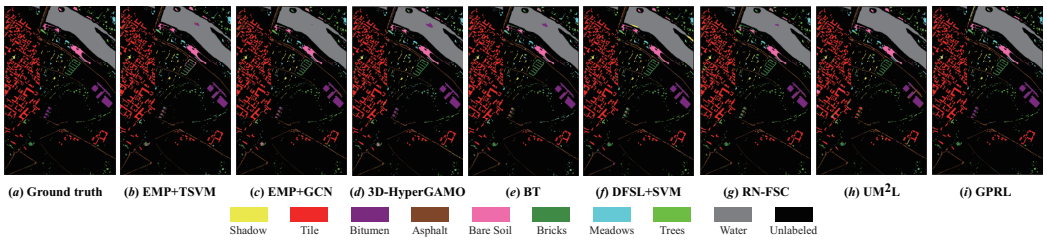



**Figure 7.** The classification maps of different methods on the PC data set.

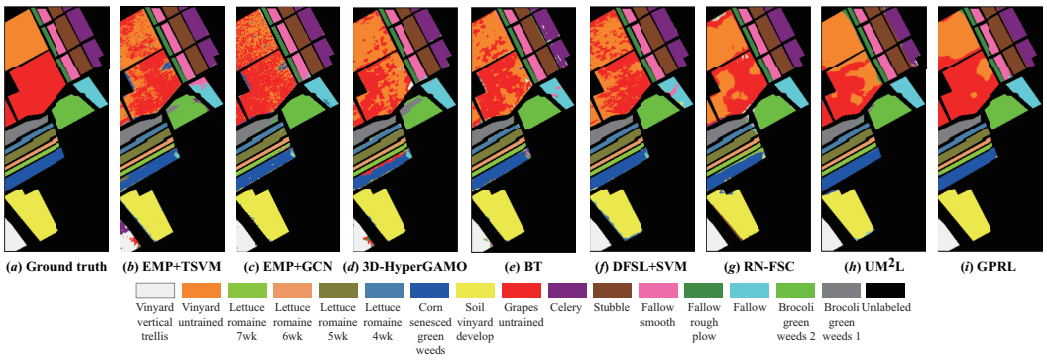



**Figure 8.** The classification maps of different methods on the SA data set.

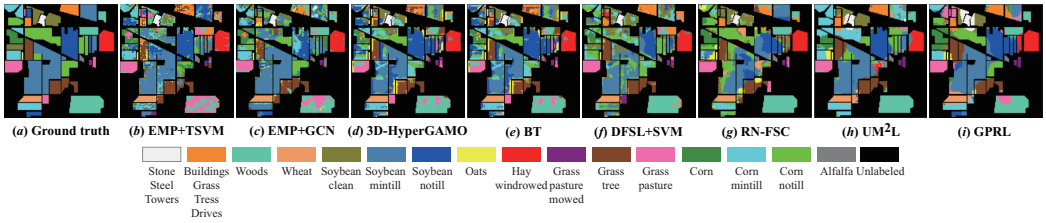



**Figure 9.** The classification maps of different methods on the IP data set.

### *4.4. General-Purpose Representations*

The proposed method focuses on learning general-purpose representations from multiple-source HSIs domains to achieve better classification results in different target domains with small samples. In this section, the learned general-purpose representations will be explained and analyzed in detail from the following three perspectives: classification accuracy, the number of required iterations, and feature separability.

#### 4.4.1. From the Perspective of Classification Accuracy

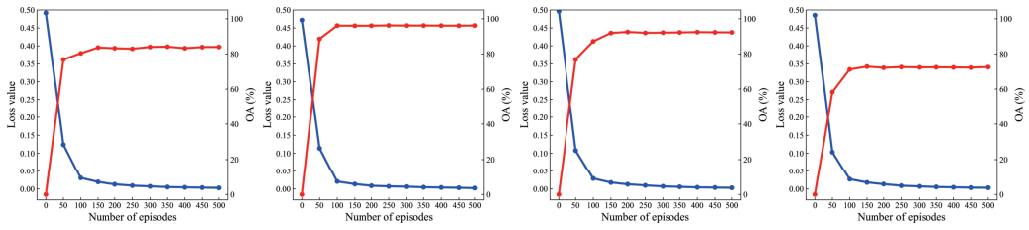
The general-purpose representations should be able to generalize to multiple different target domains and achieve higher classification accuracy with small samples. Table 3 compares the classification results of different variants of the proposed method, which is actually an ablation study on the two important parts: meta-training and knowledge distillation. In Table 3, the third column *Baseline* means that the randomly initialized model is optimized with five samples per class in target domains, while the fifth column includes the whole training and classification process. As we can see, from the third column to the fifth column, the classification accuracy and kappa coefficient gradually increase, which means that the proposed method with the combination of the meta-training process and the three-level knowledge distillation strategy can learn more general-purpose deep representations from source domains so as to achieve more accurate classification results in different target domains.

**Table 3.** The classification results of different variants of the proposed method. SD denotes the standard deviation of 10 experimental results.

| HSI | Criteria | Baseline | Baseline + Meta-Training | Baseline + Meta-Training + Knowledge Distillation |
|---|---|---|---|---|
| | | *Mean* ± *SD* | *Mean* ± *SD* | *Mean* ± *SD* |
| UP | OA | 78.73 ± 3.00 | 80.84 ± 3.88 | 84.01 ± 4.31 |
| | AA | 77.85 ± 2.86 | 78.48 ± 2.39 | 82.32 ± 2.99 |
| | kappa | 72.86 ± 3.36 | 75.56 ± 4.57 | 79.48 ± 5.14 |
| PC | OA | 94.54 ± 1.01 | 95.19 ± 0.77 | 96.17 ± 0.74 |
| | AA | 86.07 ± 2.35 | 86.37 ± 2.09 | 89.00 ± 2.30 |
| | kappa | 92.33 ± 1.39 | 93.21 ± 1.08 | 94.59 ± 1.04 |
| SA | OA | 89.45 ± 3.78 | 90.97 ± 4.07 | 92.29 ± 3.55 |
| | AA | 93.90 ± 1.73 | 94.18 ± 1.90 | 95.15 ± 1.76 |
| | kappa | 88.32 ± 4.16 | 90.00 ± 4.48 | 91.45 ± 3.92 |
| IP | OA | 67.29 ± 1.53 | 69.27 ± 4.85 | 72.27 ± 2.61 |
| | AA | 66.00 ± 3.99 | 62.42 ± 3.90 | 65.26 ± 2.56 |
| | kappa | 63.42 ± 1.65 | 65.82 ± 5.11 | 68.95 ± 2.88 |

#### 4.4.2. From the Perspective of Required Iterations

In this subsection, the number of required iterations during fast adaption is used to measure the adaptability of the learned deep representations to different target domains, considering that the more general-purpose the representations are, the fewer iterations they need to adapt to the new classes. Figure 10 shows the curves of loss value and OA when the model is fine-tuned for different target HSIs. It can be seen that as the number of iterations increases, the loss value decreases rapidly, and the classification accuracy rises rapidly. According to the curve of OA, the fast adaptation to target HSIs can be achieved at about 150 iterations. In short, the process of fine-tuning the learned representations and adapting them to the target HSIs is fast and efficient, which indirectly verifies the versatility and universality of the learned representations for different hyperspectral domains.

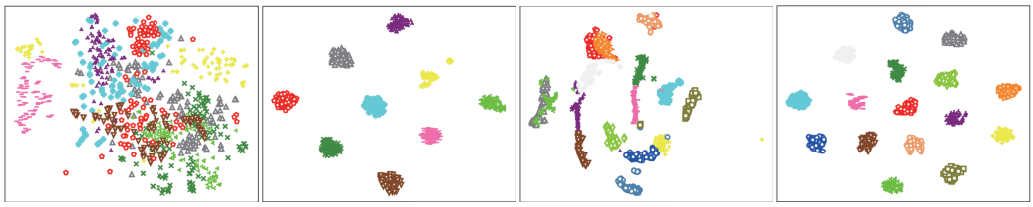



(*a*) UP (*b*) PC (*c*) SA (*d*) IP

**Figure 10.** Curves of loss value and OA during fast adaption to different target HSIs. The red and blue curves represent OA and loss value, respectively.

#### 4.4.3. From the Perspective of Feature Separability

In addition to the above two perspectives, the learned deep representations are visualized to observe the separability between different classes. Specifically, the t-SNE (t-distributed stochastic neighbor embedding) algorithm [61] is used to reduce the dimensionality of the input samples and the high-dimensional vectors generated by the feature extraction module. The UP and SA data sets are taken as examples for observation and analysis, as shown in Figure 11. It can be seen that the separability between input samples is very poor, and after the feature extraction module, the distance between feature vectors corresponding to different classes increases significantly. This shows that, after generalization to different target HSIs, the learned deep representations can effectively enhance the discrimination and separability between different classes so as to improve the accuracy of small-sample classification in the target domains.



(*a*) Input samples (UP) (*b*) Learned representations (UP) (*c*) Input samples (SA) (*d*) Learned representations (SA)

**Figure 11.** Visual presentation of the input samples and learned representations. Different colors represent different classes.

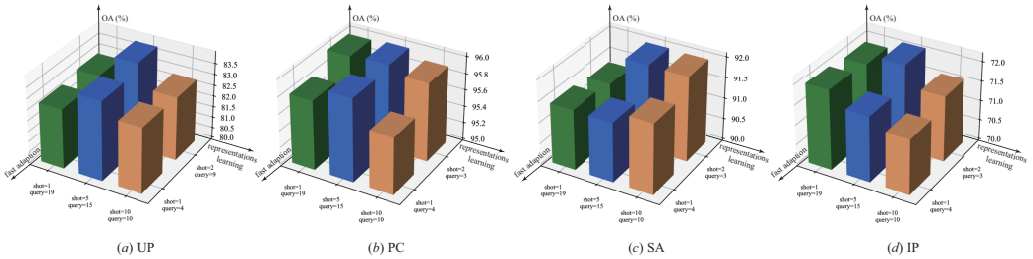
### *4.5. Hyperparameters Analysis*

In this section, the four important hyperparameters of the proposed method, including episode settings, the level of knowledge distillation, the distillation temperature and the hyperparameter $\lambda$, are explored and analyzed in detail.

#### 4.5.1. Episode Settings

As described in Section 3.5, the episode settings include the three keywords of way, shot, and query. In each episode, the value of way is equal to the number of classes in the corresponding HSIs, while the best values for shot and query require further exploration.

During general-purpose representation learning, the values of (shot, query) are set to (1, 19), (5, 15), (10, 10), respectively, and during fast adaption to target HSIs with small samples, the values of (shot, query) are set to (1, 4) and (2, 3). Consequently, the influence of six different episode settings on classification accuracy is explored, and the results are shown in Figure 12. It can be seen that on the four target HSIs, the combination of (5, 15) and (2, 3) can further improve the classification performance of the proposed method, while the other settings all lead to different degrees of decline in classification accuracy.




(*a*) UP (*b*) PC (*c*) SA (*d*) IP

**Figure 12.** The results of the proposed method with different episode settings.

### 4.5.2. The Level of Knowledge Distillation

The proposed three-level knowledge distillation strategy plays an important role in learning general-purpose representations and directly affects the cross-domain classification performance of the designed model. In the statistical results of Table 4, the symbols C, F and L denote the channel-, feature- and logit-level distillation, respectively. It can be observed that the classification accuracy corresponding to distillation only at the feature level is lower than that of other strategies, and the introduction of channel- and logit-level can effectively improve the classification accuracy. There is no doubt that the proposed method can achieve the best classification performance by simultaneously distilling knowledge at the channel-, feature- and logit-level because it provides the possibility for more sufficient knowledge transfer and fusion.

**Table 4.** The results of the proposed method with different distillation levels. The channel-, feature- and logit-level are denoted as C, F and L, respectively. SD denotes the standard deviation of 10 experimental results.

| HSI | Criteria | F | F + C | F + L | F + C + L |
|---|---|---|---|---|---|
| | | *Mean* ± *SD* | *Mean* ± *SD* | *Mean* ± *SD* | *Mean* ± *SD* |
| UP | OA | 82.94 ± 4.17 | 83.20 ± 3.43 | 83.70 ± 4.19 | 84.01 ± 4.31 |
| | AA | 81.31 ± 2.16 | 80.50 ± 2.58 | 81.43 ± 2.75 | 82.32 ± 2.99 |
| | kappa | 78.16 ± 4.80 | 78.40 ± 4.07 | 79.06 ± 4.98 | 79.48 ± 5.14 |
| PC | OA | 95.23 ± 0.96 | 95.55 ± 1.01 | 95.75 ± 0.89 | 96.17 ± 0.74 |
| | AA | 86.93 ± 2.39 | 87.86 ± 1.74 | 88.27 ± 1.73 | 89.00 ± 2.30 |
| | kappa | 93.28 ± 1.34 | 93.71 ± 1.43 | 93.98 ± 1.26 | 94.59 ± 1.04 |
| SA | OA | 91.20 ± 3.94 | 91.15 ± 3.77 | 91.67 ± 4.03 | 92.29 ± 3.55 |
| | AA | 94.12 ± 1.91 | 93.96 ± 1.91 | 94.86 ± 1.92 | 95.15 ± 1.76 |
| | kappa | 90.25 ± 4.34 | 90.19 ± 4.16 | 90.77 ± 4.44 | 91.45 ± 3.92 |
| IP | OA | 71.31 ± 1.94 | 71.79 ± 2.96 | 72.07 ± 2.47 | 72.27 ± 2.61 |
| | AA | 65.21 ± 3.67 | 66.22 ± 2.80 | 65.63 ± 4.32 | 65.26 ± 2.56 |
| | kappa | 67.94 ± 2.17 | 68.48 ± 3.09 | 68.77 ± 2.63 | 68.95 ± 2.88 |

### 4.5.3. Distillation Temperature

Distillation temperature in the logit level is also an important hyperparameter that needs to be analyzed. A higher temperature will produce a softer probability distribution

over classes, while a smaller temperature will increase the difference in classes in the probability distribution. Figure 13 shows the influence of distillation temperature on the classification results. Obviously, the optimal value of temperature corresponding to the four target HSIs is different: two for the UP data set, four for the PC and SA data sets, and six for the IP data set. Note that when the temperature is set to four, the proposed method can achieve high classification accuracy on the four different HSIs. Therefore, it is considered that setting the temperature to four has certain applicability for multiple HSIs.

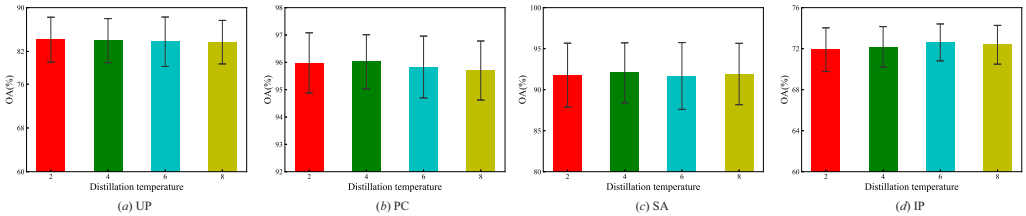



**Figure 13.** The results of the proposed method with different distillation temperatures.

#### 4.5.4. Weight of Distillation Losses

In this subsection, the influence of the weight of distillation losses on the classification results of the proposed method is explored and analyzed. According to Equation 8, the hyperparameter $\lambda$ directly determines the proportion of distillation losses in total losses. Table 5 shows the classification results of the proposed method with different values of hyperparameters $\lambda$. It can be seen that, with the gradual increase in $\lambda$, the classification results of the proposed method are gradually optimized. The $\lambda$ of 0.8 can enable the method to obtain the best classification performance. However, when the value of $\lambda$ increases further to 1.0, the marginal returns occur, and the classification accuracy decreases slightly.

**Table 5.** The results of the proposed method when hyperparameter $\lambda$ changes.

| HSI | Criteria | $\lambda$ = 0.4 *Mean* ± *SD* | $\lambda$ = 0.6 *Mean* ± *SD* | $\lambda$ = 0.8 *Mean* ± *SD* | $\lambda$ = 1.0 *Mean* ± *SD* |
|---|---|---|---|---|---|
| UP | OA | 81.87 ± 4.03 | 83.03 ± 3.63 | 84.01 ± 4.31 | 83.97 ± 4.10 |
| | AA | 79.34 ± 2.52 | 80.30 ± 2.97 | 82.32 ± 2.99 | 82.47 ± 1.33 |
| | kappa | 76.79 ± 4.96 | 78.16 ± 4.25 | 79.48 ± 5.14 | 79.46 ± 4.83 |
| PC | OA | 95.30 ± 0.97 | 95.75 ± 0.99 | 96.17 ± 0.74 | 96.02 ± 0.99 |
| | AA | 86.97 ± 2.41 | 88.03 ± 1.66 | 89.00 ± 2.30 | 88.42 ± 2.16 |
| | kappa | 93.36 ± 1.39 | 93.79 ± 1.50 | 94.59 ± 1.04 | 94.37 ± 1.40 |
| SA | OA | 91.00 ± 3.98 | 91.63 ± 3.83 | 92.29 ± 3.55 | 92.05 ± 3.66 |
| | AA | 94.01 ± 1.93 | 94.42 ± 1.97 | 95.15 ± 1.76 | 94.58 ± 1.91 |
| | kappa | 90.07 ± 4.41 | 90.65 ± 4.26 | 91.45 ± 3.92 | 91.19 ± 4.04 |
| IP | OA | 70.18 ± 1.96 | 71.53 ± 2.77 | 72.27 ± 2.61 | 72.17 ± 1.80 |
| | AA | 64.19 ± 3.52 | 66.04 ± 2.53 | 65.26 ± 2.56 | 66.04 ± 3.57 |
| | kappa | 66.96 ± 2.30 | 68.29 ± 2.99 | 68.95 ± 2.88 | 68.79 ± 2.09 |

### *4.6. Influence of Labeled Target Samples*

The number of labeled target samples for fine-tuning determines the adaptation level of the proposed method to the classification tasks on target HSI domains. Theoretically, with the increase in the number of labeled target samples, the proposed method should achieve better classification performance. Figure 14 shows the variation trend of the classification accuracy of the proposed method on four different target HSIs as the number of labeled target samples per class increases from 5 to 25. The shaded part represents the standard deviation. It can be seen that the classification accuracy of the proposed

method increases steadily, showing a certain ability to adapt to the number of labeled target samples.

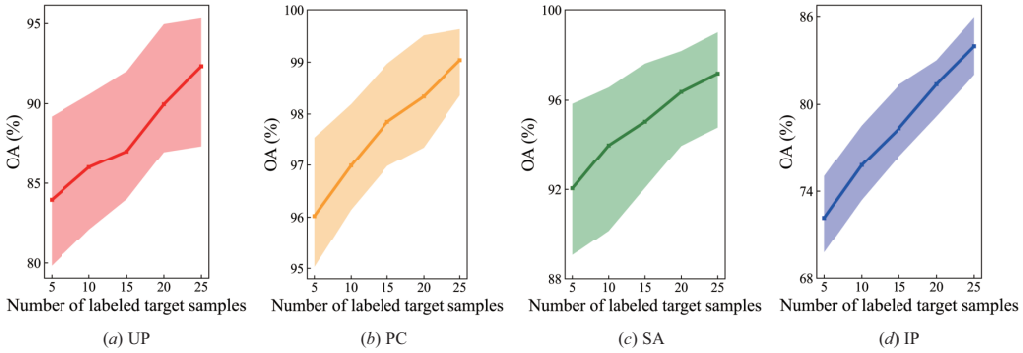



**Figure 14.** The influence of the number of labeled target samples on the classification accuracy. The shaded part represents the standard deviation.

### *4.7. Influence of Different Source HSIs*

The selection of source HSIs determines the characteristics of the learned deep representations and then affects the classification results on the target HSIs. In this subsection, the UP and SA data sets are taken as examples to explore the influence of different source HSIs on classification accuracy, and the results are shown in Figure 15. It can be seen that, with the increase in the number of source HSIs, the classification accuracy on target HSIs first rises and then tends toward stability. Note that the introduction of the CH data set can effectively improve the classification accuracy. It is believed that the CH data set has rich classes, and a total of 18 classes containing water, 3 types of soil, 7 types of vegetation and 7 types of man-made buildings are used for model training, which can provide more rich information and knowledge for learning general-purpose representations.

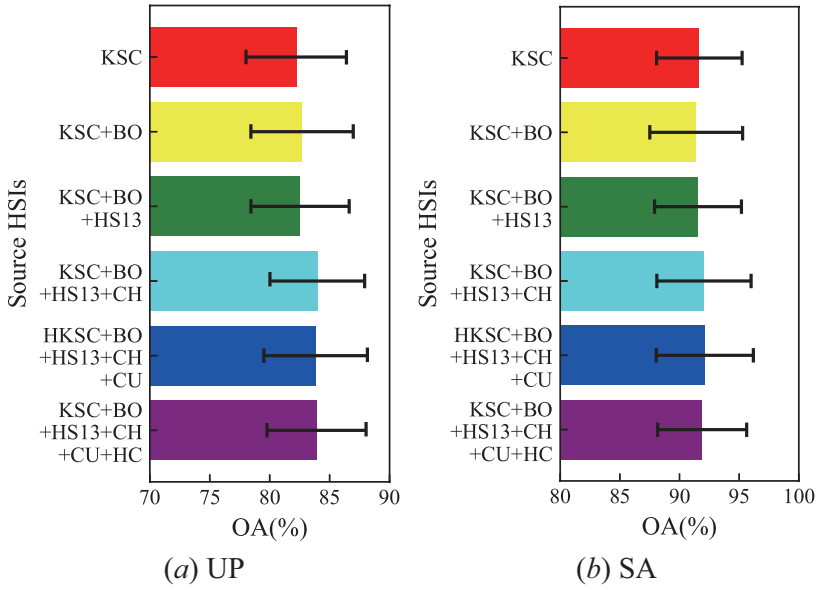



**Figure 15.** The influence of different source HSIs on the classification accuracy.

### *4.8. Efficiency Analysis*

Execution efficiency is an important metric for measuring the application potential of deep models. In this section, the efficiency of three typical cross-domain classification models and the proposed method is compared and analyzed. Specifically, with the UP dataset as the target HSI, the experimental settings are consistent with Section 4.2, and the execution times of different methods in different phases are listed in Table 6. Compared

with the other three models, although the proposed method includes four different stages, its total execution time is still the least. This is mainly because the employment of the three-level distillation strategy effectively improves the ability of the model to fuse the knowledge from different source HSI domains, and thus the model can quickly adapt to the new classification scenarios in target HSI domains with fewer training iterations. Therefore, from the perspective of execution efficiency, the proposed method is superior to the other three cross-domain classification methods.

**Table 6.** Efficiency analysis of different methods.

| Phases | DFSL + SVM | RN-FSC | UM$^2$L | GPRL |
|---|---|---|---|---|
| Pre-training | 118.23 min | 319.84 min | 316.87 min | 15.01 min |
| Knowledge distillation | / | / | / | 37.64 min |
| Fine-tuning | 9.15 s | 80.47 s | 363.32 s | 32.93 s |
| Classification | 1.88 s | 19.12 s | 141.83 s | 13.37 s |

## 5. Discussion and Future Work

Aiming at the problem of cross-domain HSI classification with small samples, this paper proposes a general-purpose representation learning method to further improve the accuracy of small-sample classification on target HSIs based on the full use of information and knowledge in source HSIs. The proposed three-level distillation strategy is the core of the proposed method, which can efficiently transfer and distill knowledge from multiple-source HSIs domains and improve the process of representation learning. The designed multi-task model can perform learning and classification on multiple HSIs simultaneously, skillfully solving the inconsistency of the number of spectral bands in different HSIs and effectively enhancing the adaptability to different target domains. The adopted episode-based learning strategy can effectively improve the generalization ability between different HSIs and data efficiency for small samples. Extensive experiments have demonstrated that by combining the advantages of knowledge distillation, multi-task learning and episode-based training, the proposed method can achieve better results in the cross-domain HSI classification with small samples.

In future work, we will draw on the ideas of self-supervised and unsupervised learning to explore how to use a large number of unlabeled samples that can be easily obtained for knowledge distillation and cross-domain learning and to further improve the classification performance while effectively reducing the dependence on deep models on a large number of source samples.

**Author Contributions:** Methodology, K.G. and A.Y.; investigation, A.Y., X.Y. and C.Q.; resources, A.Y. and X.Y.; writing—original draft preparation, K.G.; writing—review and editing, A.Y.; visualization, K.G., B.L. and W.G.; supervision, A.Y. and X.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China under grants 42130112, 42101458 and 41801388.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study, which can be found here: https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (accessed on 15 January 2023).

**Acknowledgments:** The authors would like to thank all the professionals for kindly providing the codes associated with the experiments.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Xiao, J.; Li, J.; Yuan, Q.; Zhang, L. A Dual-UNet With Multistage Details Injection for Hyperspectral Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5515313. [CrossRef]
2. Booysen, R.; Lorenz, S.; Thiele, S.T.; Fuchsloch, W.C.; Marais, T.; Nex, P.A.; Gloaguen, R. Accurate hyperspectral imaging of mineralised outcrops: An example from lithium-bearing pegmatites at Uis, Namibia. *Remote Sens. Environ.* **2022**, *269*, 112790. [CrossRef]
3. Liu, Z.; Zhong, Y.; Wang, X.; Shu, M.; Zhang, L. Unsupervised Deep Hyperspectral Video Target Tracking and High Spectral-Spatial-Temporal Resolution Benchmark Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5513814. [CrossRef]
4. Cui, Q.; Yang, B.; Liu, B.; Li, Y.; Ning, J. Tea Category Identification Using Wavelet Signal Reconstruction of Hyperspectral Imagery and Machine Learning. *Agriculture* **2022**, *12*, 1085. [CrossRef]
5. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. [CrossRef]
6. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
7. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
8. Paoletti, M.E.; Moreno-Álvarez, S.; Haut, J.M. Multiple Attention-Guided Capsule Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5520420. [CrossRef]
9. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
10. Zhao, C.; Li, C.; Feng, S.; Li, W. Spectral–Spatial Anomaly Detection via Collaborative Representation Constraint Stacked Autoencoders for Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5503105. [CrossRef]
11. Zhang, L.; Zhang, L. Artificial Intelligence for Remote Sensing Data Analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
12. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
13. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature Extraction for Hyperspectral Imagery: The Evolution From Shallow to Deep: Overview and Toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88. [CrossRef]
14. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2290–2304. [CrossRef]
15. Gao, K.; Liu, B.; Yu, X.; Qin, J.; Zhang, P.; Tan, X. Deep Relation Network for Hyperspectral Image Few-Shot Classification. *Remote Sens.* **2020**, *12*, 923. [CrossRef]
16. Lee, H.; Eum, S.; Kwon, H. Exploring Cross-Domain Pretrained Model for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5526812. [CrossRef]
17. Ma, X.; Mou, X.; Wang, J.; Liu, X.; Wang, H.; Yin, B. Cross-Data Set Hyperspectral Image Classification Based on Deep Domain Adaptation. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10164–10174. [CrossRef]
18. Ma, X.; Mou, X.; Wang, J.; Liu, X.; Geng, J.; Wang, H. Cross-Dataset Hyperspectral Image Classification Based on Adversarial Domain Adaptation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4179–4190. [CrossRef]
19. Qin, Y.; Bruzzone, L.; Li, B.; Ye, Y. Cross-Domain Collaborative Learning via Cluster Canonical Correlation Analysis and Random Walker for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3952–3966. [CrossRef]
20. Zhang, C.; Ye, M.; Lei, L.; Qian, Y. Feature Selection for Cross-Scene Hyperspectral Image Classification Using Cross-Domain I-ReliefF. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5932–5949. [CrossRef]
21. Shen, J.; Cao, X.; Li, Y.; Xu, D. Feature Adaptation and Augmentation for Cross-Scene Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 622–626. [CrossRef]
22. Miao, J.; Zhang, B.; Wang, B. Coarse-to-Fine Joint Distribution Alignment for Cross-Domain Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12415–12428. [CrossRef]
23. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
24. Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; Zhang, L. 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 Sens. Environ.* **2020**, *250*, 112012. [CrossRef]
25. Gao, K.; Liu, B.; Yu, X.; Yu, A. Unsupervised Meta Learning With Multiview Constraints for Hyperspectral Image Small Sample set Classification. *IEEE Trans. Image Process.* **2022**, *31*, 3449–3462. [CrossRef]
26. Paoletti, M.E.; Haut, J.M.; Tao, X.; Plaza, J.; Plaza, A. FLOP-Reduction Through Memory Allocations Within CNN for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5938–5952. [CrossRef]
27. Liu, B.; Yu, X.; Zhang, P.; Yu, A.; Fu, Q.; Wei, X. Supervised Deep Feature Extraction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 1909–1921. [CrossRef]
28. Gao, K.; Guo, W.; Yu, X.; Liu, B.; Yu, A.; Wei, X. Deep Induction Network for Small Samples Classification of Hyperspectral Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3462–3477. [CrossRef]

29. Xue, Z.; Yu, X.; Tan, X.; Liu, B.; Yu, A.; Wei, X. Multiscale Deep Learning Network With Self-Calibrated Convolution for Hyperspectral and LiDAR Data Collaborative Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5514116. [CrossRef]
30. Tan, X.; Gao, K.; Liu, B.; Fu, Y.; Kang, L. Deep global-local transformer network combined with extended morphological profiles for hyperspectral image classification. *J. Appl. Remote Sens.* **2021**, *15*, 038509. [CrossRef]
31. Xu, Q.; Wang, D.; Luo, B. Faster Multiscale Capsule Network With Octave Convolution for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 361–365. [CrossRef]
32. Liu, B.; Gao, K.; Yu, A.; Guo, W.; Wang, R.; Zuo, X. Semisupervised graph convolutional network for hyperspectral image classification. *J. Appl. Remote Sens.* **2020**, *14*, 026516. [CrossRef]
33. Wan, S.; Pan, S.; Zhong, P.; Chang, X.; Yang, J.; Gong, C. Dual Interactive Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5510214. [CrossRef]
34. He, X.; Chen, Y. Transferring CNN Ensemble for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 876–880. [CrossRef]
35. He, X.; Chen, Y.; Ghamisi, P. Heterogeneous Transfer Learning for Hyperspectral Image Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3246–3263. [CrossRef]
36. Liu, B.; Yu, A.; Yu, X.; Wang, R.; Gao, K.; Guo, W. Deep Multiview Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7758–7772. [CrossRef]
37. Hou, S.; Shi, H.; Cao, X.; Zhang, X.; Jiao, L. Hyperspectral Imagery Classification Based on Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5521213. [CrossRef]
38. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
39. Li, W.; Liu, X.; Bilen, H. Universal Representations: A Unified Look at Multiple Task and Domain Learning. *arXiv* **2022**. [CrossRef]
40. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
41. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. *arXiv* **2014**, arXiv:1412.6550.
42. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2016**, arXiv:1612.03928.
43. Yim, J.; Joo, D.; Bae, J.; Kim, J. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7130–7138. [CrossRef]
44. Li, W.H.; Bilen, H. Knowledge Distillation for Multi-task Learning. In *Proceedings of the Computer Vision—ECCV 2020 Workshops*; Bartoli, A.; Fusiello, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 163–176.
45. Shi, C.; Fang, L.; Lv, Z.; Zhao, M. Explainable scale distillation for hyperspectral image classification. *Pattern Recognit.* **2022**, *122*, 108316. [CrossRef]
46. Yue, J.; Fang, L.; Rahmani, H.; Ghamisi, P. Self-Supervised Learning With Adaptive Distillation for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sensing* **2022**, *60*, 5501813. [CrossRef]
47. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv* **2017**, arXiv:1703.03400.
48. Snell, J.; Swersky, K.; Zemel, R. Prototypical Networks for Few-shot Learning. *arXiv* **2017**, arXiv:1703.05175.
49. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.; Hospedales, T. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208. [CrossRef]
50. Geng, R.; Li, B.; Li, Y.; Zhu, X.; Jian, P.; Sun, J. Induction Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019. [CrossRef]
51. Gao, K.; Liu, B.; Yu, X.; Zhang, P.; Tan, X.; Sun, Y. Small sample classification of hyperspectral image using model-agnostic meta-learning algorithm and convolutional neural network. *Int. J. Remote Sens.* **2021**, *42*, 3090–3122. [CrossRef]
52. Ma, X.; Ji, S.; Wang, J.; Geng, J.; Wang, H. Hyperspectral Image Classification Based on Two-Phase Relation Learning Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10398–10409. [CrossRef]
53. Motiian, S.; Jones, Q.; Iranmanesh, S.M.; Doretto, G. Few-Shot Adversarial Domain Adaptation. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; pp. 6670–6680.
54. Bi, H.; Liu, Z.; Deng, J.; Ji, Z.; Zhang, J. Contrastive Domain Adaptation-Based Sparse SAR Target Classification under Few-Shot Cases. *Remote Sens.* **2023**, *15*, 469. [CrossRef]
55. Rostami, M.; Kolouri, S.; Eaton, E.; Kim, K. Deep Transfer Learning for Few-Shot SAR Image Classification. *Remote Sens.* **2019**, *11*, 1374. [CrossRef]
56. Lasloum, T.; Alhichri, H.; Bazi, Y.; Alajlan, N. SSDAN: Multi-Source Semi-Supervised Domain Adaptation Network for Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 3861. [CrossRef]
57. Shi, Y.; Li, J.; Li, Y.; Du, Q. Sensor-Independent Hyperspectral Target Detection With Semisupervised Domain Adaptive Few-Shot Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6894–6906. [CrossRef]
58. Wang, L.; Hao, S.; Wang, Q.; Wang, Y. Semi-supervised classification for hyperspectral imagery based on spatial-spectral Label Propagation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 123–137. [CrossRef]

59. Roy, S.K.; Haut, J.M.; Paoletti, M.E.; Dubey, S.R.; Plaza, A. Generative Adversarial Minority Oversampling for Spectral–Spatial Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5500615. [CrossRef]
60. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event, 18–24 July 2021; Meila, M., Zhang, T., Eds.; Volume 139, pp. 12310–12320.
61. van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.