Combining Spatial-Spectral Features for Hyperspectral Image Few-Shot Classification

Qiong Ran, Li Ni, Yonghao Zhou

Combining Spatial-spectral Features for Hyperspectral Image Few-shot Classification *

Yonghao Zhou[1,2], Qiong Ran[1,*], and Li Ni[2]

[1] College of Information Science and Technology
Beijing University of Chemical Technology, Beijing, 100029, China
[2] Key Laboratory of Computational Optical Imaging Technology , Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
2020210532@mail.buct.edu.cn ranqiong@mail.buct.edu.cn
nili@aircas.ac.cn

Abstract. Recently, deep learning has achieved considerable results in hyperspectral image (HSI) classification. However, when training image classification models, existing deep networks require sufficient samples, which is expensive and inefficient in practical tasks. In this article, a novel Combining Spatial-spectral Features for Hyperspectral Image Few-shot Classification (CSFF) framework is proposed, attempting to accomplish the fine-grained classification with only a few labeled samples and train it with meta-learning ideas. Specifically, firstly, the spatial attention (SPA) and spectral query (SPQ) modules are introduced to overcome the constraint of the convolution kernel and consider the information between long-distance location (non-local) samples to reduce the uncertainty of classes. Secondly, the framework is trained by episodes to learn a metric space, and the task-based few-shot learning (FSL) strategy allows the model to continuously enhance the learning capability. In addition, the designed network not only discovers transferable knowledge in the source domain (SD) but also extracts the discriminative embedding features of the target domain (TD) classes. The proposed method can obtain satisfactory results with a small number of labeled samples. Extensive experimental results on public datasets demonstrate the versatility of CSFF over other state-of-the-art methods.

Keywords: Hyperspectral Image Classification · Spatial-spectral · Few-shot Learning · Domain Adaption · Meta Learning.

1 Introduction

Hyperspectral image (HSI) contains rich spatial-spectral information and provides the possibility of accurate classification of complex features. As a result, it has been widely used in environmental monitoring and military defense, etc. Currently, it urgently requires accurate classification of HSI with the development toward big data, which demands sufficient labeled samples [8]. However, it

* This work was supported by the National Natural Science Foundation of China under Grant 62161160336 and Grant 42030111.

is extremely difficult to obtain thousands of labeled samples without great human and material resources. In earlier years, Melgai et al. [13] used the support vector machine (SVM) to explore an optimal hyperplane for classification, which slightly alleviated the "hughes" phenomenon. It only calculated the spectral information of HSI without considering spatial features. Chen et al. [2] designed a deep network to extract deep invariant features. Moreover, some strategies, such as L2 regularization and dropout were investigated to avoid overfitting during training. In contrast, humans can combine empirical knowledge and thus quickly complete new classification tasks with only a few samples.

In recent years, FSL has become popular because of its ability to perform new classification tasks with only a few labeled samples. For example, Chen et al. [1] introduced classifier baselines and FSL baselines and proposed to pretrain the classification model through a meta-learning paradigm. Lately, Gao et al. [3] designed a relation network (RN-FSC) to classify HSI, which fine-tuned the SD training model by using a few shot datasets of the TD. A deep feature extraction FSL method (DFSL) with an attached classifier was proposed in [11]. Moreover, Li et al. [10] designed a supervised deep cross-domain FSL network (DCFSL), which adopted residual 3D-CNN networks to extract local information and ignored the significance of non-local spatial features. Although the above FSL-based networks utilized convolution kernels to extract spatial-spectral features, the information is rarely obtained from long-distance location samples [5]. Secondly, due to the frequent occurrence of spectral shifts, Various discrepancies in data distribution may occur between SD and TD [15]. Therefore, it is necessary to consider the non-local relationships between samples to reduce the negative effects of domain shifts.

To overcome the above two limitations, a Combining Spatial-spectral Features for Hyperspectral Image Few-shot Classification (CSFF) framework is proposed, which is based on the mechanism of combining domain adaptation and FSL. Firstly, the episodic learning pattern of FSL is implemented on the SD and TD, which is to build a meta-task (i.e., support set and query set). Then, the spatial-spectral information is extracted by SPA and SPQ units. Moreover, the similarity between the support set S and query set Q is calculated using a metric function. Finally, a domain adaptation strategy is adopted to overcome domain shifts and achieve the accurate classification of HSI. The main contributions of this paper can be summarized as follows.

1) Unlike most existing deep networks, the proposed CSFF is learning a metric space through the episodic and task-based learning strategy, which can obtain promising HSI classifications with only a few labeled samples.
2) The SPA and the SPQ modules introduced through transformers are designed to overcome the constraint of the convolution kernel, consider the relationship between long-distance location samples, and enable the network to better extract high-level features to reduce the uncertainty of the class.
3) Rather than focusing on a specific classification task, the proposed approach is to learn a deep nonlinear and transferable metric space, where the similarity metric is implemented by comparison. Meanwhile, to reduce the distribution

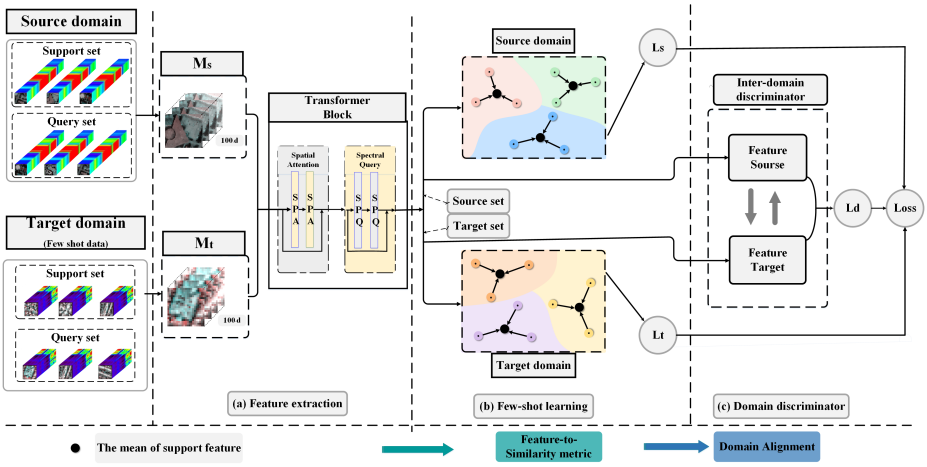

Fig. 1. Framework of the proposed CSFF, including feature extraction, few-shot learning and domain alignment.

difference between the SD and TD, we adopt a domain adaptation strategy to achieve distribution alignment of the data, which can help improve the generalization power of the model.

The remaining of the paper is arranged as follows. Section 2 introduces relevant concepts of the proposed approach CSFF. Experimental results and analyses are presented in Section 3. Finally, Section 4 draws comprehensive conclusions of this work.

2 Proposed Approach

The framework of the proposed CSFF is shown in Fig. 1, which contains three parts, i.e., feature extraction, few-shot learning, and domain alignment. During feature extraction, SPA and SPQ blocks are designed to overcome the limitations of fixed convolution kernel size. Also, an inter-domain discriminator (IDD) is used to alleviate the problem of domain shift caused by different sensors. Assuming that C_s, C_t represent the number of categories of $X_s \in \mathbb{R}^{d_s}$ and $X_t \in \mathbb{R}^{d_t}$, which denote d_s and d_t dimensional features from SD and TD. Note that TD is separated into training data $\mathcal{T}_f$ with a few labeled samples and testing data $\mathcal{T}_t$ with unlabeled samples.

2.1 Feature Extraction

Generally, HSI usually requires pre-processing with high dimensions. The mapping layers, $\mathbf{M}_s(\bullet)$ and $\mathbf{M}_t(\bullet)$ are used to map SD and TD to the same dimension d_map (d_map is set to 100 in this work). The transformer block $f_{net}(\bullet)$,

including SPA and SPQ blocks, is employed to extract the spatial-spectral features of the HSI.

Most neural networks extract features from the local space but ignore the significance of the relationship between non-local space samples [7]. However, the purpose of the spatial attention mechanism is to explore the interactions between samples at different positions. Motivated by [4, 6, 14], we design the transformer's SPA blocks to calculate query, key, and value tensors as follows.

$$\mathbf{Q} = \mathcal{F}(X, W_q) \in \mathbb{R}^{c^{'} \times h \times w} \tag{1}$$

$$\mathbf{K} = \mathcal{F}(X, W_k) \in \mathbb{R}^{c^{'} \times h \times w} \tag{2}$$

$$\mathbf{V} = \mathcal{F}(X, W_v) \in \mathbb{R}^{c \times h \times w} \tag{3}$$

where: W_q, W_k, W_v denote the training parameters of the query, key, and value tensor, respectively, c, h, w are the channel size, height, and width of the input features X, respectively, $\mathcal{F}(\bullet)$ denotes the 2D convolution operation. Thus, the output of the SPA block can be calculated as follows:

$$SPA_{out} = \mathbf{V} \cdot softmax\left(\mathbf{Q}^{\mathsf{T}} \cdot \mathbf{K}\right) \in \mathbb{R}^{c \times h \times w} \tag{4}$$

Where SPA_{out} is the output of each position on the feature map. Now, the SPA module establishes the interactions between samples at different locations but ignores the abundant spectral information of HSI. Consequently, we design another SPQ block to extract spectral features and take masks to fuse the spatial information. Specifically, the kernel and output of the SPQ block can be formulated as follows:

$$\Psi = softmax\left(\mathcal{H}(X, W_\Psi)\right) \in \mathbb{R}^{hw \times k} \tag{5}$$

$$SPQ = \Psi^{\mathsf{T}} \cdot X^{\mathsf{T}} = (X \cdot \Psi)^{\mathsf{T}} \in \mathbb{R}^{k \times c} \tag{6}$$

$$SPQ_{out} = SPQ^{\mathsf{T}} \cdot \Psi^{\mathsf{T}} = (\Psi \cdot SPQ)^{\mathsf{T}} \in \mathbb{R}^{c \times h \times w} \tag{7}$$

where $\mathcal{H}(\bullet)$ denotes 3D convolution operation to produce a tensor of size $k \times h \times w$. Finally, we use the generated mask W_Ψ to integrate it with the SPA_{out} of the input X, generating a spectral query kernel SPQ of size $k \times c$. So far, the SPQ block has established correlations between spatial locations and corresponding spectral features.

2.2 Source and Target Few-shot Learning

FSL is executed simultaneously in source and target classes in each episode. Taking SD as an example, SD is divided into support set $\mathcal{S}_s$ and query set $\mathcal{Q}_s$, where $\mathcal{S}_s$ is the training set. Then, the features f_{S_s} and f_{Q_s} are extracted by network f_{net}. In each episode, FSL calculates the similarity between f_{Q_s} and each class prototype and minimizes the predicted loss. The predicted probability of the query sample is performed as follows:

$$P(\hat{y}|\,\mathbf{x}_i) = softmax\left(-E\left(\mathbf{x_i}, \mathbf{x_s^{c_i}}\right)\right) \tag{8}$$

where $E(\bullet)$ denotes an Euclidean distance function, $\mathbf{x}_s^{c_i}$ is the c_i-th class prototype of f_{S_s}, $c_i \in C_s$. The FSL loss of a query sample in SD is calculated by cross-entropy loss.

$$\mathcal{L}_S\left(P\left(\hat{y}|\,\mathbf{x}_i\right), y_i\right) = -\sum_{Q_s, i=1}^{C_s} y_i \log P\left(\hat{y}|\,\mathbf{x}_i\right) \tag{9}$$

Equivalently, the loss $\mathcal{L}_T$ of the TD is formulated in the same way as above.

2.3 Domain Alignment

Given the effect of domain shift on classification performance in FSL episodic training, domain alignment is one of the effective measures. Inspired by [12], we design an IDD block to analyze and adjust the data distributions $P_s(x)$ and $P_t(x)$ of the SD and TD. In particular, we denote $h = (f, g)$ to represent the joint distribution of the feature $f = F(x)$ and the classifier prediction $g = G(x)$. Following this, we formulate the domain alignment network as a minimax optimization problem with a loss error term:

$$\begin{aligned}\mathcal{L}_d \leftarrow \mathcal{E}(\mathbf{D}, \mathbf{G}) = & -\mathbb{E}_{x_i^s \sim P_s(x)} \log\left[D\left(f_i^s, g_i^s\right)\right] \\ & -\mathbb{E}_{x_j^t \sim P_t(x)} \log\left[1 - D\left(f_j^t, g_j^t\right)\right]\end{aligned} \tag{10}$$

where $D(\bullet, \bullet)$ and $1 - D(\bullet, \bullet)$ denote the probability that IDD predicts SD and TD samples x, $\mathcal{E}(\mathbf{D}, \mathbf{G})$ can be considered as the loss metric of the IDD block, which is minimized over IDD but maximized over $F(x)$ and $G(x)$. By combining $h = (f, g)$, we condition IDD on g with the multilinear map as follows.

$$\mathbf{T}_{\otimes}(f, g) = f \otimes g \in \mathbb{R}^{d_f \times d_g} \tag{11}$$

where $(f \otimes g)$ defined as the outer product of multiple d_f and d_g dimensions random vectors. However, with the increasing number of training iterations, the dimension $d_f \times d_g$ of the multilinear map will become too high to be embedded the deep framework without causing parameter explosion. Luckily, according to the theoretical proof in [12], the dimension d of the randomized multilinear map($\mathbf{T}_{\odot}(f, g)$) is much smaller than $d_f \times d_g$. In other words, $\mathbf{T}_{\odot}$ is an approximate calculation of $\mathbf{T}_{\otimes}$, where $\odot$ is element-wise produc. If the dimension of $\mathbf{T}_{\otimes}$ is too large, we will adopt another strategy $\mathbf{T}_{\odot}$. Finally, the total objective function loss (together with Eq: 9) is shown as following,

$$\mathcal{L}_{oss} = \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_d \tag{12}$$

In this paper, we utilize multi-layer perceptrons in the IDD block. Furthermore, $\mathcal{T}_f$ is regarded as the support set and $\mathcal{T}_t$ as the query set in the testing stage.

Table 1. Classification results(%) on the UP data set with different methods (5 labeled samples from TD).

Class	Samples	Classification algorithms						
		SVM	3D-CNN	DFSL+NN	DFSL+SVM	RN-FSC	DCFSL	CSFF
Asphalt	6631	60.00	73.22	73.27	75.33	73.98	83.53	92.68±4.11
Meadows	18549	49.21	73.25	78.20	86.03	88.80	86.20	84.55±9.93
Gravel	2099	57.19	32.53	51.94	51.33	52.07	67.72	71.60±9.83
Trees	3064	79.49	86.30	85.95	90.91	90.64	94.26	91.28±2.72
Sheets	1345	90.74	95.35	99.37	97.64	98.94	98.85	99.58±0.43
Bare soil	5029	62.93	38.16	61.70	55.62	51.70	70.88	72.00±13.18
Bitumen	1330	80.96	43.82	69.75	71.09	71.86	79.92	82.61±9.81
Bricks	3682	62.55	49.24	53.34	55.46	58.62	65.92	85.66±3.57
Shadow	947	99.71	94.22	97.13	91.74	98.90	98.60	93.08±5.91
OA		59.60	62.50	73.44	76.84	77.89	82.39	**84.88±3.33**
AA		71.42	65.12	74.52	75.02	76.17	82.88	**85.89±2.10**
Kappa		50.77	52.67	65.77	69.61	70.83	77.06	**80.40±3.88**

3 Experimental Results

To prove the validity of the proposed framework CSFF, two publicly available HSI data sets were collected. The details of the two data sets are listed as following. Several state-of-the-art classification methods are adopted for comparison algorithms, SVM, 3D-CNN [9], DFSL+NN [11], DFSL+SVM [11], relation network (RN-FSC) [3], and DCFSL [10].

Source domain: the Chikusei data contains 19 classes and has 128 bands in the spectral range from 363 nm to 1018 nm. It has 2517×2335 pixels and a spatial resolution of 2.5m. Target domain: the University of Pavia data(UP) has 9 classes and 103 spectral bands in the spectral range from 430 nm to 860nm. The size of the image is 610×340 pixels with a spatial resolution is 1.3m per pixel.

3.1 Experimental Setting and Performance

In CSFF, 9×9 neighborhoods are selected as the spatial size of the input data. The learning rate is set to 0.001 and the number of training iterations is 10000 with being trained via Adam optimizer. For each meta-task of C-way K-shot in episodic training, C is set to the same number of classes as in TD. K for SD FSL and TD FSL is set to 1 in FSL-based experiments. In addition, the number of the query samples in $\mathcal{Q}$ is set to 19 to evaluate the learned classifier. Note that SVM and 3D-CNN only utilize the few-shot data set from the TD can to train a classifier. Furthermore, 5 labeled samples are randomly selected from each class of TD for FSL, and the data $\mathcal{T}_f$ is augmented by adding random Gaussian noise to the current known samples.

Table 1 reports the performance of all methods with overall accuracy (OA%), average accuracy (AA%), and kappa coefficients (Kappa%) in TD. Compared with SVM and 3D-CNN [9], several other FSL-based methods, including the proposed CSFF, provide over 9% improvement in both OA and AA. It indicates that FSL methods trained with meta-learning ideas in SD can better address the

problem of few labeled samples in TD. Compared with DFSL+NN(SVM)[11] and RN-FSC[3] methods (without domain adaptation), DCFSL[10] and CSFF increased Kappa by 6.23% to 14.63%, which demonstrates that domain adaptation is essential. Furthermore, compared with the DCFSL [10] that only focuses on local features, the classification accuracy of CSFF is slightly lower than that of DCFSL for a few classes (i.e., Meadows, Trees and Shadow), which may be explained by the fact that the experimental UP HSI was taken during a period of lush green vegetation, in which some of the trees and pasture are similar in visual color. At the same time, some trees, shadows, and meadows overlap each other in the spatial distribution. Both of them can trigger serious spectral confusion. However, due to the transformer block can integrate non-local sample information, CSFF performs significantly better than DCFSL on non-vegetation classes. In particular, spectral shifts are significantly mitigated in some classes to enhance the classification performance, such as the 3-th class (Gravel), 6-th class (Bare soil), and 8-th class (Bricks) in the UP.

4 Conclusions

In this article, Combining Spatial-spectral Features for Hyperspectral Image Few-shot Classification (CSFF) has been proposed to address the issues of HSI classification with only a few labeled samples. It attempts to overcome the geometric constraints of the convolution kernel and reduce the negative effect of domain shift on FSL. Specifically, the spatial attention and spectral query modules are designed to extract and aggregate information from non-local samples in SD and TD. In addition, the framework is trained by episodes to learn a metric space, and a conditional domain adaptation strategy is utilized to achieve domain distribution alignment. The experimental results demonstrate that the proposed method has presented significant improvements over the state-of-the-art models. In the future, we will consider integrating local and non-local information (e.g., topological structure) and designing a multi-constrained domain distribution discrepancy metric to further reduce the data distribution differences. Meanwhile, a deep combined domain adaptation network will be constructed to achieve accurate classification of cross-domain hyperspectral images with a few labeled instances.

References

1. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9042–9051 (2021). https://doi.org/10.1109/ICCV48922.2021.00893
2. Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P.: Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing **54**(10), 6232–6251 (2016). https://doi.org/10.1109/TGRS.2016.2584107

3. Gao, K., Liu, B., Yu, X., Qin, J., Zhang, P., Tan, X.: Deep relation network for hyperspectral image few-shot classification. Remote Sensing **12**(6) (2020). https://doi.org/10.3390/rs12060923
4. He, J., Zhao, L., Yang, H., Zhang, M., Li, W.: Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers. IEEE Transactions on Geoscience and Remote Sensing **58**(1), 165–178 (2020). https://doi.org/10.1109/TGRS.2019.2934760
5. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., Chanussot, J.: Graph convolutional networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **59**(7), 5966–5978 (2021). https://doi.org/10.1109/TGRS.2020.3015157
6. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B.: More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 4340–4354 (2021). https://doi.org/10.1109/TGRS.2020.3016820
7. Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J.: Spectralformer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022). https://doi.org/10.1109/TGRS.2021.3130716
8. Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A.: Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing **57**(9), 6690–6709 (2019). https://doi.org/10.1109/TGRS.2019.2907932
9. Li, Y., Zhang, H., Shen, Q.: Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. Remote Sensing **9**(1) (2017). https://doi.org/10.3390/rs9010067
10. Li, Z., Liu, M., Chen, Y., Xu, Y., Li, W., Du, Q.: Deep cross-domain few-shot learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2022). https://doi.org/10.1109/TGRS.2021.3057066
11. Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G., Wang, R.: Deep few-shot learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **57**(4), 2290–2304 (2019). https://doi.org/10.1109/TGRS.2018.2872830
12. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Domain adaptation with randomized multilinear adversarial networks. ArXiv (2017). https://doi.org/10.48550/arXiv.1705.10667
13. Melgani, F., Bruzzone, L.: Support vector machines for classification of hyperspectral remote-sensing images. In: IEEE International Geoscience and Remote Sensing Symposium. vol. 1, pp. 506–508 vol.1 (2002). https://doi.org/10.1109/IGARSS.2002.1025088
14. Sun, H., Zheng, X., Lu, X., Wu, S.: Spectral–spatial attention network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **58**(5), 3232–3245 (2020). https://doi.org/10.1109/TGRS.2019.2951160
15. Tuia, D., Persello, C., Bruzzone, L.: Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geoscience and Remote Sensing Magazine **4**(2), 41–57 (2016). https://doi.org/10.1109/MGRS.2016.2548504