Domain Adaptive Remote Sensing Scene Classification With Middle-Layer Feature Extraction and Nuclear Norm Maximization

Ruitong Du, Guoqing Wang, *Graduate Student Member, IEEE*, Ning Zhang, Liang Chen, and Wenchao Liu

***Abstract*—Unsupervised domain adaptation (UDA) methods have become a research hotspot in remote sensing scene classification to reduce dependence on labeled samples. However, most current methods focus on extracting domain invariant features, ignoring the problem of large intraclass differences and the imbalanced sample numbers between categories in remote sensing images. To address these issues, we propose a remote sensing scene domain adaptive method based on middle-layer feature extraction and nuclear norm maximization (MFE-NM). In the MFE module, the middle-layer features of the feature extractor are randomly extracted and processed. Since the receptive field of the middle-layer features is smaller and the resolution is higher, the effective use of the middle-layer features can reduce the impact of image feature heterogeneity caused by large intraclass differences in remote sensing images. In addition, it can be concluded that the constrained nuclear norm can simultaneously improve the prediction diversity and discriminability of the model through theoretical derivation. Therefore, the NM module is proposed to solve the problem of reduced prediction diversity caused by entropy minimization methods when dealing with scene classification problems with imbalanced sample numbers between categories. Extensive experiments and analyses on three public remote sensing datasets demonstrate the effectiveness and competitiveness of our proposed method.**

***Index Terms*—Middle-layer feature extraction (MFE), nuclear norm maximization (NM), remote sensing scene classification, unsupervised domain adaptation (UDA).**

I. Introduction

With the advancement of satellite and remote sensing technology, the number of remote sensing images is rapidly increasing [1], [2], [3]. Scene classification is a method that can effectively process remote sensing images, which aims to classify remote sensing images into different semantic categories. Remote sensing scene classification plays an important role in urban planning, geological hazard detection, and other fields [4]. However, due to differences in geographical distribution, imaging conditions, sensors, etc., the data distribution of different remote sensing datasets is distinct [5]. To adapt to the distribution differences between different datasets, domain adaptation (DA) methods have been proposed [6]. DA consists of two domains, the source domain and the target domain. Due to differences in data distribution between different domains, it is often difficult to obtain satisfactory results when the model trained on the source domain is tested directly on the target domain [7]. In DA, the data from the source and target domains are mapped to the same feature space, and the distribution differences between the two domains are minimized in this feature space, enabling the target domain to fully utilize the rich information in the source domain.

At present, deep learning plays a significant role in the field of DA [8], [9], [10], [11], [12], [13]. Convolutional neural network (CNN) is one of the most representative models in deep learning methods. CNN usually needs to use a large amount of labeled data during the training process. Although the number of remote sensing images that can be obtained has greatly increased, labeling these images not only relies on expert knowledge, but also consumes a large amount of human resources, which is usually uneconomical [14]. To reduce the dependence on labeled data, unsupervised learning is introduced [15]. Compared with supervised or semisupervised methods, the target domain in unsupervised domain adaptation (UDA) technology does not contain labeled samples, but learns relevant information from the labeled source domain [16], [17].

UDA methods can generally be divided into two categories. The first type is based on statistical methods, which use mean or higher order moments to measure the distribution differences between domains and minimize this statistical measure to align different domains [8]. The second type is based on adversarial learning methods. However, due to the complex features of remote sensing images, manually designed statistical metrics are difficult to characterize complex feature distribution information. Therefore, researchers mostly focus on methods based on adversarial learning. Domain adaptive neural network (DANN) [18] is the first method to introduce the generative adversarial networks (GANs) into the field of transfer learning.

Manuscript received 13 July 2023; revised 23 October 2023; accepted 28 November 2023. Date of publication 5 December 2023; date of current version 8 January 2024. This work was supported in part by the National Natural Science Foundation for Young Scientists of China under Grant 62201059 and in part by the Foundation under Grant SAST2021-040. *(Corresponding author: Liang Chen.)*

The authors are with National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China (e-mail: 3220210578@bit.edu.cn; gqwang@bit.edu.cn; 3120205375@bit.edu.cn; chenl@bit.edu.cn; 6120220028@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3339336

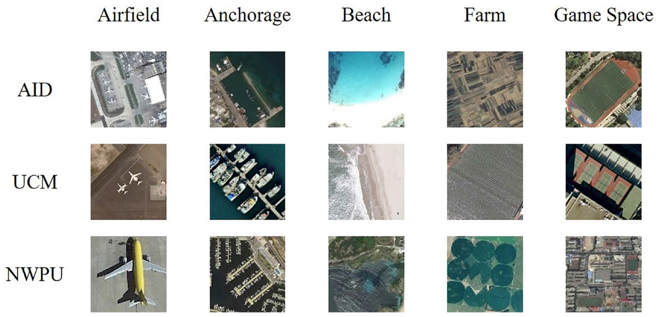


Fig. 1. Examples of five types of images in different datasets. Shown are five types of images in the UC Merged land use dataset (UCM) [20], Aerial Image Datasets (AID) [21], and NWPU-RESISC45 (NWPU) [22]. It can be found that even for the same category, there are differences in key parts of images in different datasets.

In DANN, the domain discriminator is used to distinguish whether the sample comes from the source domain or the target domain, while the feature extractor extracts features that are difficult for the domain discriminator to distinguish as much as possible. When the domain discriminator and feature extractor reach dynamic equilibrium, the source domain and target domain are aligned. Although this adversarial learning-based methods can reduce the distribution differences between different domains in the feature space, it is likely to sacrifice the discriminability of features [19].

In traditional adversarial training methods, only the last layer features output by the feature extractor is usually selected to represent image features. However, as shown in Fig. 1, due to the difference in shooting satellite, location, and other factors, the key parts of remote sensing images are quite different, so it is not enough to only extract the deep features of the image. Compared with the deep features, the shallow features of the image have a smaller receptive field and higher resolution, which enables the network to capture detailed information. The selection method of middle-layer features can also have a significant impact on performance. Therefore, we adopt the method of randomly extracting the middle-layer feature to enhance the discrimination of the model in the feature extraction module. By fully utilizing the middle-layer features, it is possible to capture key parts in remote sensing images and reduce the distribution differences between the source and target domains.

At the same time, as shown in Fig. 2, in the same remote sensing dataset, there may be an imbalance in the number of images between categories. In this case, the method based on entropy minimization currently used has side effects [23]. It tends to judge a small number of category samples as a large number of categories, which will reduce the prediction diversity of unlabeled data. In response to the issue of category imbalance, Bai et al. [24] used focal loss to reallocate the losses of samples from different categories. Although the focal loss can effectively suppress the negative impact of imbalanced category numbers on model training, it suppresses well-classified samples and inevitably introduces difficult-to-classify samples. Cui et al. [23] proposed that in an ideal state, the F-norm and rank of the output matrix can be used to measure the prediction discriminability and diversity of unlabeled samples, thereby reducing the impact of reduced prediction diversity caused by the imbalanced sample numbers between remote sensing image categories. However, the actual training results often deviate from the ideal state. Therefore, we introduce a scale factor to make the prediction output closer to the assumed ideal state. Overall, we propose an improved nuclear norm maximization module to address the issue of reduced prediction diversity of unlabeled samples caused by imbalanced sample numbers between categories in remote sensing datasets.

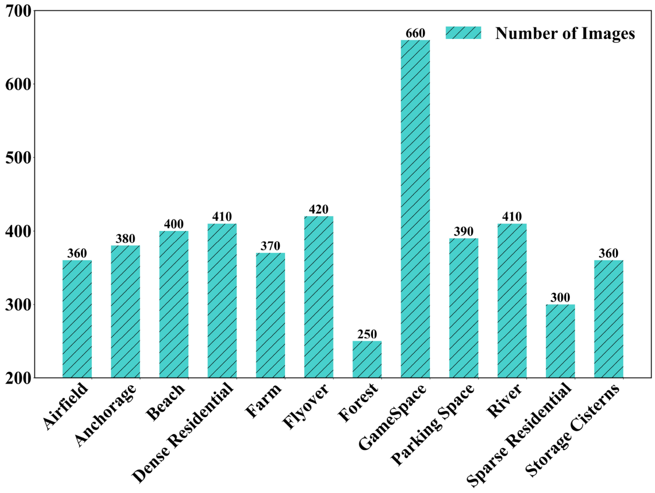


Fig. 2. Image number of different categories in AID dataset.

In general, the contributions of this article are as follows.

1) This article proposes a novel DA framework based on middle-layer feature extraction and nuclear norm maximization (MFE-NM) for remote sensing scene classification, which can effectively improve the accuracy of the model and enhance the prediction diversity of unlabeled samples.
2) The MFE module is proposed to solve the problem of feature heterogeneity caused by large intraclass differences in remote sensing images. During the model training process, middle-layer features are randomly selected and fully utilized, which can achieve domain alignment.
3) The NM module is proposed to solve the problem of reducing prediction diversity caused by entropy minimization methods when dealing with scene classification problems with imbalanced sample numbers between remote sensing image categories. By constraining the nuclear norm, the prediction discriminability and diversity of unlabeled samples are effectively improved, thereby improving the classification accuracy of the model.

The rest of this article is organized as follows. In Section II, the related work is introduced. Section III describes our proposed method in detail. Section IV introduces the dataset, experimental setup, and experimental results. The MFE-NM method is compared with existing methods and its effectiveness is verified through a large amount of experiments. Finally, Section V concludes this article.

II. Related Work

A. Unsupervised Domain Adaptation

In recent years, a large number of DA methods have been proposed, which can be roughly divided into two categories. The first is to achieve DA by measuring the distribution difference between the source domain and the target domain; the second category is mainly based on GANs. The feature representation with both domain invariance and class discriminability is obtained through the game process between the feature extractor and the domain discriminator.

Most methods based on statistical criteria use mean or higher-order moments to measure the distance between the source domain and the target domain. In [8], [18], and [25], researchers applied maximum mean discrepancy (MMD) to the DA classification. Tzeng et al. [25] proposed a deep domain confusion (DDC) by introducing a common adaptive layer to align the feature representations of the source domain and target domain, and using an additional domain confusion loss to automatically learn the feature representations during joint training between the domains, enabling the network to learn domain invariant features and optimize classification results. Long et al. [8] proposed a deep adaptation network (DAN), which is further developed on the basis of DDC. Compared with the method of only one layer of network adaptation and using single-core MMD in DDC, DAN adopts multilayer adaptation and uses multiple kernel variants of MMD, which can enhance the transfer ability of the model.

The adversarial-based method applies the idea of GANs to the field of DA. Ganin and Lempitsky [18] first applied GANs to transfer learning. In addition, [9], [10], [11], [12], [13], and other articles also try to use this method to solve the domain shift issues. In [26], they found through principal component analysis that the largest singular value represents the most important component. However, cross-domain classification problems require the proximity of the primary components and the alignment of the secondary components. Therefore, they proposed the batch spectral penalization (BSP) method. By adding the penalty term, the difference between the singular penalty will not be too large, thus effectively enhancing the feature discriminability of the model. Adversarial discriminative domain adaptation [27] is also one of the representative methods in the adversarial-based methods. This method adopts an asymmetric mapping approach, where the source domain and the target domain use two feature extractors with the same network structure but independent of each other. The target domain feature extractor uses pretrained parameters from the source domain for initialization operations, making the extracted features more domain specific.

The above methods are all based on natural scenes, but due to the fact that remote sensing scene images usually come from satellites, their image characteristics differ significantly from natural scene images. Therefore, when the above unsupervised domain adaption methods are directly applied to remote sensing scenes, there may be a decrease in classification accuracy. Therefore, it is necessary to develop specialized UDA methods tailored to the characteristics of remote sensing scenes.

B. UDA in Remote Sensing

In the field of remote sensing, many effective DA methods have also emerged. For example, Elshamli et al. [28] fully applied the DANN method in the field of remote sensing, studied the end-to-end mode of DA in the background of pixel-based remote sensing image classification, and evaluated its performance. In [29], based on the BSP method, a new multisource DA method was proposed, which maps different groups of source and target domains into a group-specific subspace using adversarial learning with metric constraints, and aligns the remaining source and target domains in the subspace. In addition, DA methods based on remote sensing scenes have been proposed in [19], [30], [31], [32], [33], and [34].

In the practical application of remote sensing images, the source domain data is sometimes inaccessible due to confidentiality or privacy considerations. At this time, the existing DA methods cannot be applied. Considering this situation, the authors in [30] and [35] proposed a novel source data generation-based universal DA model. The model distills the source domain knowledge from the pretrained model, and redescribes it as the conditional distribution of the source domain data, thereby obtaining unknown source domain information. Afterward, transferable weights are further used to distinguish between shared and private label sets for each domain. The authors in [31] and [33] take into account the current situation where remote sensing DA methods lack effective utilization of target domain information. Zheng et al. [31] proposed a DA via a task-specific classifier (DATSNET) framework that uses two neural networks as task-specific classifiers, and used task-specific decision boundary in the target domain to align the distribution of source domain and target domain features. Ma et al. [33] proposed an error-correcting boundaries mechanism with a feature adaptation metric (ECB-FAM) structure. The ECB structure can fully utilize the information of the target domain, effectively correcting classification errors when the classifier acts on the target domain, and reducing the uncertainty of difficult-to-classify samples; the FAM structure targets fuzzy features and is used to construct domain invariant features.

In the context of remote sensing, due to the fact that most of the methods based on statistical criteria only measure the distance between the source domain and the target domain through mean or variance, this simple method is difficult to achieve accurate alignment between the two domains. Since the idea of GANs was introduced into the field of DA, this method can effectively obtain the common feature representation between two domains, and has more advantages in domain alignment. However, the methods proposed above did not fully consider the significant differences in key parts of remote sensing images, and aligning the source and target domains directly may not achieve satisfactory results. At the same time, due to the imbalance in the number of images between classes, the use of entropy-based methods will reduce the diversity of class prediction in the target domain.

In response to the above issues, we propose a new network framework to achieve two goals: 1) learning the key representation of features to obtain accurate domain invariant features and

2) improving the reduction of prediction diversity of unlabeled samples caused by entropy minimization.

III. Methods

A. Preliminary Knowledge

In UDA, the source domain is usually represented as $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, containing N_s labeled samples. Represent the target domain as $D_t = \{x_i^t\}_{i=1}^{N_t}$, containing N_t unlabeled samples. The source domain and the target domain are sampled from the joint distribution $p(x^s, y^s)$ and $p(x^t, y^t)$ of features and labels, respectively. The goal of DA is to design a network that can align the distribution between the source domain and target domain, so that the classifiers trained on the labeled source domains can be applied to the unlabeled target domains. The UDA network implemented in the article uses an adversarial training method, and its structure mainly includes three parts: a feature extractor F for image feature extraction, a classifier G for categories classification, and a domain discriminator D for judging the domain to which the sample belongs, where $f = F(x)$ and $g = G(x)$ are used to represent features and classifier predictions, and $\hat{y}$ is the prediction label. Long et al. [36] proposed that when the joint distributions p and q of the source and target domains are not similar, it is difficult to achieve domain alignment by adjusting only the features. At the same time, the distribution of features often contains complex multimodal structural information, and only adjusting the features is not conducive to the training of adversarial networks. Based on these two points, it was found that classifier prediction g may contain discriminant information that can interpret multimodal structures. Therefore, modeling the feature representation f and classifier prediction g enables the network to reveal the multimodal information behind the data. f and g are modeled using a multilinear mapping method, formulated as follows:

$$h = f \otimes g. \tag{1}$$

The proposed framework is shown in Fig. 3. First, both the source domain data and the target domain data are input into the feature extractor, and while outputting the last layer, the middle-layer features output by the feature extractor are iteratively and randomly selected. Pass the randomly selected features and the last layer of features through a classifier to obtain the corresponding classifier prediction g. The feature and its corresponding classifier prediction through a multilinear mapping to obtain a joint representation h, which is used as input to the domain classifier. The entire model mainly includes two parts: MFE-NM. The next two parts will specifically introduce its mechanism of action.

B. Random Middle-Layer Feature Extraction

In the context of remote sensing, there are still significant differences in the key parts of the same category images in the source and target domains due to differences in factors, such as shooting satellites, time, and location. At this time, it is not sufficient to process only the last layer of features output by the feature extractor. According to the research in [37], even under different datasets and training tasks, the features of the first layer of the model are very similar. As the network progresses from shallow to deep, the features also shift from general to specific, commonly referred to these shallow features as fuzzy features [8]. In the field of transfer learning, previous research has often focused on aligning high-level features in CNN networks, which often contain unique information about task objects, which will cause overall alignment attenuation [37]. Compared with advanced features, fuzzy features of images have smaller receptive fields and higher image resolution. At the same time, fuzzy features are general and contain less obvious semantic information specific to task objects. If fuzzy features are aligned, the negative impact on domain alignment is relatively small [33]. This means that the network trained on the source domain dataset can also achieve satisfactory results when tested in the target domain. Therefore, for remote sensing images with large interdomain differences, using multilayer features can effectively represent invariant information, thereby improving classification accuracy.

Making full use of shallow features can obtain accurate domain invariant information, but using all feature layers in the training process not only consumes a large amount of computing resources, but may also have a negative impact on the domain transfer process. To address this issue, a random middle-layer feature extraction method is proposed. As shown in Fig. 3, the last layer of features output by the feature extractor is task-based, therefore maintaining this layer of features unchanged. On this basis, in order to fully utilize shallow features, n middle-layer features are randomly selected from the m middle-layer features output by the feature extractor during each training period, and input these n middle-layer features into a classifier to obtain their corresponding classifier predictions. The random extraction process is represented as

$$f_i = R_n \left(f_1, \ldots, f_m\right) \tag{2}$$

where R_n is an iterative random selection, f_i is the randomly extracted middle-layer features, i represents the number of feature layers, the last layer feature is represented as f_{final}, and the corresponding classifier predictions are represented as g_i and g_{final}. Both feature and classifier predictions are connected through a multilinear mapping, and the corresponding joint variables are represented as

$$h_i = \left(f_i, g_i\right) \tag{3}$$

$$h_{\text{final}} = \left(f_{\text{final}}, g_{\text{final}}\right). \tag{4}$$

At this point, the domain discriminator loss can be expressed as

$$\mathcal{L}_{\text{MFE}} = \frac{1}{N_s + N_t} \sum_{x_i \sim (\mathcal{D}_s \cup \mathcal{D}_t)} \left(L\left(D\left(h_{\text{final}}\right), d_i\right)\right.$$
$$\left. + \lambda \sum_{j=1}^{n} L\left(D\left(h_j\right), d_i\right)\right) \tag{5}$$

where $d_i (i = 0, 1)$ represents the domain label, when $i = 0$, it indicates that the corresponding sample comes from the source domain, and when $i = 1$, it indicates that the corresponding sample comes from the target domain.

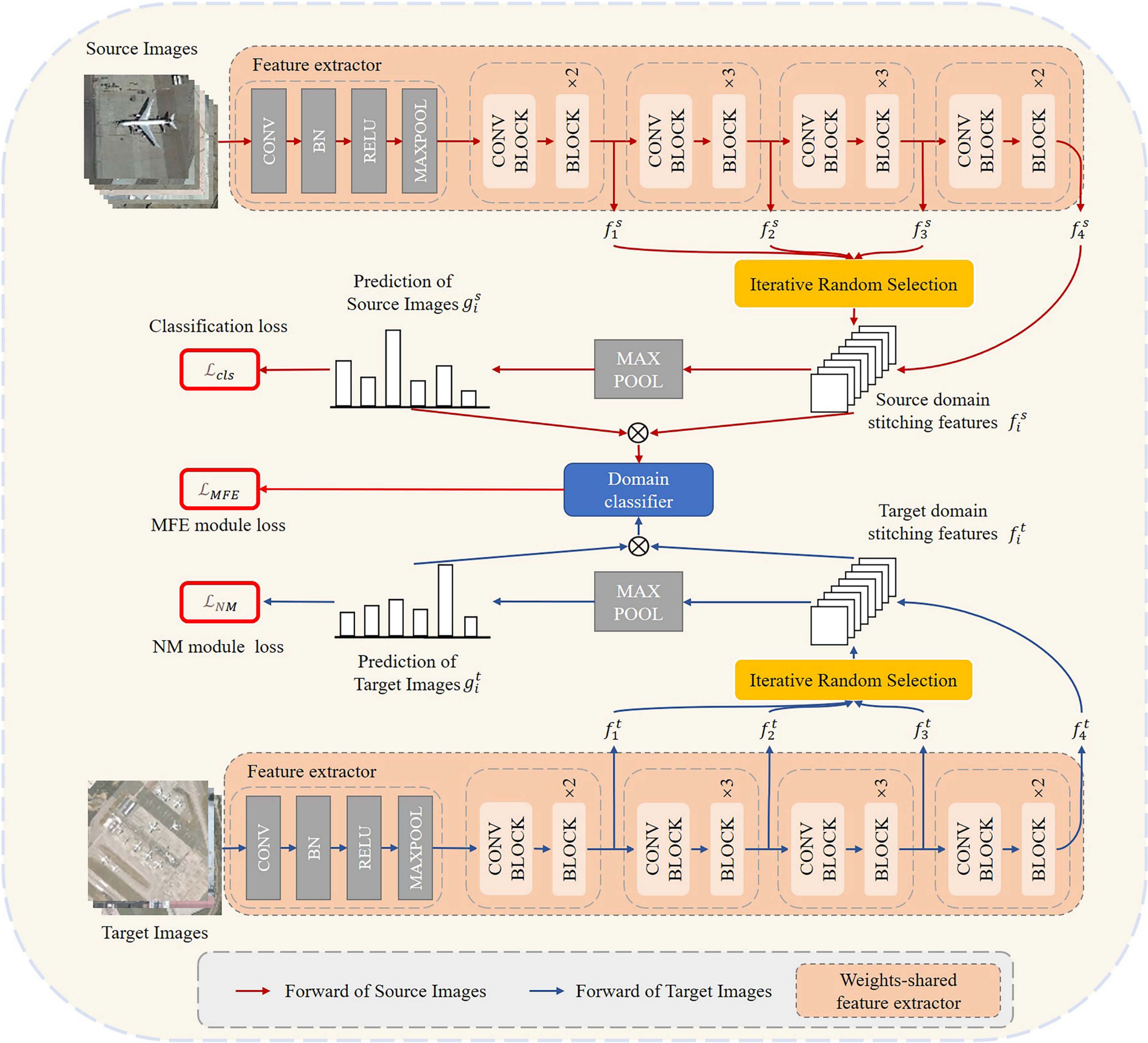


Fig. 3. Framework based on MFE-NM.

C. Nuclear Norm Maximization

In the field of remote sensing, labeling samples not only requires a lot of time, but also requires expert knowledge, so the number of labeled samples is relatively scarce. In this context, researchers endeavor to obtain a deep neural network that can be trained directly on unlabeled samples. But directly using unlabeled samples for training can lead to an increase in data density near the decision boundary, thus reducing prediction discriminability. To balance data density, most methods use entropy minimization methods. However, the method based on entropy minimization also has a certain negative effect [38], which can lead to a decrease in the number of predicted categories, that is, it is easy to judge the samples from the category with fewer samples into the category with more samples, thereby reducing the prediction accuracy, as shown in Fig. 4.

Cui et al. [23] found that the Frobenius-norm and rank of the output matrix can constrain the prediction discriminability and diversity of unlabeled data. Shannon entropy $H(P)$ is usually used to measure the uncertainty of the prediction. When $H(P)$ reaches its minimum value, the prediction uncertainty is minimized, which means the prediction discriminability is maximum. At this point, the value in the prediction matrix P is fixed, and this fixed matrix is represented as P^*. According to the derivation in [23], the F-norm of a matrix is strictly opposite to the monotonicity of Shannon entropy. Therefore, when the F-norm reaches its maximum value, the corresponding prediction matrix P reaches P^*, and $H(P)$ reaches its minimum value, it can be considered that the F-norm and $H(P)$ are equivalent in representing the prediction discriminability. The expression and upper limit of the F-norm are as follows, where L is the prediction output number of unlabeled sample data batches and

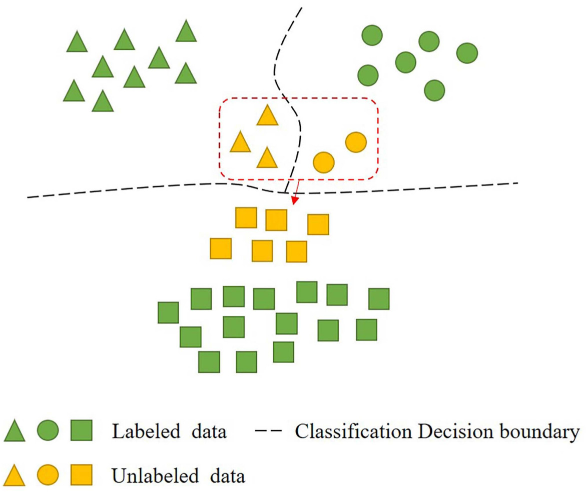


Fig. 4. Entropy minimization methods tend to classify triangles and circles with fewer sample numbers in the graph as block, resulting in a decrease in prediction diversity and classification accuracy.

K is the number of categories:

$$\begin{aligned}\|P\|_F &= \sqrt{\sum\nolimits_{i=1}^{L}\sum\nolimits_{j=1}^{K}|p_{i,j}|^2}\\ &\le \sqrt{\sum\nolimits_{i=1}^{L}\left(\sum\nolimits_{j=1}^{K}p_{i,j}\right)^2}\\ &\le \sqrt{L}.\end{aligned} \tag{6}$$

The prediction diversity can be reflected by the number of predicted categories. The overall number of predicted categories E should be the expected value of the number of categories in all prediction matrices. In an ideal state, according to the correlation principle of matrix rank, when two randomly selected row vectors in the prediction matrix belong to different categories, they can be considered linearly independent, and when they belong to the same category, they can be considered linearly correlated. Therefore, the rank of the prediction matrix can be approximately equal to the number of predicted categories.

However, due to the similarity of ground features in remote sensing images, there may be some correlation between images of different categories, so the line vectors corresponding to different categories in the prediction matrix cannot be completely linear independent. To solve this problem, we introduce a scale factor η and obtain an improved prediction matrix

$$p = \frac{\exp(z_j/\eta)}{\sum_m^K \exp(z_m/\eta)} \tag{7}$$

$$P_\eta = [p_1, \ldots p_j, \ldots p_K]^T \quad \forall j \in (1 \ldots K). \tag{8}$$

When η taking values within the range of (0,1) and $\eta \longrightarrow 0$, it can be obtained that

$$\begin{aligned}\lim_{\eta\to 0} p_{i=c} &= \lim_{\eta\to 0}\frac{\exp(z_i/\eta)}{\sum_m^K \exp(z_m/\eta)}\\ &= \lim_{\eta\to 0}\frac{1}{1+\sum_{m\neq i}^K \exp((z_m - z_i)/\eta)} = 1\end{aligned} \tag{9}$$

$$\begin{aligned}\lim_{\eta\to 0} p_{i\neq c} &= \lim_{\eta\to 0}\frac{\exp(z_i/\eta)}{\sum_m^K \exp(z_m/\eta)}\\ &= \lim_{\eta\to 0}\frac{1}{1+\sum_{m\neq i}^K \exp((z_m - z_i)/\eta)} = 0\end{aligned} \tag{10}$$

where c represents the class with the highest probability of class prediction in the row vector. The row vectors corresponding to images belonging to different categories each obtain the maximum probability value of 1 at the corresponding category. At this point, the two vectors are completely linearly independent. If the value of the scale factor η is between (0,1), the linear independence between the row vectors belonging to different categories in the prediction matrix can be higher, so as to be closer to the ideal state. At this time, the rank of the prediction matrix can be used to approximate the number of predicted categories E. For the classic low-rank matrix completion optimization problem, a common method is to use nuclear norm $\|P\|_*$ to approximate the rank of the matrix, where all singular values of the matrix are required to be less than or equal to 1 [39]. By combining the (6), the convex envelope of the prediction matrix rank can be obtained. According to [37], the relationship between the F-norm and nuclear norm is as follows:

$$\|P\|_F \le \|P\|_* \le \sqrt{Q}\,\|P\|_F. \tag{11}$$

Among them, $Q = \min(L, K)$, where the nuclear norm can simultaneously constrain the discriminability and diversity of predictions. When the nuclear norm reaches its maximum value, the discriminability and diversity of predictions are also maximized. The loss function is expressed as

$$\mathcal{L}_{\text{NM}} = -\frac{1}{L_t}\|P_\eta\|_*. \tag{12}$$

D. MFE-NM Framework

In summary, our proposed MFE-NM method consists of two parts: MFE-NM. The MFE module can accurately capture domain invariant information in DA problems by fully utilizing middle-layer features. The NM module improves the diversity and discriminability of prediction by constraining the nuclear norm of the prediction matrix. In the actual training process, the source domain data is input into the feature extractor and category classifier to obtain classification loss

$$\mathcal{L}_{\text{cls}} = \frac{1}{N_s}\sum_{x_i\sim\mathcal{D}_s} -\sum_{j=1}^{K} y_{x_i,j}\log\left(G\left(F\left(x_i\right)\right)\right) \tag{13}$$

where $y_{x_i,j}$ represents the true value probability of class K on the source domain sample. In order to train the network as a whole, we synchronously optimize multiple losses. The loss of the middle-layer feature extraction module and the nuclear norm maximization module can be determined through the parameter λ_{MFE} and λ_{NM} combined with classification loss, the overall loss function is represented as follows:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{MFE}}\mathcal{L}_{\text{MFE}} + \lambda_{\text{NM}}\mathcal{L}_{\text{NM}}. \tag{14}$$

Algorithm 1 demonstrates the main steps of training the MFE-NM model.

Algorithm 1: MFE-NM Model Training Process.

```
Data: Labeled source domain sample set $\{\mathbf{x}^\mathbf{s}, \mathbf{y}^\mathbf{s}\}$ and
      unlabeled target domain sample set $\{\mathbf{x}^\mathbf{t}\}$ ,
      feature extractor $\mathbf{F}$ , classifier $\mathbf{G}$ and domain
      discriminator $\mathbf{D}$, the feature representation $\mathbf{f}_\mathbf{i}^\mathbf{s}$
      and $\mathbf{f}_\mathbf{i}^\mathbf{t}$, where $\mathbf{f}_\mathbf{final}^\mathbf{s}$ and $\mathbf{f}_\mathbf{final}^\mathbf{t}$ represents the last
      layer of features, $\mathbf{g}_\mathbf{i}^\mathbf{s}$ and $\mathbf{g}_\mathbf{i}^\mathbf{t}$ are the
      corresponding classifier outputs.
Result: Classification accuracy of $\mathbf{x}^\mathbf{t}$.
for $i$ in epochs do
    $\mathbf{x}^\mathbf{s}$ and $\mathbf{x}^\mathbf{t}$ are passed through $\mathbf{F}$ and obtain $\mathbf{f}_\mathbf{i}^\mathbf{s}$ and
      $\mathbf{f}_\mathbf{i}^\mathbf{t}$.
    Pass $\mathbf{f}_\mathbf{i}^\mathbf{s}$ and $\mathbf{f}_\mathbf{i}^\mathbf{t}$ through $\mathbf{G}$ to obtain the
      corresponding $\mathbf{g}_\mathbf{i}^\mathbf{s}$ and $\mathbf{g}_\mathbf{i}^\mathbf{t}$.
    Splicing $\mathbf{f}_\mathbf{i}^\mathbf{s}$ and $\mathbf{f}_\mathbf{i}^\mathbf{t}$ to obtain $\mathbf{f}_\mathbf{i}$, and performing the
      same operation on $\mathbf{g}_\mathbf{i}^\mathbf{s}$ and $\mathbf{g}_\mathbf{i}^\mathbf{t}$ to obtain $\mathbf{g}_\mathbf{i}$.
    Randomly select from $\mathbf{f}_\mathbf{i}$ other than $\mathbf{f}_\mathbf{final}$.
    Multiply the $\mathbf{f}_\mathbf{final}$ and $\mathbf{g}_\mathbf{final}$ matrices, and
      perform the same operation on the randomly
      selected middle-layer features $\mathbf{f}_\mathbf{i}$ and its
      corresponding $\mathbf{g}_\mathbf{i}$ in the previous step.
    Enter the joint representation into $\mathbf{D}$ and calculate
      $\mathcal{L}_{MFE}$ according to Eq.(5)
    Calculate the nuclear norm of $\mathbf{g}_\mathbf{4}^\mathbf{t}$ to obtain the loss
      $\mathcal{L}_{NM}$ according to Eq.(12).
    Calculate classification loss $\mathcal{L}_{cls}$ according to
      Eq.(13).
    Obtain the total loss $\mathcal{L}_{all}$ according to Eq.(14), and
      update the network.
    Obtain the classification accuracy of the target
      domain.
end
```

IV. Experiments

In this section, specific information, such as the dataset and experimental settings, used in the experimental process was introduced, and a large number of experiments were conducted to demonstrate the effectiveness of the proposed method.

A. Dataset Description

Due to the lack of a remote sensing dataset specialized for transfer learning research, in order to evaluate the method, three commonly used remote sensing datasets are used to form a subdataset with 12 common classes for testing. The three remote sensing datasets used are UCM [20], AID [21], and NWPU [22]. The UCM dataset includes 21 types of earth scenes, each consisting of 100 images with a pixel size of 256×256, with a ground resolution of 0.3 m. The images in this dataset are all extracted from the USGS National Map Urban Area Imagery series. The AID dataset contains 30 types of scenes, the number of each type is about 220–420, and the total number of all types of images is 10 000. Image pixel size is 600×600, with ground resolution ranging from 0.5 to 8 m. These pictures were collected from Google Earth and released by Wuhan University and Huazhong University of Science and Technology. The NWPU dataset contains a total of 31 500 images with a pixel size of 256×256, covering a total of 45 categories. Each category contains 700 images, and the ground resolution ranges from 30 to 0.2 m.

The categories of the constructed remote sensing cross domain dataset are airfield, anchorage, beach, dense residential, farm, flyover, forest, game space, parking space, river, sparse residential, and storage cities. Due to the fact that different remote sensing datasets are labeled by different experts, categories may have different names, and such images need to be merged, such as the ground track field and stage in NWPU to form the game space, circular farms, and rectangular farms to form the farm, and airplane and airport to form the airfield. Table I gives the formed remote sensing cross domain dataset and its detailed information. Abbreviate the three datasets as A (AID), U (UCM), and N (NWPU). During the experiment, two of the three remote sensing datasets were randomly selected as the source and target domains, respectively. The proposed model was trained on the source domain dataset and tested on the target domain dataset, resulting in six adaptive tasks: A→U, U→A, N→U, U→N, N→A, and A→N.

B. Experimetal Setting

We use the Pytorch framework to implement our model on NVIDIA TITAN RTX GPU. In order to train faster, a pretrained ResNet-50 network on ImageNet is used as a feature extractor. Due to the different sizes of images in different datasets, the size of the input images is unified to 224 × 224 and the batch size is set to 32. The training process uses an Adam optimizer with a learning rate set to 10^{-4}. During the training process, 80% of the source domain data is divided into training sets and the remaining 20% is divided into test sets. Simultaneously divide 50% of the target domain data into training sets [40].

C. Hyperparameters Determination

It is necessary to determine the number of middle-layer features used in the MFE module. In the experiment, resnet-50 is used as the feature extractor, which includes three additional shallow layer features in addition to the last layer feature output. The shallow features are represented as $l_i (i = 1, 2, 3)$. In order to determine the number of randomly selected middle-layer features, fixed middle-layer feature extraction, and random middle-layer feature extraction were performed on the cross domain dataset. The experimental results obtained are given in Table II. According to the results in the table, it can be found that using fixed middle-layer features is difficult to achieve optimal classification results for multiple cross domain datasets simultaneously. Compared with randomly selecting one middle-layer feature and aligning all middle-layer features, randomly selecting two middle-layer features yielded the best experimental results. Therefore, the method of randomly selecting two middle-layer features was ultimately chosen.

In the NM module, it is necessary to determine the value of the hyperparameter scale factor η. According to Section III-C, η is taken at intervals of 0.05 within (0.6,1), and experiments are conducted on six cross-domain tasks. The relationship curve

TABLE I
DETAILED INFORMATION AND COMMON CATEGORIES OF THE THREE REMOTE SENSING DATASETS USED IN THE MFE-NM METHOD

Datasets	Classes	Number of images per class	Total number of images	Spatial resolution (m)	Image size	Data source	Public class
UCM	21	100	2100	0.3	256×256	USGS	airfield, anchorage, beach, dense residential, farm, flyover, forest, game space, parking space, river, sparse residential and storage cities
AID	30	220−420	10000	0.5−8	600×600	Google Earth	
NWPU	45	700	31500	0.2−30	256×256	Google Earth	

TABLE II
CROSS DOMAIN CLASSIFICATION RESULTS WHEN SELECTING DIFFERENT MIDDLE-LAYER FEATURES, INCLUDING A SINGLE MIDDLE-LAYER FEATURE, TWO MIDDLE-LAYER FEATURES, AND THREE MIDDLE-LAYER FEATURES

Methods	A→U	U→A	N→U	U→N	N→A	A→N
Backbone	81.63	57.68	92.25	51.39	91.21	83.32
Backbone+l_1	**87.58**	62.02	90.17	52.56	90.96	84.04
Backbone+l_2	81.58	60.11	**92.25**	53.32	90.25	**84.53**
Backbone+l_3	82.42	62.34	91.83	**55.24**	**91.21**	84.38
Backbone+l_i	83.83	**64.61**	91.92	51.50	90.87	82.97
Backbone+l_1+l_2	**88.50**	60.93	91.58	57.31	91.40	83.61
Backbone+l_1+l_3	87.83	59.89	90.67	52.14	90.53	83.76
Backbone+l_2+l_3	81.42	65.61	91.67	54.62	90.91	83.37
Backbone+l_i+l_j	87.58	**65.67**	**93.83**	**58.54**	**91.55**	**84.09**
Backbone+l_1+l_2+l_3	84.83	60.31	92.75	54.14	90.51	84.24

The bold part represents the best accuracy result when using single and two middle-layer features respectively.

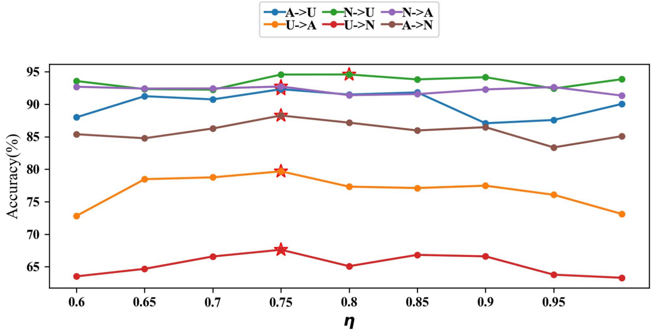


Fig. 5. Result of accuracy changing with the hyperparameter η.

TABLE III
COMPARATIVE EXPERIMENTS WITH AND WITHOUT THE SCALE FACTOR

Methods	A→U	U→A	N→U	U→N	N→A	A→N
Backbone	81.63	57.68	92.25	51.39	91.21	83.32
Backbone+NM*	90.05	73.12	93.86	63.29	91.34	85.09
Backbone+NM	**91.44**	**75.97**	**95.03**	**67.33**	**92.63**	**85.24**

*Indicates the situation where no scale factor is added.
The bold part represents the best classification accuracy.

between prediction accuracy and η is shown in Fig. 5. When $\eta = 0.75$, among the five cross domain tasks except for N→U, the prediction accuracy reached the maximum value. Based on this, we chose $\eta = 0.75$ for the experiment. To verify the effectiveness of the scale factor, we conducted comparative experiments with and without the scale factor on six DA tasks. The experimental results are given in Table III. It can be seen

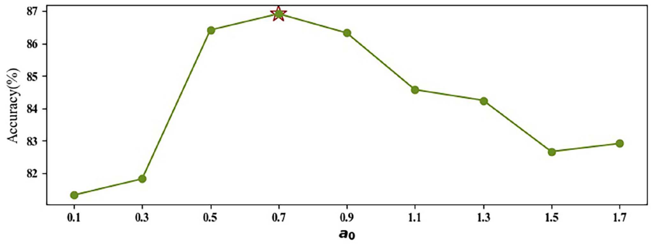


Fig. 6. Result of accuracy changing with the hyperparameter a_0.

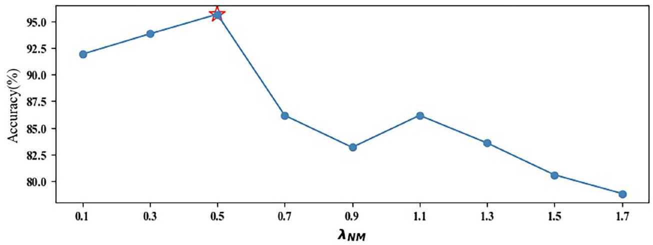


Fig. 7. Result of accuracy changing with the hyperparameter λ_{NM}.

that after adding the scale factor, the classification performance has been improved by $0.15\% - 4.04\%$.

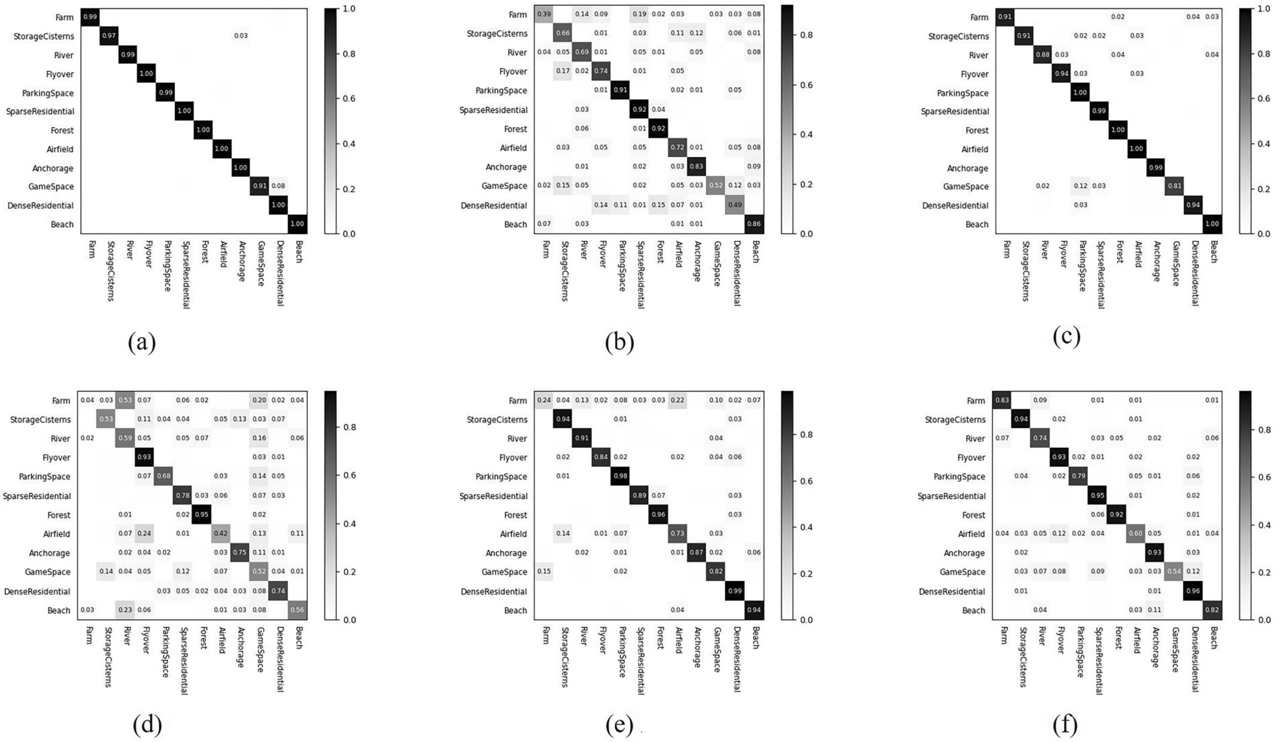
There are two hyperparameters, λ_{NM} and λ_{MFE} in the method. By adjusting the two hyperparameters on A→U, appropriate hyperparameter values can be obtained. Set λ_{MFE} in form of $\frac{a_0\times(\text{epoch}+1)}{\text{epoches}}$, which means that the variable parameter in λ_{MFE} is a_0. The transformation range of the hyperparameter is set to 0.1–1.7, and the variation interval is 0.2. The experimental results are shown in Figs. 6 and 7, when $a_0 = 0.7$ and $\lambda_{\text{NM}} = 0.5$, the accuracy reaches its peak, so we will set these two fixed values in the following experiments.

D. Ablation Study

To verify the effectiveness of the above two modules, we conducted ablation experiments on cross domain remote sensing datasets, including the backbone, MFE module, and NM module. The final experimental results are presented in Table IV. According to the results in the table, it can be seen that when using the backbone, the experimental accuracy is the lowest. When adding MFE module and NM module respectively based on the backbone, and conducting data transfer on six cross domain datasets, the classification accuracy is improved. Compared with using only a single module at the same time, when two modules are combined, the improvement effect of the model is better, with the highest being 20.81% accuracy improvement was achieved on the U→A. The classification

TABLE IV
ABLATION STUDY OF OUR METHOD

Methods	A → U	U → A	N → U	U → N	N → A	A → N
Backbone	81.63±1.38	57.68±1.10	92.25±1.20	51.39±1.04	91.21±0.87	83.32±1.27
Backbone+MEF	85.38±0.85	65.70±2.02	92.39±0.84	56.88±0.53	91.55±0.98	84.68±0.97
Backbone+NM	91.44±0.36	75.97±0.60	94.83±1.23	67.33±1.98	92.63±0.73	85.24±1.15
our method	**94.28±1.25**	**78.49±1.61**	**95.03±0.97**	**68.59±1.49**	**93.03±0.36**	**85.57±0.52**



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 8. Confusion matrices of MFE-NM module. (a) A → U. (b) U → A. (c) N → U. (d) U → N. (e) N → A. (f) A → N.

TABLE V
CLASSIFICATION ACCURACY (%) OF EACH CATEGORY ON THE A → U

Methods	Farm	Storage Cisterns	River	Flyover	Parking Space	Sparse Residential	Forest	Airfield	Anchorage	Game Space	Dense Residential	Beach	Average
DeepCORAL	52	64	64	28	98	76	100	68	75	80	48	100	71.08
DDC	89	69	73	21	97	53	94	44	63	82	30	95	67.50
DAN	77	58	63	77	95	69	85	68	89	85	60	96	76.83
DANN	90	52	66	88	98	66	82	92	85	93	62	95	80.75
GVB	90	68	74	68	100	84	100	**98**	86	**100**	78	100	87.17
CGDM	71	63	73	39	100	64	100	76	100	91	30	100	75.58
CDTrans	89	64	29	39	100	75	89	97	100	66	70	100	76.50
MFE-NM(ours)	**91**	**85**	**90**	**97**	**100**	**87**	**100**	94	**100**	85	**92**	**100**	**93.42**

The bold part represents the best classification accuracy.

accuracy of each category in the MFE-NM module is shown in Fig. 8 by the confusion matrices.

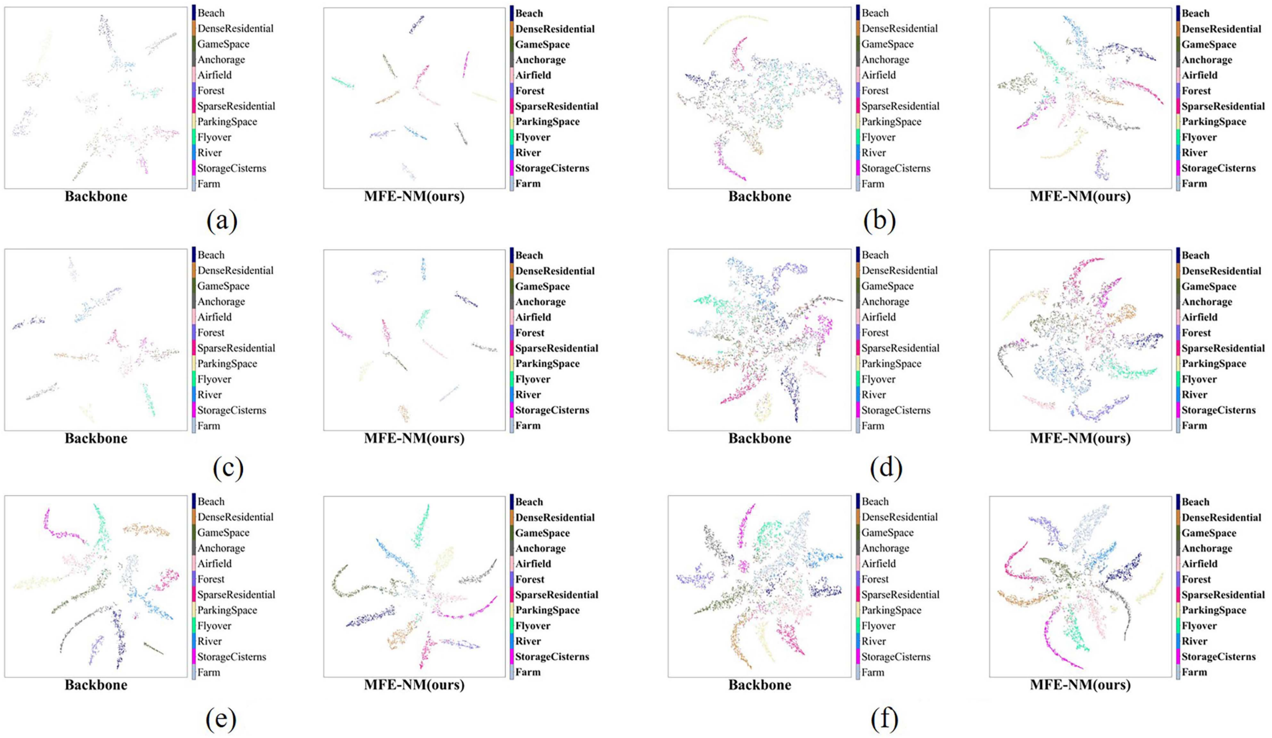
E. Comparison With Other Methods

To demonstrate the progressiveness of the MFE-NM method, it is compared with some recent classical DA methods, including DANN [18], DDC [25], DAN [8], DeepCORAL [41], gradually vanishing bridge (GVB) [42], cross-domain gradient discrepancy minimization (CGDM) [43] and cross-domain transformer (CDTrans) [44]. According to the original paper of these methods, hyperparameters are set to obtain the best performance. Experiments were conducted using these methods on the cross domain remote sensing dataset produced, and the obtained experimental results are given in Tables V and VI. Table V gives the prediction accuracy for each category in

TABLE VI
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON THE CROSS DATASETS

Methods	A→U	U→A	N→U	U→N	N→A	A→N
DeepCORAL	73.42	65.33	83.5	52.78	81.62	82.67
DDC	69.00	54.42	78.58	54.74	71.63	80.65
DAN	78.08	77.75	76.58	56.23	82.99	79.10
DANN	83.08	72.46	84.58	65.71	89.89	85.60
GVB	89.33	57.37	84.25	58.06	88.89	85.06
CGDM	75.58	61.57	80.67	47.90	92.82	83.28
CDTrans	76.50	67.30	91.84	67.70	92.53	85.40
MFE-NM(ours)	**94.28**	**78.49**	**95.03**	**68.59**	**93.03**	**85.57**

The bold values indicate the best performance.



(a) (b) (c) (d) (e) (f)

Fig. 9. T-SNE visualization results of backbone and MFE-NM (ours). (a) A→U. (b) U→A. (c) N→U. (d) U→N. (e) N→A. (f) A→N.

the MFE-NM method and other methods on the A→U task. The MFE-NM method achieves optimal results in all categories except for airfield and game space, and its average accuracy is superior to other methods. Table VI gives the comparison results between the MFE-NM method and other methods on six cross domain classification tasks. It can be seen that MFE-NM method is significantly superior to previous methods. Compared with DA methods based on distance measurement, such as DAN and DeepCORAL, the performance improvement is more significant. Compared with other adversarial-based DA methods, such as DANN, although our method is slightly lower in A→N task, the gap is only 0.03, and on the whole, the MFE-KM method is better than the DANN method, especially in the A→U task, the performance improvement is 11.20%. Compared with the recent DA method CDTrans, MFE-NM method also exhibits certain advantages. This indicates that our proposed MFE-NM model can fully utilize feature information, achieve alignment between source and target domains, and effectively solve the problem of DA in remote sensing scene classification.

F. Visual Analysis

The feature clustering effect can effectively reflect the classification results, and the distributed stochastic neighbor embedding (t-SNE) [45] technique is a very effective feature visualization display method [46]. Therefore, t-SNE is used to visualize the distribution of the target dataset. T-SNE is a tool for dimensionality reduction and visualization of high-dimensional data. We used this tool to visualize the backbone and our proposed MFE-NM method on cross domain datasets, and the results are shown in Fig. 9. It can be seen that the MFE-NM model can effectively cluster sample features. As shown Fig. 9(a), the classification boundary between the airfield class and other classes in the backbone is relatively fuzzy, while in the MFE-NM

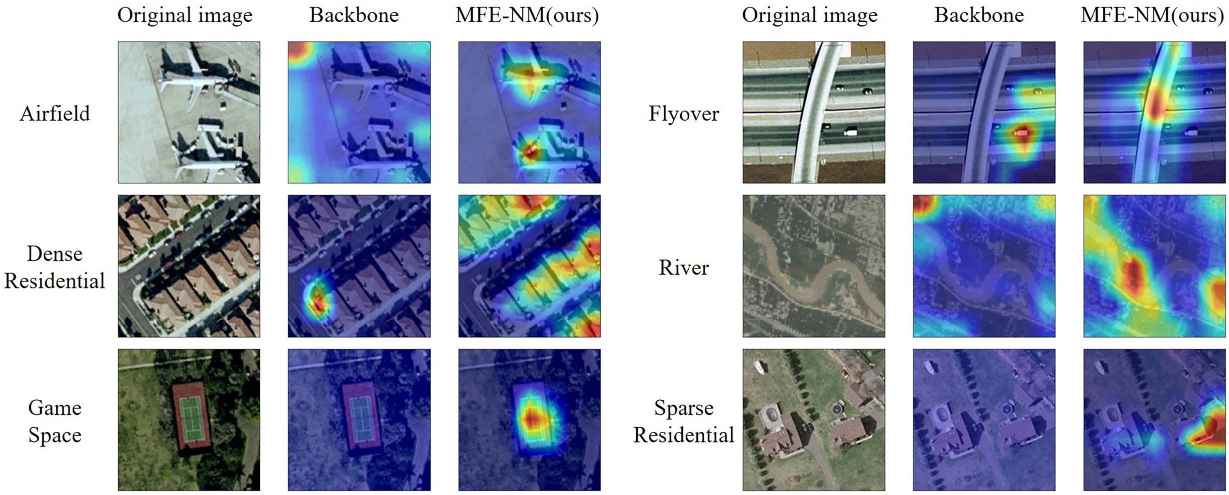


Fig. 10. Cam results of six types of images in the public data category. From left to right are the original image, the CAM corresponding to the backbone, and the CAM corresponding to the MFE-NM.

method, the sample feature aggregation degree is higher and the classification boundary is clearer.

Moreover, the Grad-class activation map (cam) [47] method is also used for the visual interpretation of images. After inputting the original image, a CAM can be obtained, which is a thermal map that generates class activation for the input image and can be understood as the contribution distribution to the prediction output. The higher the score, the higher the response of the corresponding position in the original image to the deep neural network, indicating a higher contribution to the category. In the image, the color tends to be closer to deep red. To observe the importance of different positions in the image sample and the effectiveness of our proposed model, we conducted experiments on the remote sensing cross domain datasets. As shown in Fig. 10, it can be clearly seen that after the use of MFE-NM method, the range of the effective part in the cam map is expanded and its contribution to the prediction output is improved, which proves that MFE-NM method can effectively capture key information in remote sensing images and have a positive impact on DA.

V. Conclusion

Because the data of different remote sensing datasets may be quite different, there may be DA problems in the actual application process. To address the issue of significant differences in data distribution between different remote sensing datasets, we propose a UDA method based on adversarial learning, namely, MFE-NM. The MFE module is proposed to address the issue of feature heterogeneity caused by large intraclass differences in remote sensing images, which effectively utilizes middle-layer features to achieve alignment between the source and target domains. Simultaneously using the NM module to solve the problem of reduced prediction diversity when applying entropy minimization methods to remote sensing datasets with imbalanced sample numbers between categories. Evaluated on the cross domain remote sensing dataset produced, a large number of experimental results have proven that MFE-NM method can effectively capture key information in remote sensing images, obtain domain invariant features, and effectively improve the prediction diversity of unlabeled samples. Verified the effectiveness and competitiveness of the MFE-NM model.

References

[1] Y. Li, D. Kong, Y. Zhang, R. Chen, and J. Chen, “Representation learning of remote sensing knowledge graph for zero-shot remote sensing image scene classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1351–1354, doi: 10.1109/IGARSS47720.2021.9553667.

[2] Z. Sha and J. Li, “MITformer: A multiinstance vision transformer for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 6510305, doi: 10.1109/LGRS.2022.3176499.

[3] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, “Remote sensing image scene classification based on an enhanced attention module,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1926–1930, Nov. 2021, doi: 10.1109/LGRS.2020.3011405.

[4] B. Zhao, Y. Zhong, and L. Zhang, “Hybrid generative/discriminative scene classification strategy based on latent dirichlet allocation for high spatial resolution remote sensing imagery,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 196–199, doi: 10.1109/IGARSS.2013.6721125.

[5] J. Wang, Y. Zhong, Z. Zheng, A. Ma, and L. Zhang, “RSNet: The search for remote sensing deep neural networks in recognition tasks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2520–2534, Mar. 2021, doi: 10.1109/TGRS.2020.3001401.

[6] Y. Liu, X. Kang, Y. Huang, K. Wang, and G. Yang, “Unsupervised domain adaptation semantic segmentation for remote-sensing images via covariance attention,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 6513205, doi: 10.1109/LGRS.2022.3189044.

[7] M. Xu, M. Wu, K. Chen, C. Zhang, and J. Guo, “The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data,” *Remote Sens.*, vol. 14, no. 17, Sep. 2022, Art. no. 4380, doi: 10.3390/rs14174380.

[8] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[9] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 95–104, doi: 10.1109/CVPR.2017.18.

[10] Y. Zhang, H. Tang, K. Jia, and M. Tan, “Domain-symmetric networks for adversarial domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5026–5035, doi: 10.1109/CVPR.2019.00517.

[11] J. Wang and J. Jiang, "Conditional coupled generative adversarial networks for zero-shot domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3374–3383, doi: 10.1109/ICCV.2019.00347.

[12] Q. Chen, Y. Liu, Z. Wang, I. Wassell, and K. Chetty, "Re-weighted adversarial adaptation network for unsupervised domain adaptation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7976–7985, doi: 10.1109/CVPR.2018.00832.

[13] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1498–1507, doi: 10.1109/CVPR.2018.00162.

[14] Y. Wan, Y. Zhong, A. Ma, J. Wang, and L. Zhang, "MAE-Net: A micro network architecture evolutionary search method for remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 2223–2226, doi: 10.1109/IGARSS46834.2022.9883201.

[15] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 746–750, Apr. 2021, doi: 10.1109/LGRS.2020.2982783.

[16] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*.

[17] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2715–2724, doi: 10.1109/CVPR.2019.00283.

[18] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[19] S. Zhu, F. Luo, B. Du, and L. Zhang, "Adversarial fine-grained adaptation network for cross-scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2369–2372, doi: 10.1109/IGARSS47720.2021.9554195.

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

[21] G. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[22] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, doi: 10.1109/JPROC.2017.2675998.

[23] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Toward discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3940–3949, doi: 10.1109/CVPR42600.2020.00400.

[24] H. Bai, J. Cheng, Y. Su, S. Liu, and X. Liu, "Calibrated focal loss for semantic labeling of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6531–6547, Aug. 2022, doi: 10.1109/JSTARS.2022.3197937.

[25] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, U. Lowell, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[26] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs discriminability: Batch spectral penalization for adversarial domain adaptation," in *Proc. Int. Conf. Mach Learn.*, 2019, pp. 1081–1090.

[27] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971, doi: 10.1109/CVPR.2017.316.

[28] A. Elshamli, G. W. Taylor, A. Berg, and S. Areibi, "Domain adaptation using representation learning for the classification of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4198–4209, Sep. 2017, doi: 10.1109/JSTARS.2017.2711360.

[29] C. Ren, Y. Liu, X. Zhang, and K. Huang, "Multi-source unsupervised domain adaptation via pseudo target domain," *IEEE Trans. Image Proc.*, vol. 31, pp. 2122–2135, Feb. 2022, doi: 10.1109/TIP.2022.3152052.

[30] Q. Xu, Y. Shi, and X. Zhu, "Universal domain adaptation without source data for remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 5341–5344, doi: 10.1109/IGARSS46834.2022.9884889.

[31] Z. Zheng, Y. Zhong, Y. Su, and A. Ma, "Domain adaptation via a task-specific classifier framework for remote sensing cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5620513, doi: 10.1109/TGRS.2022.3151689.

[32] J. Zhang, J. Liu, L. Shi, B. Pan, and X. Xu, "An open set domain adaptation network based on adversarial learning for remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1365–1368, doi: 10.1109/IGARSS39084.2020.9323944.

[33] C. Ma, D. Sha, and X. Mu, "Unsupervised adversarial domain adaptation with error-correcting boundaries and feature adaption metric for remote-sensing scene classification," *Remote Sens.*, vol. 13, no. 7, pp. 1270, Mar. 2021, doi: 10.3390/rs13071270.

[34] B. Fang, R. Kou, L. Pan, and P. Chen, "Category-sensitive domain adaptation for land cover mapping in aerial scenes," *Remote Sens.*, vol. 11, no. 22, pp. 2631, Nov. 2019, doi: 10.3390/rs11222631.

[35] Q. Xu, Y. Shi, X. Yuan, and X. X. Zhu, "Universal domain adaptation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4700515, doi: 10.1109/TGRS.2023.3235988.

[36] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1647–1657.

[37] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Int. Conf. Neural Inf. Proc. Syst.*, 2014, pp. 3320–3328.

[38] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Fast batch nuclear-norm maximization and minimization for robust domain adaptation," 2021, *arXiv:2107.06154*.

[39] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Aug. 2010, doi: 10.1137/070697835.

[40] N. Makkar, L. Yang, and S. Prasad, "Adversarial learning based discriminative domain adaptation for geospatial image analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 150–162, Dec. 2022, doi: 10.1109/JSTARS.2021.3132259.

[41] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[42] S. Cui, S. Wang, J. Zhuo, C. Su, Q. Huang, and Q. Tian, "Gradually vanishing bridge for adversarial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12452–12461, doi: 10.1109/CVPR42600.2020.01247.

[43] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu, "Cross-domain gradient discrepancy minimization for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3936–3945, doi: 10.1109/CVPR46437.2021.00393.

[44] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, and R. Jin, "CDTrans: Cross-domain transformer for unsupervised domain adaptation," 2022, *arXiv:2109.06165*.

[45] L. Van Der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[46] Y. Zhao, S. Li, C. H. Liu, Y. Han, H. Shi, and W. Li, "Domain adaptive remote sensing scene recognition via semantic relationship knowledge transfer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 2001013, doi: 10.1109/TGRS.2023.3267149.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

Ruitong Du received the B.S. degree in electronic information engineering, in 2021, from the Beijing Institute of Technology, Beijing, China, where she is currently working toward the master's degree in electronic information engineering.

Her research focuses on remote sensing scene classification.

Guoqing Wang (Graduate Student Member, IEEE) received the B.S. degree in electronic information engineering, in 2020, from Beijing Institute of Technology, Beijing, China, where he is currently working toward the Ph.D. degree in information and communication engineering.

His research interests include remote sensing scene classification, change detection, and real-time processing.

Ning Zhang received the B.S. degree in electronic information engineering from the Wuhan University of Technology, Wuhan, China, in 2018. He is currently working toward the Ph.D. degree in information and communication engineering with the Beijing Institute of Technology, Beijing, China.

His research interests are spaceborne real-time information processing technology and remote sensing image processing.

Liang Chen received the B.S. degree in information engineering and Ph.D. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2003 and 2008, respectively.

He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology. His research interests include spaceborne real-time information processing technology and remote sensing image processing.

Wenchao Liu received the Ph.D. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2019.

He is currently an Assistant Professor with the Beijing Institute of Technology. His research focuses on remote sensing image processing.