ZHAO, C., QIN, B., FENG, S., ZHU, W., ZHANG, L. and REN, J. 2022. An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification. *IEEE transactions on geoscience and remote sensing* [online], 60, article 5546216. Available from: https://doi.org/10.1109/TGRS.2022.3230378

An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification.

ZHAO, C., QIN, B., FENG, S., ZHU, W., ZHANG, L. and REN, J.

2022

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

OpenAIR @RGU

This document was downloaded from https://openair.rgu.ac.uk

SEE TERMS OF USE IN BOX ABOVE

DISTRIBUTED UNDER LICENCE

An Unsupervised Domain Adaptation Method Towards Multi-level Features and Decision Boundaries for Cross-Scene Hyperspectral Image Classification

Chunhui Zhao, Boao Qin, *Student Member, IEEE*, Shou Feng, *Member, IEEE*, Wenxiang Zhu, Lifu Zhang, *Senior Member, IEEE*, and Jinchang Ren, *Senior Member, IEEE*

***Abstract*—Despite success in the same-scene hyperspectral image classification (HSIC), for the cross-scene classification, samples between source and target scenes are not drawn from the independent and identical distribution, resulting in a significant performance drop. To tackle this issue, a novel unsupervised domain adaptation (UDA) framework towards multi-level features and decision boundaries (ToMF-B) is proposed for the cross-scene HSIC, which can align task-related features and learn task-specific decision boundaries in parallel. Based on the maximum classifier discrepancy, a two-stage alignment scheme is proposed to bridge the interdomain gap and generate discriminative decision boundaries. In addition, to fully learn task-related and domain-confusing features, a CNN and Transformer-based multi-level features extractor (generator) is developed to enrich the feature representation of two domains. Furthermore, to alleviate the harmless even the negative transfer to UDA caused by task-irrelevant features, a task-oriented feature decomposition method is leveraged to enhance the task-related features while suppressing task-irrelevant features, enabling the aligned domain-invariant features to explicitly improve the classification. Extensive experiments on three cross-scene HSI benchmarks have validated the effectiveness of the proposed framework.**

***Index Terms*—Hyperspectral image, cross-scene classification, unsupervised domain adaptation, task-specific, task-irrelevant.**

I. Introduction

HYPERSPECTRAL images (HSIs), delivering advantages stemming from continuous spectrum data and rich spatial information [1], [2], span a broad range of applications [3], such as environmental monitoring [4], ecological science [5] and smart city [6]. Similar to dense prediction tasks, hyperspectral image classification (HSIC), hyperspectral image classification (HSIC) aims at assigning one of the predefined categories to each pixel [7], which is the most common downstream task in hyperspectral earth observation or remote sensing. In the past decades, an impressive number of HSIC methods for the same scene have been investigated. These HSIC methods for the same scene generally follow a common assumption, that is, the training and test data are drawn from independent and identical distribution within the same HSI. These methods can be divided as traditional machine learning category, such as support vector machine (SVM) [8], [9], sparse representation [10], [11] and clustering-based methods [12], [13], and later deep learning-based (DL-based) category, such as multilayer perceptron (MLP) based [14], [15], convolutional neural network (CNN) based [16], [17], graph neural network (GNN) based [18], [19] and Transformer-based methods [20], [21]. The superiority of these supervised and semi-supervised machine learning based methods has been demonstrated on many HSIC datasets. However, most mainstream DL-based methods usually require big datasets to achieve excellent performance, but collecting and annotating plenty of samples for each new sense or each new task are costly and time-consuming for the HSIC [22], [23].

The remarkable success of aforementioned methods benefits from the massive annotated data and same distribution between training and test data, and it will yield a significant performance drop when the model lacks enough labeled training data or generalizes to another unlabeled scene. But in practice, it is often encountered that training data and testing data are originated from different scenes in realistic applications. This is a new challenge task called cross-scene HSIC, aiming to train a classifier on source scenes yet test it on target scenes. Therefore, it is natural and reasonable to explore transferable models for the cross-scene HSIC. How to transfer the same or task-related knowledge to other scenes or tasks is crucial for addressing the problem of cross-scene HSIC. In the DL field, transferability lies at the core of the whole lifecycle of deep learning. The most commonly used task-generic transfer paradigm of computer vision (CV) tasks and natural language processing (NLP) tasks is that pre-trained on the upstream task and fine-tuning to the specific downstream task [24]–[26]. However, the technique of **pretrain and finetune** may be

This work is supported by National Natural Science Foundation of China Grant 62002083 and 61971153, Open Fund of State Key Laboratory of Remote Sensing Science Grant OFSLRSS202210, and Heilongjiang Provincial Natural Science Foundation Grant LH2021F012, and the Fundamental Research Funds for the Central Universities Grant 3072021CFT0801,3072022QBZ0805 and 3072022CF0808. (Corresponding author: Shou Feng.)

Chunhui Zhao, Boao Qin, Shou Feng and Wenxiang Zhu are with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China, and also with Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin 150001, China.(e-mail: zhaochunhui@hrbeu.edu.cn; qinboao@hrbeu.edu.cn; fengshou@hrbeu.edu.cn;zhuwenxiang@hrbeu.edu.cn).

Shou Feng and Lifu Zhang are with the State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences. Beijing 100101, China; (e-mail: zhanglf@radi.ac.cn)

Jinchang Ren is with the National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, UK; (e-mail: jinchang.ren@ieee.org)

suboptimal for the cross-scene HSIC, as we may be impossible to build a large-scale HSI dataset like Imagenet [27] and CoCo [28]. Moreover, the spectral drift problem for the HSIs from different regions and sensors is quite serious, and different land covers may have similar spectral reflectance. Thus the generic and transferable representations learned from upstream tasks may be hardly to generalize to downstream specific tasks due to the interclass similarity and intraclass variability. Consequently, specific task-oriented transfer learning, also known as domain adaptation (DA), which aims to bridge the distribution gap between source and target domains to transfer the task-related knowledge, may be more suitable for the cross-scene HSIC task.

In recent years, DA-based HSIC methods have been developed to remedy the poor generalization of conventional DL models when the data comes from different scenes. These methods can be divided into two categories: 1) Semi-supervised DA [29]–[32], aiming to leverage both the small number of labeled target samples and massive unlabeled target samples to alleviate the inter-domain shifts; 2) Unsupervised domain adaptation (UDA), which predicts the target data without leveraging any labeling information from the target domain. In this work, we will focus on UDA, because it is more in line with the real HSIC task. From the viewpoint of general machine learning, prior UDA methods can be roughly categorized as two seminal lines: 1) Domain discrepancy minimization-based methods [33]–[37], e.g., maximum mean discrepancy (MMD) [34], [38], concentrating on explicitly aligning the distribution by mitigating the domain discrepancy. 2) Adversarial-based methods [39]–[45], e.g., Domain Adversarial Neural Networks (DANN), borrowing ideas from Generative Adversarial Network (GAN) [46]. DANN instances the hypothesis-induced domain divergence into a binary classifier and encourages the model to learn the domain-confused features in an adversarial manner, thereby reducing the domain divergence. Apart from these mainstream methods, some researches attend to seek a common subspace to implicitly transfer the knowledge [47], [48], for instance, Yao et al. [49] proposed a method based on the tensor alignment to project original tensors into an invariant subspace. Encouraged by the excellent performance of cross-attention in the multi-modal, Xu et al. proposed CDTrans [50] to reduce the gap between domains by aligning the instances from source-target pairs. Moreover, due to the MMD-based methods cannot take into account the geometric distribution of data when estimating the discrepancy between two domains, Zhang et al. [51] first investigate a novel topological structure and Semantic information Transfer network (TSTnet) by considering the alignment of statistical distribution and geometric distribution at the same time. Motivated by the semi-supervised learning, Fang et al. [52] proposed a confident learning-based DA for HSIC.

Numerous advances have manifested that both the domain discrepancy minimization based methods and the **DANN-based** methods can perform well when there only existed the difference of marginal distribution between domains However, in many cases, when the joint distributions of features and classes changes, only taking the alignment of marginal distribution into consideration may be not enough. Therefore, some advanced methods tried to align the marginal distribution and joint distribution between the source and target domain. For example, JDA [53] aligns conditional distribution through the simultaneous alignment of marginal distribution and joint distribution. Conditional domain adversarial networks (CADNs) [54] aim to align the joint distribution of features and classes through the conditional GAN. Liu et al. [41] proposed a cross-scene HSIC method to align the conditional distribution of each class via combined **class-wise DANN** and MMD

In general, the core of UDA is to find the discriminative features for downstream tasks of target domains, so how to optimize the domain alignment task and discrimination task in parallel is the key point of UDA. To fit this gap, Maximum Classier Discrepancy (MCD) [55] starts to reduce the domain discrepancy and improve the discrimination of decision boundaries, which is the first technique to estimate and optimize $\mathcal{H}\Delta\mathcal{H}$-Divergence [56] in a fully parameterized manner. Lately, many variants of hypothesis adversarial learning [57]–[59] are introduced for UDA. Although these advanced adversarial-based methods have taken into account both the domain alignment and classification tasks, their motivation is to provide additional inductive information to the discriminator in the domain level or decision level [60]. The alignment of entire features prone to lead the negative transfer of the model, that is, the task-irrelevant features may harm the results of transfer learning and degrade the classification performance. Especially for the methods based on multiple different initialized classifiers, task-irrelevant features may result in the amplification of the errors between classifiers.

To address the aforementioned problems, a novel unsupervised domain adaptation (UDA) framework towards multi-level features and boundaries (ToMF-B) is proposed for the cross-scene HSIC in this paper. The ToMF-B mainly aims to align the task-related features and learn the task-specific decision boundaries in parallel. Specifically, to coordinate the task of domain adaptation and the classification task, ToMF-B adopts the maximum classifier discrepancy algorithm to carry out the task of domain alignment by considering classification-specific decision boundaries simultaneously. To achieve the task of domain alignment, a two-stage alignment scheme is introduced to bridge the interdomain distribution gap of local features and instance-level features. Moreover, to build richer and more diverse feature spaces for learning the task-related domain-confused features, a hybrid model based on CNN and Transformer is designed as the feature extractor of ToMF-B, which can retain the local properties and the long-distance contextual dependency of the HSI data. Finally, to avoid the negative transfer caused by task-irrelevant features, multi-level features obtained by the feature extractor (generator) are decomposed into task-related features and task-irrelevant features. By enhancing task-related features while suppressing task-irrelevant features, features related to the classification task are fed into the two classifiers to perform the domain alignment, which enables the aligned domain-confused features to explicitly serve classification tasks of each classifier. The main contributions of this paper can be summarized as follows:

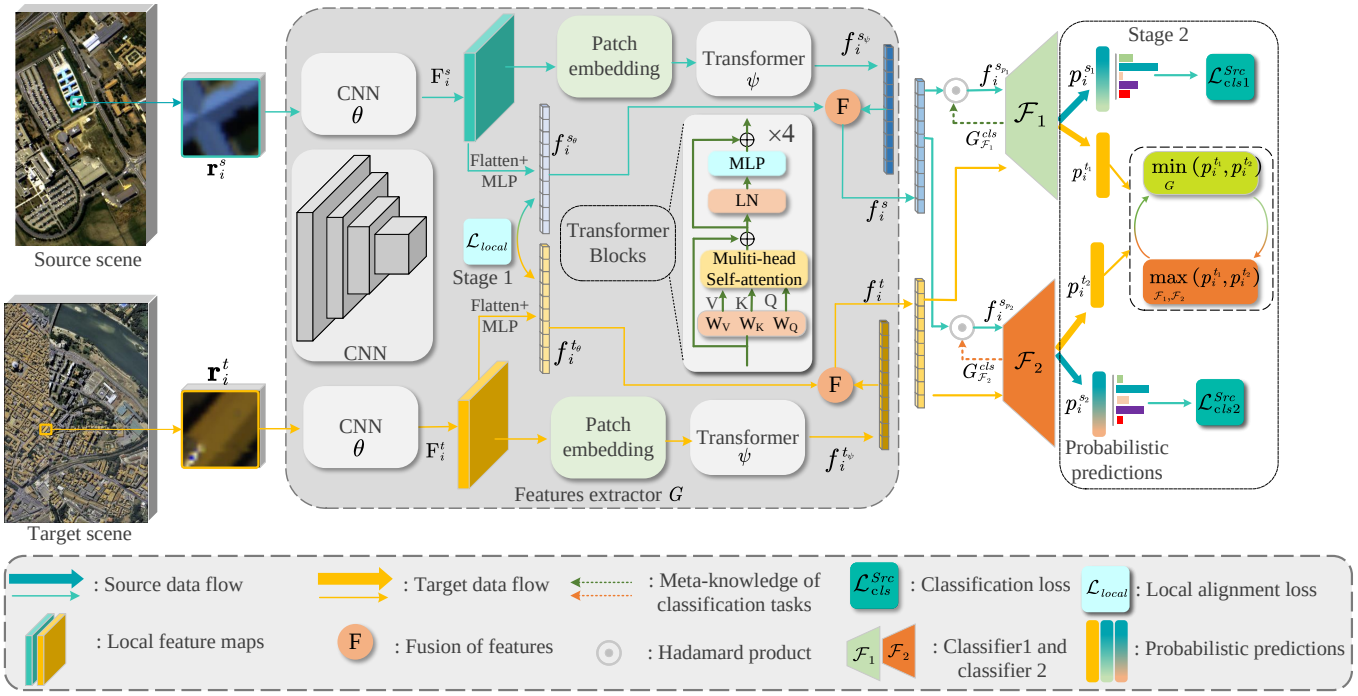


Fig. 1. An overview of the proposed ToMF-B, which consists of (a) feature extractor (generator), (b) two classifiers with different initialization. The feature extractor (generator) combines CNN and Transformer, aiming to enrich the feature representation of two domains. Two classifiers are employed to perform the domain align task and to obtain the task-specific decision boundaries. The classification meta-knowledge is adopted to obtain the task-related features. Note that two branches are weight-sharing, and joint predictions of two classifiers are taken as test results of target scenes in the inference stage.

1) A novel cross-scene HSIC framework is proposed, which is an unsupervised domain adaptation framework towards multi-level features and decision boundaries (ToMF-B). The design of ToMF-B is based on the maximum classifier discrepancy criterion, which is consisted of a feature extractor and two classifiers, so that the ToMF-B can learn task-specific decision boundaries while can align the distributions of source and target. Then a two-stage alignment scheme allows the ToMF-B to bridge the interdomain gap of local features and instance-level features, that is, the local alignment ensures the similar distribution of local features from CNN, and the global alignment performed by minmax the discrepancy of two classifiers will alleviate the interdomain distribution gap of multi-level features space.
2) A hybrid model based on CNN and Transformer is designed as the feature extractor of ToMF-B to generate multi-level features. Specifically, the fusion of the local features extracted by CNN and the long-range contextual features built by the subsequent Transformer retains the local properties and the long-distance contextual relationship of the data, enriching the features representation of two domains. Such multi-level features provide the richer and more diverse feature spaces, which is helpful for learning the commonly task-related features when performing the task of domain alignment.
3) A task-oriented feature decomposition method based on meta-knowledge of classification tasks is leveraged to enhance the task-related features while suppressing the task-irrelevant features. The task-related features can be obtained by adopting the gradients of the predicted score corresponding to the ground-truth class as the attention weights of multi-level features, and the task-irrelevant features are on the opposite side. By taking advantage of the remaining task-related features, the aligned domain-invariant features can explicitly serve the classification tasks of each classifier, which will avoid the negative transfer or mode collapse caused by task-irrelevant features.

The remainder of this paper is organized as follows. In Section II, the proposed ToMF-B is presented in detail. Sections III presents and analyzes the experimental results in detail. Finally, conclusions are reported in Section IV.

II. Methodology

The ToMF-B framework is illustrated in Fig. 1, which is composed of a feature extractor and two classifiers. The goal of ToMF-B is to align the task-related features and learn the task-specific decision boundaries in parallel. At first, a multi-level features extractor (generator), which hybrids the CNN and Transformer, is designed to build the multi-level features representation. The local feature maps are fed into the Transformer to build the long-range contextual relationship after the patch embedding. Moreover, the feature maps are transformed into corresponding feature vectors with the same dimension as the class token. Then the source multi-level feature vectors fusing the local and long-range features are decomposed into two task-related features based on the meta-knowledge of each classifier. The source task-related and target feature vectors are feed into two classifiers, and a two scheme

is employed to bridge the interdomain gap. It is noteworthy that we decompose the holistic feature based on the weights of two classifiers to obtain their respective task-related features. Therefore, the source features fed into two classifiers are related to the classification task of each classifier. In the final inference stage, the joint predictions of two classifiers are taken as test results of target scenes.

A. Multi-level features extractor hybrid CNN and Transformer

The purpose of the multi-level features extractor is to enrich the features representation of two domains and provide the multi-level features for alignment. The features extractor hybrid the CNN model and the Transformer model, thus the fusion of the local features extracted by CNN and the long-range contextual features built by the subsequent Transformer will retain both the local properties and the long-range contextual relationship. Indeed, there are some advanced hybrid models based on CNN and ViT, these works mainly focus on combining the strengths of these two architectures while avoiding their respective limitations. However, there are two main purposes for us to adopt the hybrid model. Firstly, we aim to enrich the feature representation via the hybrid model so that task-related features and domain-invariant features can be fully learned. Secondly, by aligning the local features from the CNN and the instance-level features extracted by Transformer, the domain shifts will be further bridged in two-stage manner. Assumed the labeled samples $X^s = \{x_i^s\}_i^{n_s} \in \mathbb{R}^D$ with corresponding class labels $Y^s = \{y_i^s\}_i^{n_s}$ are from the source HSI $\mathbf{R}^s \in \mathbb{R}^{H_s \times W_s \times D}$ and the target samples $X^t = \{x_i^t\}_i^{n_t} \in \mathbb{R}^D$ with no access to the labels are from the target HSI $\mathbf{R}^t \in \mathbb{R}^{H_t \times W_t \times D}$. n_s and n_t denote the size of source samples and target samples, respectively. D is the number of spectral bands, and the spectral dimension of HSIs from two scenes/domains have been sampled to be the same. Importantly, we will only focus on the close-set cross-scene classification, hence the $Y^s, Y^t \in \mathbb{R}^C$, where the C is the number of class.

Initially, a CNN with four 2-D convolutional blocks is utilized to extract the local features of the source and target samples. As for the each convolutional block, it consists of a convolution layer with 3×3 kernel size, a batch normalization (BN) layer and ReLU. It is noteworthy that the following derivations take the source as an example, and the corresponding results of the target domain can be obtained in the same way. Denoted θ as the CNN, taking the sample x_i^s and surrounding pixels as the input HSI cube $\mathbf{r}_i^s \in \mathbb{R}^{\mathrm{P}\times\mathrm{P}\times\mathrm{D}}$, where the $\mathrm{P}\times\mathrm{P}$ is the spatial size of HSI cube, then the local feature map can be obtained by

$$\mathrm{F}_i^s = \theta(\mathbf{r}_i^s), \qquad \mathrm{F}_i^s \in \mathbb{R}^{\mathrm{P}\times\mathrm{P}\times\mathrm{d}} \tag{1}$$

where d is the output dimension of the last convolution block and the spatial size of feature maps is same as the input cube due to the padding operation. The obtained feature map F_i^s is transformed as the local feature vector $f_i^{s_\theta}$ via the flatten operation and MLP's mapping. Moreover, following the Vision Transformer (ViT), the feature map F_i^s is also separated into M non-overlapping patches $\mathrm{F}_{i,m}^s \in \mathbb{R}^{p\times p\times d}, m \in M$ by a patch embedding module. Each patch is treated as a token $f_{i,m}^s \in \mathbb{R}^{1\times1\times\hat{d}}$, where $\hat{d}$ is the embedding dimension. It is noteworthy that $f_i^{s_\theta}$ is also a $\hat{d}$-dimensional feature vector.

Then a Transformer network with four Transformer blocks is employed to model the long-range contextual representation. Let ψ represent the Transformer network, then the final class token that aggregates the classification useful information can be denoted as

$$f_i^{s_\psi} = \psi(\mathrm{F}_i^s), \qquad f_i^{s_\psi} \in \mathbb{R}^{1\times1\times\hat{d}} \tag{2}$$

After that, the multi-level feature can be acquired by the fusion of local feature vector $f_i^{s_\theta}$ and the class token $f_i^{s_\psi}$ contained the long-range contextual relationship. To control the contribution of the two features for different hyperspectral data, we adopted a trade-off parameter α to balance the importance of local properties and the long-range contextual relationship. As such, the output feature vector of the generator can be represented as

$$f_i^s = \mathcal{G}(\mathbf{r}_i^s) = \alpha f_i^{s_\theta} + (1-\alpha) f_i^{s_\psi} \tag{3}$$

where $f_i^s \in \mathbb{R}^{1\times1\times\hat{d}}$ has the same dimension as the class token. $\mathcal{G}(\cdot) = \psi(\theta(\cdot))$ is a composite function of CNN model $\theta(\cdot)$ and transformer network $\psi(\cdot)$.

B. Domain alignment towards multi-level features and classification boundaries

To learn task-specific decision boundaries while aligning the distributions of source and target, the design of ToMF-B is based on maximum classifier discrepancy [55], which is consisted of a feature extractor and two classifiers. In UDA, a common belief is that the source and target domain share the same or close feature space but are drive from different marginal distributions, that is, $P, Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, $P_x \neq Q_x$. P and Q are the source and target distribution over $\mathcal{X} \times \mathcal{Y}$, where the $\mathcal{X}$ is the instance space and the $\mathcal{Y}$ is the label space. P_x and Q_x indicate the marginal distribution of source and target domain over $\mathcal{X}$.

1) The UDA theory based on $\mathcal{H}\Delta\mathcal{H}$-Divergence: The core idea of the domain adaptive theory is to build the relationship of generalization error between two domains, which can implicitly reduce target domain error by optimizing the source domain. Base on this objective, Ben-David et al. [56] proposed a seminal theory by characterizing the disagreement between any pair of labeling hypotheses. The error of hypothesis h (e.g., models or classifiers) on the target domain Q is denoted as

$$\begin{gathered}\mathcal{E}_Q(h) \leq \mathcal{E}_P(h) + d(P,Q) + \lambda, \\ \lambda = \mathcal{E}_P(h^*) + \mathcal{E}_Q(h^*)\end{gathered} \tag{4}$$

where the $h^* = \underset{h\in\mathcal{H}}{arg\,\min}\left[\mathcal{E}_P(h^*) + \mathcal{E}_Q(h^*)\right]$. To sum up this inequality, the error $\mathcal{E}_Q(h)$ is constrained by the three terms: 1) the error of h on the source domain, which can be optimized by source domain labeled data; 2) the shared error of ideal joint hypothesis λ, which can be viewed as a constant because it should be sufficiently small; 3) the disparity difference

between P and Q, which cannot be estimated directly due to the target labels is inaccessible. Therefore, the goal of the UDA is to reduce the bound of disparity difference, the paper [56] proposed the $\mathcal{H}\Delta\mathcal{H}$-Divergence to measure the upper bound, and the $\mathcal{H}\Delta\mathcal{H}$-Divergence between P and Q is denoted as

$$d_{\mathcal{H}\triangle\mathcal{H}}\left(P,Q\right)\triangleq\underset{h,h'\in\mathcal{H}}{sup}\left|\mathcal{E}_s\left(h,h'\right)-\mathcal{E}_t\left(h,h'\right)\right| \tag{5}$$

The key innovation of [56] presents that a distribution distance can be estimated based on a hypothesis space $\mathcal{H}^{\{0,1\}}$ of binary classification, therefore, the disparity can be expressed as

$$d_{\mathcal{H}\triangle\mathcal{H}}\left(P_x,Q_x\right)\triangleq\underset{h,h'\in\mathcal{H}^{\{0,1\}}}{sup}\left|E_{Q_x}\mathbb{I}\left[h\neq h'\right]-E_{P_x}\mathbb{I}\left[h\neq h'\right]\right| \tag{6}$$

where the $E_{Q_x}\mathbb{I}\left[h\neq h'\right]=E_{x\sim Q_x}\mathbb{I}\left[h\left(x\right)\neq h'\left(x\right)\right]$, the disparity between h and h' can be specified a measurable subset $\{x\in X|h\left(x\right)\neq h'\left(x\right)\}$. In this way, the distribution distance between P_x and Q_x can be measured on the subsets by taking the supremum over all pairs of $h,h'\in\mathcal{H}^{\{0,1\}}$. To this end, we adopted the L_1-distance draw from the MCD [55] to measure the disagreements,

$$E_{\boldsymbol{x}\sim D}\frac{1}{K}\left\|\left(h(\varphi(\boldsymbol{x}))\right)-\left(h'(\varphi(\boldsymbol{x}))\right)\right\|_1 \tag{7}$$

where φ is the learnable network, and D is the distribution of $\boldsymbol{x}$. To close the supremum of target domain error, the optimization objective of UDA can be defined as

$$\begin{aligned}\min_{h,\varphi}\quad&\mathcal{E}_{P_x^{\varphi}}(h)+E_{x\sim Q_x}\mathbb{I}\left[h\left(\varphi(x)\right)\neq h'\left(\varphi(x)\right)\right],\\\max_{h,h'}\quad&E_{x\sim Q_x}\mathbb{I}\left[h\left(\varphi(x)\right)\neq h'\left(\varphi(x)\right)\right]\end{aligned} \tag{8}$$

which suggests the model to reduce the error of target domain in an adversarial manner. As a result, the goal of the domain alignment is to optimize the three items in formula (8).

2) Specific procedures for the optimization task: At beginning, three modules for classifier $\mathcal{F}_1$, classifier $\mathcal{F}_2$ and generator $\mathcal{G}$ are built, and we ensemble the two classifiers into the features extractor via build two classification heads similar to linear probe. To improve the nonlinear capability of classification heads, the additional Batch normalization (BN) layers and hidden layers are added to the heads. This design allows the gradient of classification heads to feed back to the backbone $\mathcal{G}$ in the process of backward, which develops a pathway for the subsequent method of features decomposition guided by classification meta-knowledge. The training of our model are mainly divided into three steps, minimizing the error of source domain, maximizing the discrepancy of classifiers $\mathcal{F}_1,\mathcal{F}_2$ and the alignment of distribution (minimize discrepancy by $\mathcal{G}$), which are responsible for the three optimization terms of (8), respectively. In addition, the first-stage alignment is employed into step one to reduce the interdomain gap of local features, and the step two and three can be viewed as the second-stage alignment, i.e., the alignment instance-level features.

Step one, for the term of source error $\mathcal{E}_{P_x^{\varphi}}(h)$, the model can be trained by minimizing the loss of labeled source data as

$$\min_{\mathcal{G},\mathcal{F}_1,\mathcal{F}_2}\mathcal{L}_{cls}^{Src}\left(X^s,Y^s\right) \tag{9}$$

where the $\mathcal{L}_{cls}^{Src}\left(X^s,Y^s\right)$ is the classification loss of source domain. Defined the cross entropy loss as

$$\mathcal{L}(\mathrm{p}_i,\mathrm{y}_i)=-\sum_{c=1}^{C}y_i^c\log p_i^c \tag{10}$$

where the y_i is the one-hot encoding of labels and the p_i is the probabilistic predictions obtained by the softmax layer. Hence the classification loss of classifier can be defined as

$$\mathcal{L}_{cls1}^{Src}\left(X^s,Y^s\right)=\frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}(p_i^{s_1},y_i) \tag{11}$$

Noted that the $p_i^{s_1}$ is obtained by the task-related feature, and the details of task-related features will be explained in the section C. In addition to the minimization of source error, a MMD loss is employed to align local feature vectors in this step, which will ensure that local feature vectors of two domains have similar statistical distribution. Denoted the $\mathcal{L}_{local}$ as the loss of local alignment, then local alignment loss can be drown as

$$\mathcal{L}_{local}\left(X^s,X^t\right)=\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}f_i^{s_\theta}-\frac{1}{n_t}\sum_{j=1}^{n_t}f_i^{t_\theta}\right\|_H^2, \tag{12}$$

where the $f_\theta(i)$ is the local feature vectors from the CNN. The meaning of the formula is to calculate the mean discrepancy of local features between SD and TD in reproducing kernel Hilbert space.

Step two, to maximize the domain discrepancy, e.g., the term of $\max_{h,h'}E_{x\sim Q_x}\mathbb{I}\left[h\left(x\right)\neq h'\left(x\right)\right]$, freezing the $\mathcal{G}$, two independent classifier heads $\mathcal{F}_1,\mathcal{F}_2$ with different initialization are trained to maximize the discrepancy of each other while minimizing the source error, then the optimal objective can be denoted as

$$\min_{\mathcal{F}_1,\mathcal{F}_2}\mathcal{L}\left(X^s,Y^s\right)-\mathcal{L}_{\mathrm{adv}}\left(X^t\right) \tag{13}$$

where the $\mathcal{L}(X^t)$ is written as

$$\begin{aligned}&\mathcal{L}_{\mathrm{adv}}\left(X^t\right)=E_{x^t\sim X^t}\left[d\left(p^{t_1}\left(y\mid\varphi_1(x^t)\right),p^{t_2}\left(y\mid\varphi_2(x^t)\right)\right)\right],\\&d\left(p^{t_1}\left(y\mid\varphi_1(x^t)\right),p^{t_2}\left(y\mid\varphi_2(x^t)\right)\right)=\\&\qquad\frac{1}{C}\sum_{c=1}^{C}\left|p^{t_1}\left(y=c|\varphi_1(x^t)\right)-p^{t_2}\left(y=c|\varphi_2(x^t)\right)\right|,\end{aligned} \tag{14}$$

where the $\varphi_1(x^t)=\mathcal{F}_1(f_i^t)$ and $\varphi_2(x^t)=\mathcal{F}_2(f_i^t)$ are the target outputs of two classifiers, and C is the class dimension of final linear layer. After that, the decision boundary will be far from the samples without discriminative representation in the feature space.

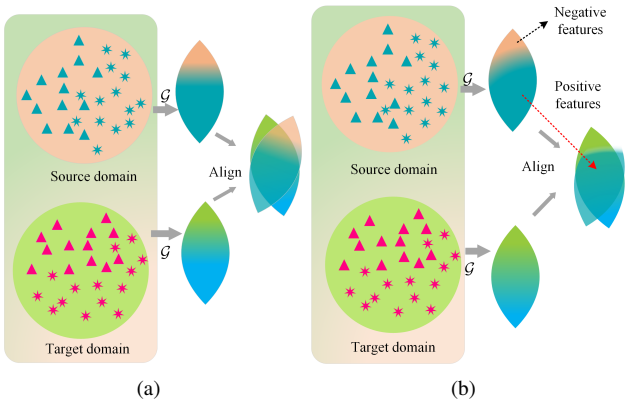


Fig. 2. An illustration of difference between prior UDA methods of HSIC and ours. (a) Prior hyperspectral UDA methods generally align the holistic feature representation. (b) Task-oriented features decomposition only takes task-related features into consideration via discarding task-irrelevant features.

Step three, to minimize the domain discrepancy, e.g., the term of $\min_{h,h'} E_{x\sim Q_x}\mathbb{I}\left[h\left(x\right)\neq h'\left(x\right)\right]$, fixing the $\mathcal{F}_1$ and $\mathcal{F}_2$, the generator $\mathcal{G}$ is updated to reduce the interdomain gap. After closing the distribution of local outputs of CNN in the first stage (step one), the second-stage alignment aims to minimize the discrepancy of instance-level features, that is

$$\min_{G}\mathcal{L}_{\text{adv}}\left(X^t\right). \tag{15}$$

As such, the features extractor $\mathcal{G}$ will try to generate the target features near/within the boundaries. Finally, repeat the step two and step three for k times, and the ToMF-B can align the task-related features and learn the task-specific decision boundaries in parallel. To clarify the overall optimization process, the three-step losses are summarized as follows

$$\min_{\mathcal{G},\mathcal{F}_1,\mathcal{F}_2}\mathcal{L}_{cls}^{Src}\left(X^s,Y^s\right)+\mathcal{L}_{local}\left(X^s,X^t\right), \tag{16}$$

$$\min_{\mathcal{F}_1,\mathcal{F}_2}\mathcal{L}\left(X^s,Y^s\right)-\mathcal{L}_{\text{adv}}\left(X^t\right), \tag{17}$$

$$\min_{G}\mathcal{L}_{\text{adv}}\left(X^t\right). \tag{18}$$

C. Task-oriented features decomposition for UDA

Task-irrelativant features may damage the discrimination power of aligned features and further hurt the entire transfer process, such as negative transfer or mode collapse. Accordingly, it is crucial to learn the task-related domain-invariant features. To address this, a task-oriented features decomposition based on the classification meta-knowledge is adopted to enhance the task-related features while suppressing the task-irrelevant features. Motivated by the Toalign [60] and the CAM [61], a holistic feature vector can be decomposed into a task-discriminative vector and a task-irrelevant (negative) one. According to the formula (3), the source multi-feature vector $f_i^s\in\mathbb{R}^{1\times1\times\hat{d}}$ can be viewed as the holistic feature vector. Therefore, our goal is to seek the task-related features from the diversely multi-level feature representation, and the procedures to get final features is designed as follows.

At first, the logits of all source classes are predicted by the classification heads $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$. Taking the classifier $\mathcal{F}_1$ as an example, based on the response of final linear layer, the gradient $\mathbf{G}_{\mathcal{F}_1}^{cls}\in\mathbb{R}^C$ of y^c can be denoted as

$$\mathbf{G}_{\mathcal{F}_1}^{cls}=\frac{\partial y_c^s}{\partial f_i^s}, \tag{19}$$

where y_c^s is the predicted score corresponding to the class c, and C is the class dimension of final linear layer.

Then, $\mathbf{G}_{\mathcal{F}_1}^{cls}$ can drive the model to find the classification-discriminative features from the view of channel-wise attention. In turn, the task-irrelevant feature can be drown by

$$f_i^{s_{n_1}}=-\mathbf{G}_{\mathcal{F}_1}^{cls}\odot f_i^s=-\eta\mathbf{G}_{\mathcal{F}_1}^{cls}\odot f_i^s \tag{20}$$

where the $\odot$ is the Hadamard product and the η is an adaptive parameter to modulate the $\mathbf{G}_{\mathcal{F}_1}^{cls}$. The detailed η is drawn from the Toalign, so the details of η see [60].

Finally, the task-related features can be obtained by suppressing the task-irrelevant features, which is

$$f_i^{s_{p_1}}=f_i^s-\lambda f_i^{s_{n_1}} \tag{21}$$

where λ is a regularization parameters controlling the suppressed degree of task-irrelevant features (contribution of negative features). The task-related feature vectors of classifier $\mathcal{F}_1$ can be obtained in the same manner. In this way, the source feature vectors that will be fed into the classifiers is the task-related features, which will avoid the harmless of task-irrelevant features to the transfer process. The overall different between the task-oriented features decomposition and prior related studies is presented in Fig.2.

Overall, in order to achieve the distribution alignment of two domains and the adaptation of classification boundaries in parallel, the overall design of ToMF-B adopts MCD algorithm, which consists of a feature extractor and two classifiers with different initialization. However, the negative transfer caused by some difficult samples may be amplified by two classifiers with discrepancy when the input features learned from the features extractor (generator) are unexpected. We wanted to avoid this problem via feature decomposition-based approaches, thus a task-based feature decomposition method is leveraged to compress the task-irrelevant features. Furthermore, MCD mainly focuses on the distribution alignment of instance-level features, and a rich feature space will provide better task-related and domain-invariant feature representation. Therefore, a hybrid model is designed to enrich the feature representation, and the alignment of local features can help the instance-level alignment to further reduce the domain shift.

III. Experimental Results and Analysis

In this section, the details of three experimental datasets and experimental settings are first introduced. Then, the results of the proposed method and nine classical and state-of-the-art methods on three HSI datasets are compared. Finally, the ablation of the ToMF-B is analyzed in detail and some parameters of the ToMF-B are also discussed.

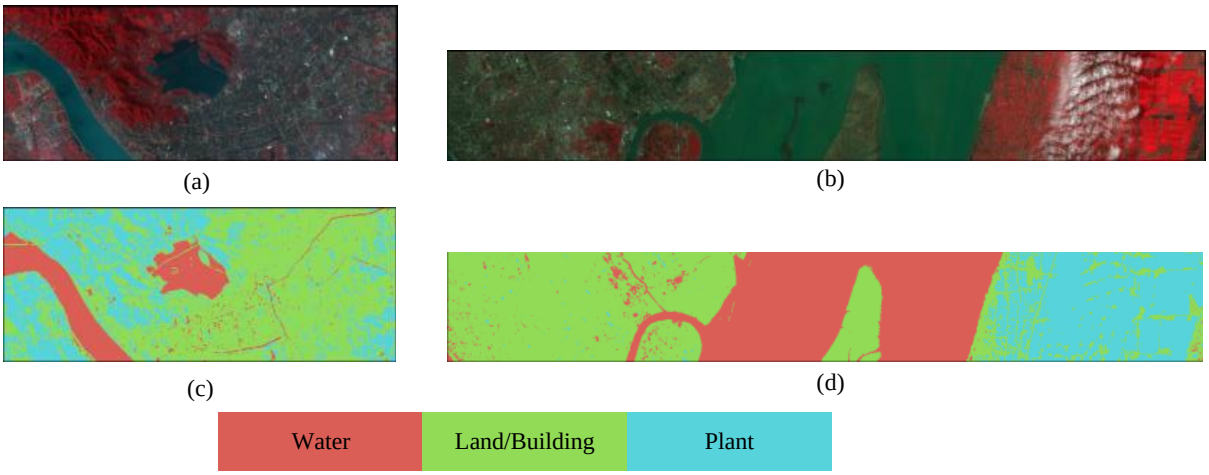


Fig. 3. An overview of the H-S dataset. (a) Pseudo color image of HangZhou. (b) Pseudo color image of ShangHai. (c) Ground-truth map of HangZhou. (d) Ground-truth of ShangHai.

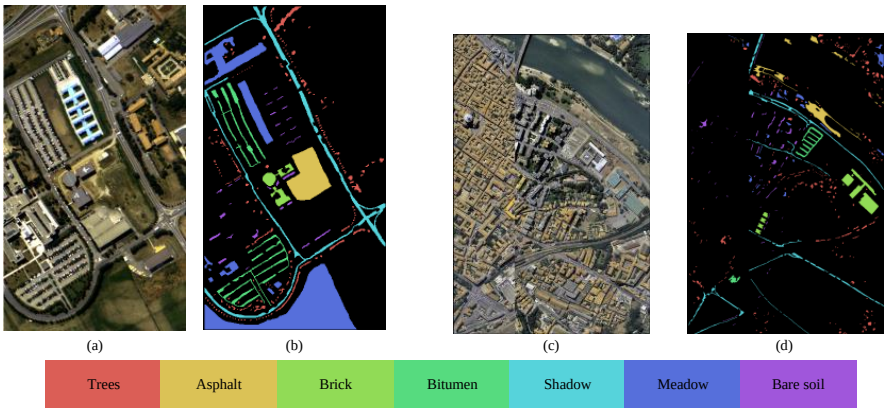


Fig. 4. An overview of the PaviaU-C dataset. (a) Pseudo color image of University of Pavia. (b) Ground-truth map of University of Pavia.. (c) Pseudo color image of Pavia Center. (d) Ground-truth of Pavia Center.

TABLE I
LAND COVER CLASSES AND THE NUMBER OF SAMPLES FOR THE HANGZHOU-SHANGHAI DATASET

No.	Land Cover Type	Source / Train	Target
C1	Water	18043 / 133	123123
C2	Land/Building	77450 / 571	161689
C3	Plant	40207 / 296	83188
	Total	135700 / 1000	368000

TABLE II
LAND COVER CLASSES AND THE NUMBER OF SAMPLES FOR THE PAVIAU-PAVIAC DATASET

No.	Land Cover Type	Source / Train	Target
C1	Trees	3064 / 78	7598
C2	Bare soil	5029 / 24	6584
C3	Bitumen	1330 / 94	7287
C4	Brick	3682 / 34	2685
C5	Asphalt	6631 / 128	9248
C6	Meadow	18649 / 474	3090
C7	Shadow	947 / 168	2863
	Total	39332 / 1000	39355

A. Datasets

Three benchmark HSI datasets, e.g., ShangHai-HangZhou data [62], PaviaU-PaviaC[1] data and HyRANK [63] data, are employed for performance assessment. These datasets are collected by different sensors over different land covers, and have been selected for experiments by many related papers. The details of these datasets are given as follows.

TABLE III
LAND COVER CLASSES AND THE NUMBER OF SAMPLES FOR THE DIONI-LOUKIA DATASET

No.	Land Cover Type	Source / Train	Target
C1	Dense Urban Fabric	1262 /63	206
C2	Mineral Extraction Sites	204 / 5	54
C3	Non Irrigated Arable Land	614 / 31	426
C4	Fruit Trees	150 / 7	79
C5	Olive Groves	1768 / 88	1107
C6	Coniferous Forest	361 / 18	422
C7	Dense Sderophyllous Vegetation	5035 / 251	2996
C8	Sparce Sderophyllous Vegetation	6374 / 318	2361
C9	Sparcely Vegetated Areas	1754 / 88	399
C10	Rocks and Sand	492 / 25	453
C11	Water	1612 / 81	1393
C12	Coastal Water	398 / 20	421
	Total	20024 / 1000	10317

1) *ShangHai-HangZhou (S-H)*: This dataset was acquired by

[1]http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

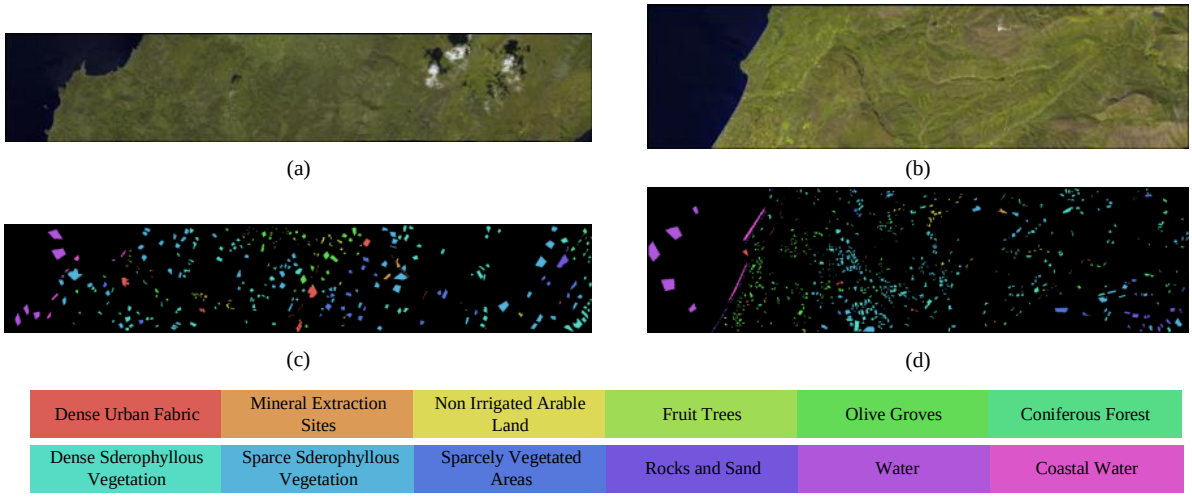


Fig. 5. An overview of the HyRANK dataset. (a) Pseudo color image of Dioni. (b) Pseudo color image of Loukia. (c) Ground-truth map of Dioni. (d) Ground-truth of Loukia.

the REO-1 Hyperion hyperspectral sensor, which consists of 198 spectral bands after removing 22 bad bands. ShangHai data wass collected over ShangHai city on 1 April 2002, which includes roads, buildings, plants, and the waters of the Yangtze River and Huangpu River. The HangZhou data was acquired over HangZhou city on 2 November 2002, including roads, buildings, plants, West Lake, and the Qiantang River basin The size of HangZhou and ShangHai is 590×230 and 1600×230, respectively. The HangZhou is taken as source scene, and the ShangHai is viewed as the target scene in this paper. Three common land cover classes were selected for cross-scene classification, which are listed in Table I, and their pseudocolor iamge and ground-truth maps are presented in Fig.3.

2) *PaviaU-PaviaC (PaviaU-C)*: This dataset was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the urban area surrounding the University of Pavia and Pavia center, Italy. The spatial size of two data are 610×340 and 1096×715, respectively. The spatial resolution is as high as 1.3 m/pixel, and both datasets are sampled as same 102 spectral bands. The PaivaU is taken as source scene, and the PaivaC is viewed as the target scene in this paper. Seven common classes are selected for this experiment, which are listed in Table II, and their pseudocolor iamge and ground-truth maps are presented in Fig.4.
3) *HyRANK*: The HyRANK dataset was collected by the Hyperion sensor (EO-1, USGS), which has 176 spectral bands. The size of two labeled scenes, Dioni and Loukia, are 250×1376 and 246×945, respectively. There are 12 common classes, which are listed in Table III. The Dioni is taken as source scene, and the Loukia is viewed as the target scene in this paper. The pseudocolor representations and ground-truth maps of two scenes are shown in Fig.5.

B. Implementation Details

All the experiments were implemented on the same hardware platform GPU: GTX-3090, CPU: Intel 4210R and memory: 32G. The total of 1000 labeled source samples are uniform sampled from per class for training, and the specific number of training samples per class are listed in Tables I-III. All methods for comparison are trained for 100 epochs without any data augmentation, and the patch size of all experiments is 12×12. The SGD is adopted as the optimizer, and the batch size is 100. Meanwhile, the trade-off parameter of features fusion α, the learning rate and parameter λ are empirically set to 0.5, $1e^{-3}$, $1e^{-1}$, respectively. The repeat times k of step three are empirically set as 5,5,15 for the S-H, PaviaU-C and HyRANK datasets, respectively. The overall accuracy (OA), class-specific accuracy, and the Kappa coefficient are employed to evaluate the classification performance. To avoid biased estimation, all experiments were conducted with ten independent tests, and the average values were reported for all the evaluation metrics.

C. Comparison With State-of-the-Art Methods

1) The settings of different methods: To show the overall classification performance of the proposed method, nine representative methods belong to different categories with different advantages are selected to compare with ToMF-B. Specifically, these compared methods are: a baseline based on the CNN, an advanced CNN-based deep network DBDA [64], a Vision Transformer based network (ViT) [21], a semi-supervised deep cross-domain few-shot learning(DCFSL) [30], an confident learning-based unsupervised domain adaptation (UDA) method (CLDA) [39], a deep metric learning based UDA method (S-DMM) [37], a subspace learning-based UDA method (DCA) [47], a classification boundaries-oriented UDA method (MCD) [55], a UDA-based method for segmentation tasks (CLAN) [65], and a network coupling CCN and GCN for the alignment of statistical and geometric distribution (TSTnet) [51]. In terms of source-only methods, i.e.,the CNN, DBDA and ViT, they are trained on source scenes and directly test on

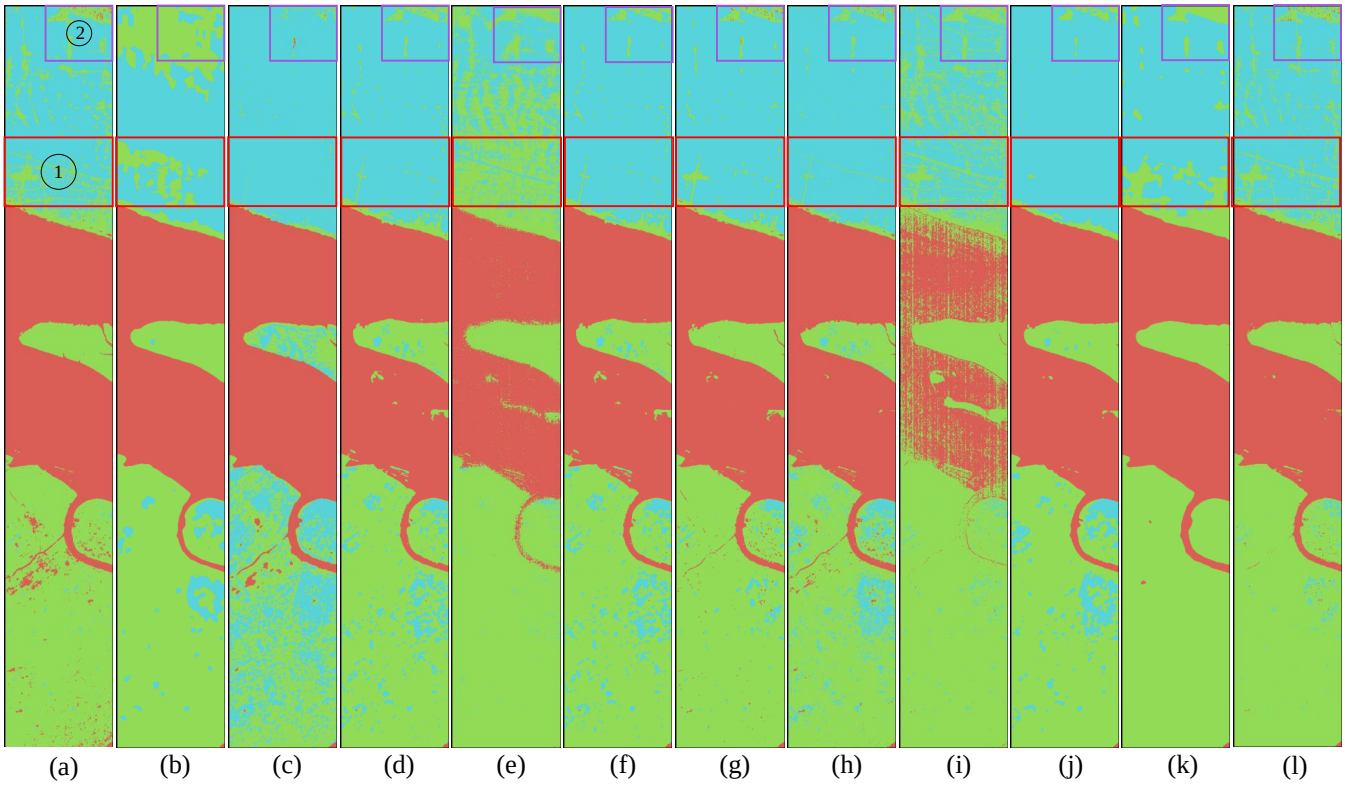


Fig. 6. Classification maps of different methods on the target scene ShangHai. (a) Ground truth. (b) CNN. (c) DBDA. (d) ViT. (e) DCFSL. (f) CLDA. (g) S-DMM. (h) DCA. (i) MCD. (j) CLAN. (k) TSTnet. (l) ToMF-B(ours). Compared with other methods, (g) and (l) have the smoother classification maps, but the visualization results of (l) ours in the red and purple rectangle boxes contain more details and are more consistent with the ground truth (a) than other methods.

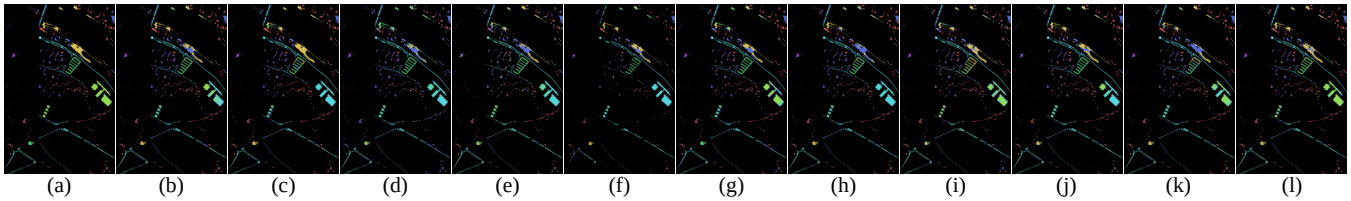
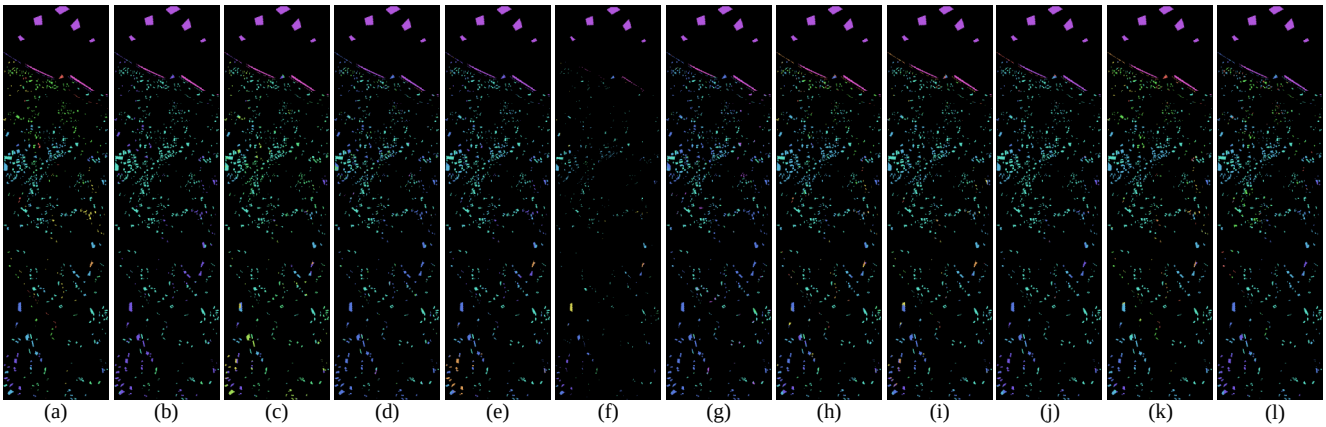


Fig. 7. Classification maps of different methods on the target scene Pavia Center. (a) Ground truth. (b) CNN. (c) DBDA. (d) ViT. (e) DCFSL. (f) CLDA. (g) S-DMM. (h) DCA. (i) MCD. (j) CLAN. (k) TSTnet. (l) ToMF-B(ours).

target scenes. The $12 \times 12 \times 4$ HSI cube is selected as a group-level token of the ViT. For the CLAN and MCD, the VGG is chosen as the backbone for three datasets, and segmentation-tailored classifiers of CLAN are replaced with general classifiers for the classification tasks. The hyper-parameters of all methods are consistent with the original settings, such as λ_1 and λ_2 in TSTnet, the embedding dimension in S-DMM and et al. It is noted that the one-shot sample is given to DCFSL for comparison.

2) The quantitative results of compared methods: Quantitative results of different methods are reported in Table IV-VI. Compared with other methods, the proposed ToMF-B achieves the best performance in terms of the OA, the class-specific accuracy and the Kappa coefficient. For the S-H dataset, it can be observed that methods without domain adaptive (DA) technique (CNN, DBDA and ViT) are generally lower than DA-based methods, but DBDA and ViT also achieve competitive performance. This can be inferred that the DBDA and ViT are capable of extracting the features with fine transferability on the H-S dataset. MCD and CLAN adopt the same backbone (VGG) as CNN, and their results achieve significant improvement when compared with CNN, which may demonstrate the effectiveness of the UDA technique. The performance of ToMF-B is better than other UDA-based methods in the H-S dataset, which might owe to the well-design multi-level feature extractor and two-stage UDA technique.

For the PaviaU-C dataset, the performance of some UDA-based methods are even worse than those of general networks, suggesting that the unfavorable UDA technique may lead to negative transfer on some datasets. The ToMF-B can still obtain the best performance, indicating that the feature-oriented (task-related features) transfer may remedy the risk of negative transfer. Note that the results of CLDA significantly drop compared with the results of the original paper (92.80% *vs.* 77.65%, OA), but the total number of labeled samples is close (1260 *vs.* 1000). This is because the sampling criteria are different, the original CLDA selected 180 labeled samples per class, and our experimental setting is to select total 1000



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 8. Classification maps of different methods on the target scene Loukia. (a) Ground truth. (b) CNN. (c) DBDA. (d) ViT. (e) DCFSL. (f) CLDA. (g) S-DMM. (h) DCA. (i) MCD. (j) CLAN. (k) TSTnet. (l) ToMF-B(ours).

TABLE IV
CLASSIFICATION ACCURACIES AND COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS ON THE TARGET SCENE SHANGHAI BY USING 1000 LABELED SAMPLES FROM THE SOURCE SCENE HANGZHOU.

Class Name	CNN	DBDA	ViT	DCFSL	CLDA	S-DMM	DCA	MCD	CLAN	TSTnet	**ToMF-B**
Water	96.84	94.81	95.65	97.13	92.04	94.72	97.28	80.08	95.68	90.24	**97.48±0.37**
Land/Building	56.56	84.62	86.01	93.26	82.35	90.05	91.33	83.71	85.36	87.90	**93.51±0.62**
Plant	62.41	75.46	83.86	90.15	89.13	89.24	60.93	87.52	82.68	94.96	**95.19±0.75**
OA(%)	71.36	85.96	88.75	93.85	91.46	91.43	86.45	83.36	88.21	90.28	**95.22±0.57**
Kappa	0.59	0.80	0.83	0.91	0.87	0.89	0.85	0.74	0.82	0.86	**0.93±0.01**
Params	126,212	599,646	2,781,124	56,369	187,728	200,274	-	**8,673,800**	1,184,713	5,370,372	4,468,680
FLOPS	7,584,000	**432,806,076**	138,629,376	43,147,560	4,095,800	201,402	-	47,162,368	8,641,600	5,343,752	210,881,280
Time (s)	62.65	321.95	82.92	803.45	288.96	1043.58	328.17	128.87	93.82	**1329.17**	166.07

TABLE V
CLASSIFICATION ACCURACIES AND COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS ON THE TARGET SCENE PAVIAC BY USING 1000 LABELED SAMPLES FROM THE SOURCE SCENE PAVIAU

Class Name	CNN	DBDA	ViT	DCFSL	CLDA	S-DMM	DCA	MCD	CLAN	TSTnet	**ToMF-B**
Trees	84.25	93.36	92.52	98.84	93.69	89.39	**99.66**	88.82	93.67	83.78	84.26±2.30
Asphalt	65.57	81.98	71.83	82.70	85.11	75.19	53.63	86.49	80.76	72.70	**89.21±3.23**
Brick	68.86	22.97	52.16	78.35	12.66	73.90	5.29	40.00	34.61	33.83	**80.03±2.91**
Bitumen	50.60	87.53	0	45.83	79.87	49.71	9.05	77.85	76.32	6.06	**84.80±6.44**
Shadow	97.79	82.09	73.26	99.37	99.89	80.99	99.42	97.85	85.74	89.74	**100.00±0.00**
Meadow	66.89	75.78	65.08	68.29	**88.13**	61.29	53.75	54.81	71.65	68.40	69.02±4.50
Bare Soil	73.20	**91.08**	35.03	57.30	83.43	73.05	90.43	56.83	73.21	82.71	70.98±8.41
OA(%)	76.99	73.83	64.41	83.43	76.58	76.07	63.97	75.77	74.31	67.49	**85.89±1.32**
Kappa	0.72	0.68	0.57	0.73	0.71	0.73	0.55	0.71	0.70	0.65	**0.84±0.02**
Params	102,664	321,346	2,781,896	46,769	182,558	181,074	-	**8,651,280**	1,165,521	5,343,752	2,818,152
FLOPS	4,823,296	**217,649,820**	72,035,328	42,369,960	3,633,983	3,633,402	-	36,168,704	5,885,248	10,060,032	156,249,600
Time (s)	55.54	185.54	103.80	**931.67**	161.54	335.92	31.12	73.62	75.58	301.41	121.14

labeled samples via a same percentage per class. To improve the objectivity of the experiment, we performed an additional experiment with 180 source labeled samples on PaviaU-C dataset, and the performance gain of ToMF-B is significant (85.89 *vs*. 91.04, OA). Comparatively, the HyRANK dataset has more complicated spatial land-cover distribution and similar spectral distribution, which is a more challenging dataset for cross-scenes HSIC methods. Thus it is more objective to validate the effectiveness of different approaches. Obviously, the performance of the general backbones is poorer than those of other DA-based methods, especially for the evaluation index of the class-specific accuracy. The ToMF-B not only obtains the highest classification accuracy, but also distinguishes some samples hard to transfer, such as the samples belong to the Fruit Trees and Arable Land which cannot be correctly classified by most methods. Overall, the ToMF-B outperforms non-UDA based methods, e.g., CNN and ViT, which shows that the effectiveness of our method benefits from the proposed UDA technique rather than the backbones. Moreover, the ToMF-B consistently outperforms those UDA-based counterparts, yielding the superiority of the domain alignment method towards multi-level features and classification boundaries.

3) The qualitative results of compared methods: In addition to the quantitative analysis, the qualitative classification maps are shown in Fig. 6-8. We mainly compare the qualitative performance of different methods on H-S dataset, including the region ① in red rectangle box and the region ② in purple

TABLE VI
CLASSIFICATION ACCURACIES AND COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS ON THE TARGET SCENE LOUKIA BY USING 1000 LABELED SAMPLES FROM THE SOURCE SCENE DIONI.

Class Name	CNN	DBDA	ViT	DCFSL	CLDA	S-DMM	DCA	MCD	CLAN	TSTnet	**ToMF-B**
Dense Urban Fabric	11.21	0	9.18	4.39	1.70	5.37	**54.17**	21.38	7.96	9.05	30.28±11.46
Mineral Extraction Sites	3.17	9.44	0	17.13	45.63	11.64	0	10.16	1.44	17.44	**47.50±14.71**
Non Irrigated Arable Land	1.59	0.05	2.02	29.76	16.13	2.94	26.38	14.13	15.61	14.14	**28.64±11.43**
Fruit Trees	23.41	13.80	0	16.67	1.25	14.91	2.37	9.22	2.05	0	**16.66±9.48**
Olive Groves	5.23	0	7.23	12.75	0.18	27.48	17.14	50.53	32.68	65.31	**67.92±6.34**
Coniferous Forest	0	**59.03**	3.01	37.18	10.21	0	0	14.09	3.25	6.73	36.12±12.54
Dense Sclerophyllous Vegetation	66.68	70.73	**74.71**	42.43	65.32	72.18	65.85	71.17	71.10	69.03	71.21±2.49
Sparce Sclerophyllous Vegetation	52.16	6.07	31.34	37.84	72.96	56.65	**74.14**	64.78	65.07	62.36	51.92±4.37
Sparcely Vegetated Areas	12.19	15.84	27.37	**75.72**	63.06	28.96	53.83	34.25	31.01	41.09	38.27±9.28
Rocks and Sand	29.36	65.78	34.17	62.37	21.53	0	66.06	56.41	28.35	25.90	**67.82±13.18**
Water	**100.00**	**100.00**	**100.00**	85.47	**100.00**	93.58	**100.00**	88.00	95.25	99.63	**100.00±0.00**
Coastal Water	**100.00**	72.54	98.31	87.53	99.75	67.61	**100.00**	98.08	**100.00**	**100.00**	**100.00±0.00**
OA(%)	51.69	44.46	51.36	46.10	63.21	53.79	62.68	62.31	59.39	62.71	**65.53±1.85**
Kappa	0.42	0.36	0.44	0.37	0.54	0.43	0.53	0.52	0.49	0.55	**0.57±0.02**
Params	102,664	536,847	2,782,861	54,169	187,728	183,328	-	**8,661,128**	1,197,403	5,366,349	4,090,434
FLOPS	4,823,296	**383,500,404**	124,757,184	42,969,360	4,095,800	22,835,400	-	44,627,968	8,027,008	12,192,512	198,363,072
Time (s)	42.89	238.25	67.25	**1430.12**	86.95	302.57	37.69	159.32	92.50	353.34	361.59

rectangle box. Compared with other methods, the results of ToMF-B and MCD are closer to the real land-covers in red box even with the complex spatial distribution. The clear boundaries may be benefited from the boundaries-oriented domain alignment, which refines the classification boundaries. MCD is also based on the boundaries-oriented domain alignment, but the map of MCD is extremely noisy, this may be due to that the simple CNN cannot provide the rich feature representation for learning the discriminative domain-confused features like our multi-level feature extractor. Furthermore, the region ② in purple box is a region that a small amount of water is mixed in the vast land, thus samples belong to the Water prone to be misclassified as the Land since the model tends to learn a trivial solution. It is noteworthy that ToMF-B still has the capacity to identify them, indicating that the decision boundary of the proposed method may be more discriminative.

The ToMF-B achieve the excellent classification performance closing to the supervised results on S-H dataset, because that a large number of bands of H-S data provide more spectral information to help the model learn the transferable knowledge. The two scenes of PaviaU-C data have the higher spatial resolution but serious spectral drift, and the performance gap between ToMF-B and other UDA-based is significantly bigger than the other two datasets. The HyRANK dataset has the lower spatial resolution and most land cover categories, which is very challenging to UDA, and the ToMF-B still achieve the best classification performance. Overall, compared with some advanced methods, quantitative and qualitative experiments on three datasets convey that ToMF-B not only delivers the promising classification accuracy but also has more clear classification maps.

4) The computational complexity of compared methods: To clarify the trade-off between the performance of models and model sizes, the computational complexities of different methods are presented in Tables IV-VI. Note that the MCD has no learnable parameters since the DCA is based on the subspace transformation (i.e., traditional machine learning based method), we can only give the processing times. The results yield that the computational complexity of our model is slightly high, but overall processing times of ToMF-B on three datasets are acceptable. Moreover, the running time of some methods does not match their sizes of parameters, such as CLDA and S-DMM, because they need to predict the pseudo labels or calculate the metric distance of target scenes, which requires a lot of running time and memory. The model size of ToMF-B is not particularly large compared with other methods (i.e., fewer learnable parameters), which will avoid the mismatch between the model complexity and the size of the training set.

D. Ablation study and visualization

To further explore the effectiveness of our method and different modules, three ablation studies are carried out in this section. Firstly ,the ablation study of different components are performed. Then the performance gains or drops brought by different features are explored. Meanwhile, to intuitively evaluate the transfer performance of ToMF-B, the distribution of original samples and aligned features between domains are visualized via t-SNE.

1) Effects of different components: Table VII shows the classification performance achieved by ToMF-B with different modules. The symbol "✓" represents that the corresponding component is added into the ToMF-B, while the symbol "×" does not. The stage1, stage2 and Ta-R FE refer to local alignment, global (instance-level) alignment and task-related features decomposition, respectively. The results of Exp1 can be viewed as the *Baseline* without any additional components, that is, the results of hybrid model training by source-only data. The experimental results of Exp2 show that the UDA technique towards feature and classification boundaries brings the significant performance improvement. Note that the stage1 aims to reduce the distribution gap of local features output

TABLE VII
OA (%) OF TOMF-B WITH DIFFERENT MODULES ON THREE DATASETS.

Module	Stage1	Stage2	Ta-R FE	S-H	PaviaU-C	HyRANK
Exp1	×	×	×	89.71±1.45	79.66±3.38	57.19±3.16
Exp2	×	✓	×	94.35±0.76	83.73±2.16	63.80±2.43
Exp3	✓	✓	×	94.95±0.66	84.07±1.87	64.58±2.45
Exp4	✓	✓	✓	**95.54±0.53**	**85.89±1.32**	**65.53±1.85**

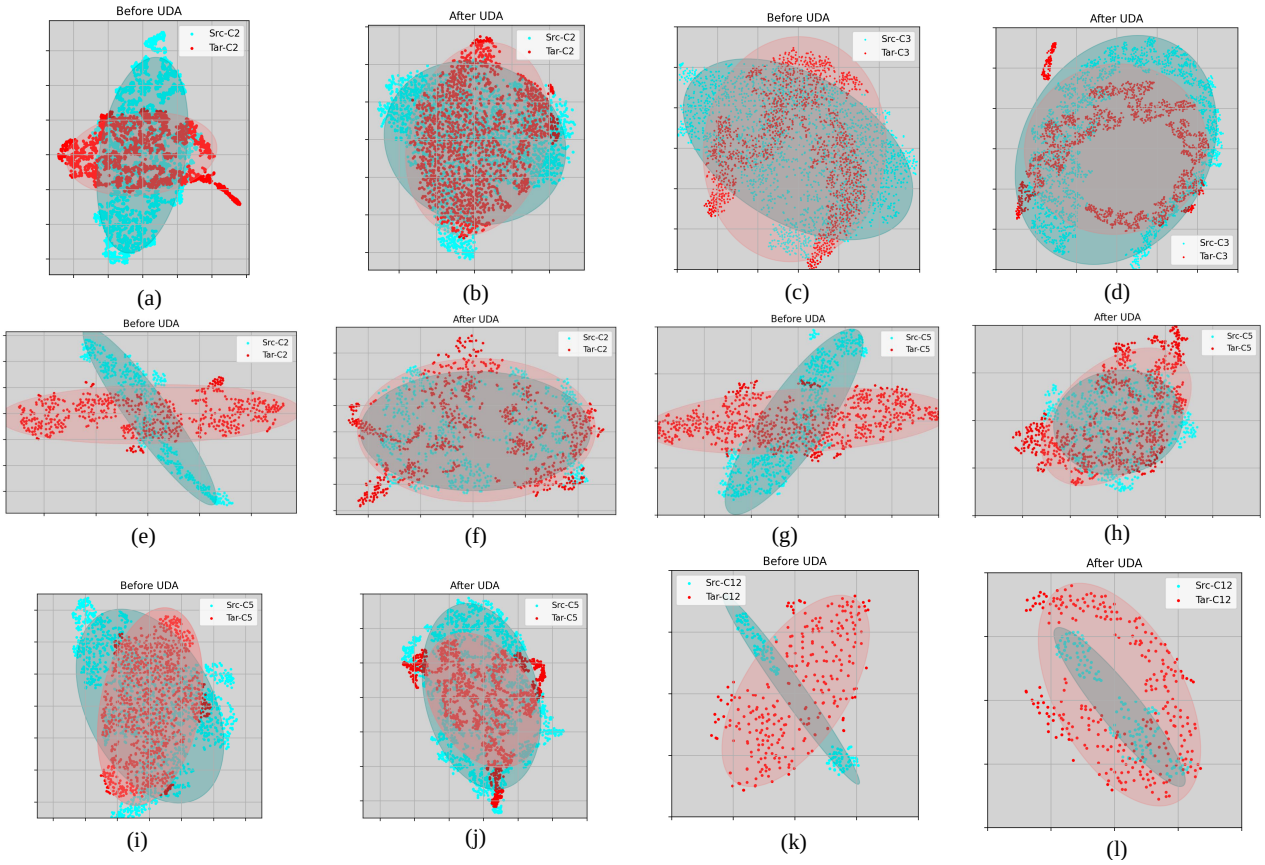


Fig. 9. The t-SNE visualization of embedding features on three datasets. The top row (a)-(d), middle row (e)-(h) and bottom row (i)-(l) represent the results of H-S dataset, PaviaU-C dataset and HyRANK dataset, respectively. The first and third columns are the original distribution gap, and the second and forth columns present the alignment performance after ToMF-B.

by CNN, but the alignment of final instance-level features is the basis of cross-domain classification tasks. Therefore, after carrying out the alignment of stage2, the performance gains brought by the local alignment will meaningfully demonstrate the effectiveness of stage1. Overall, the results of Exp2, Exp3 and Exp4 yield remarkable performance improvements after adopting the alignment of stage2, stage1 and the task-related features decomposition, indicating the effectiveness of different components.

2) UDA performance based on different features: To better explore the effectiveness of our UDA method and task-oriented feature decomposition, the performance with four different settings is investigated in this section. Specifically, the multi-level features extractor without UDA is termed as the *Baseline*, and this setting aims to present the improvement of our UDA method. *Baseline*+UDA, *Baseline*+UDA (Neg) and *Baseline*+UDA(Pos) represent the domain alignment guided by holistic feature vectors, task-irrelevant feature vectors and task-related feature vectors, respectively. Fig.9 presents the detailed results after alignment based on different features. The *Baseline* represents that a model trained on the source domain is directly applied to the target domain. The *Baseline+UDA* means the ToMF-B driven by holistic features, that is, the pure contribution of boundaries-oriented alignment. Compare with the *Baseline*, the boundaries-oriented alignment (*Baseline*+UDA) achieves the 7.93%, 8.25%, 11.03% improvement on H-S, PaviaU-C and HyRANK datasets, respectively, which shows the effectiveness of our boundaries-oriented UDA.

To verify the effectiveness of task-oriented feature decomposition method for UDA, the ToMF-B driven by task-related features are built for comparison. It can be observed that ToMF-B with task-related features yields the best performance, and it significantly boosts the accuracy of *Baseline*. Furthermore, compared with the UDA guided by holistic feature, the task-related features based transfer boosts the accuracy of *Baseline*+UDA 1.85%, 3.63% and 1.79% on three datasets, respectively. To see the destruction of task-irrelevant features, the performance ToMF-B with task-irrelevant features is reported. Note that, to avoid the training collapse, we just added the task-irrelevant features to the original features. It can be seen that the task-irrelevant features drops the accuracy even compare with *Baseline*, suggesting the harm to the transfer task even it is not in the extreme case, e.g, guided by purely task-irrelevant features.

3) T-SNE visualization of ToMF-B: In order to have an intuitive understanding for the effectiveness of ToMF-B, in Fig.10, two classes from each dataset are taken as toy examples to visualize the distribution gap before and after ToMF-B. The red and blue points represent the samples from source and target domain with same classes, and the corresponding ellipses are their 95% confidence regions. It is noteworthy that the original data distribution of fractional samples between

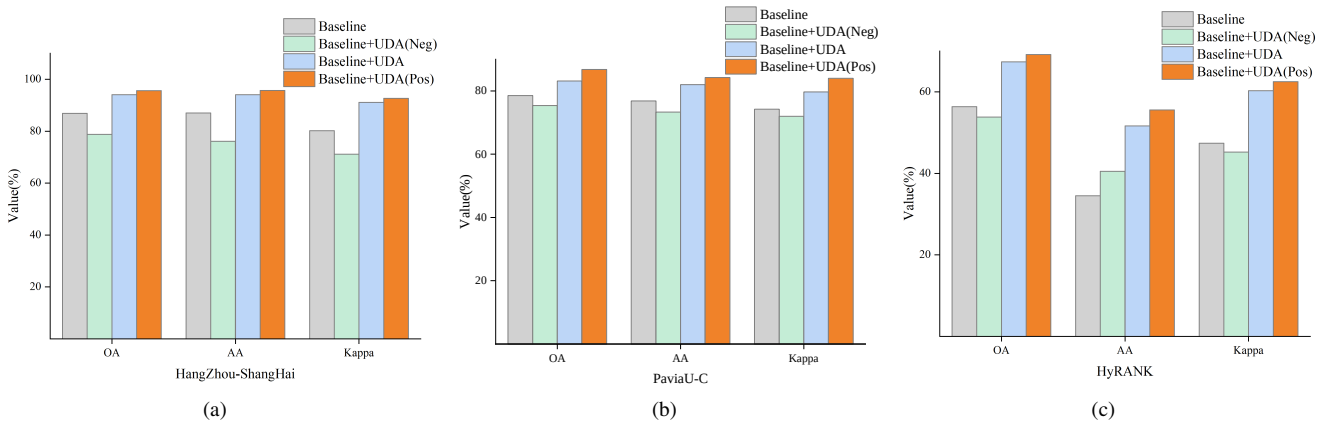


Fig. 10. Classification performance of ToMF-B with different ablation schemes on three datasets. (a) HangZhou-ShangHai. (b) PaviaU-C. (c) HyRANK.

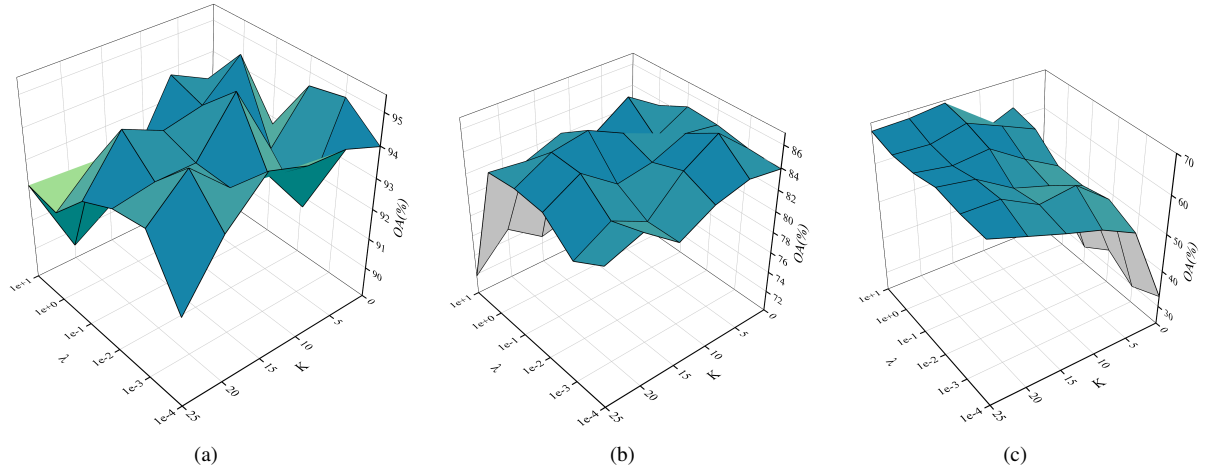


Fig. 11. Sensitivity analysis of parameter λ and K for the proposed ToMF-B model on three HSI datasets. (a) HangZhou-ShangHai. (b) PaviaU-C. (c) HyRANK.

source and target domains are overlapped, which is the reason that model of training on source domain can obtain the acceptable performance even directly test on the target domain. However, it can be seen the obviously distribution gap between two original domains via the confidence ellipses. Through the domain alignment of ToMF-B, the domain discrepancy is significantly reduced especially for the PaviaU-C and HyRANK datasets. Notably, the distribution gaps of the class 2 from H-S data, the class 2 from PaviaU-C data and class 12 from HyRANK data are almost orthogonal, which is extremely challenging for the model. After the alignment of ToMF-B, those distribution become uniform and co-directional, yield the superiority and robustness of our method.

E. Parameter analysis

In this section, the hyper-parameters sensitivity of ToMF-B model on is mainly discussed, including the trade-off parameter α in Eq(3) and λ in Eq(21), the repeat times of minimizing the Eq(15), k, and the base learning rate lr.

1) The analysis of parameter α: At beginning, the lr, λ and k are empirically fixed as $1e^{-3}$, $1e-1$ and 15 to explore the effect of parameter α. For the parameter α, the candidate interval is set to $[0:0.25:1]$. To show the advantage of hybrid model for UDA, based on the same UDA methods as ToMFB, the multi-level features extractor is replaced by the CNN (VGG model) network and Vision Transformer (ViT) model for comparison. The results of parameter α are reported in Tabel VII. Obviously, although based on the same UDA framework, multi-level feature fusion significantly boosts the performance of UDA when compared with the CNN- and ViT-only methods, indicating that the superiority of our hybrid features extractor. Especially for the PaviaU-C and HyRANK datasets, the the performance CNN-only method yields a significantly drop on PaviaU-C dataset, and the ViT-only exhibits the same case on HyRANK dataset. For the multi-level feature extractor based methods, the performance of ToMF-B with different α yields that the best performance on different datasets come from different settings of α, which can be inferred that the discriminative domain-confused representation may be learned from the feature space of different levels.

2) The sensitivity analysis of parameter λ and k: Then, the lr and α are fixed as $1e^{-3}$ and 0.25 to explore the sensitivity of λ and k. The sensitivity surfaces of λ and k are reported in Fig.9. The candidate interval of λ and K are set to $[1e-4:10:1e+1]$ and [0:5:25], respectively. There are several observations from the change trend of λ and k. First, on the PaviaU-C and HyRANK datasets, when the λ

TABLE VIII
OA (%) OF TOMF-B WITH DIFFERENT α ON THREE DATASETS.

α	0	0.25	0.5	0.75	1.0	CNN-only	ViT-only
H-S	95.06	**96.83**	95.22	95.86	94.71	94.82	92.58
PaviaU-C	84.93	**86.42**	85.89	85.23	76.32	76.92	83.35
HyRANK	63.69	64.18	65.53	**66.03**	60.27	60.91	56.52

is in the range of $1e+0$ to $1e-4$, the model can achieve the suboptimal results, which can be argued that the ToMF-B is less sensitive to the λ on these two datasets. For the H-S dataset, the performance of $\lambda = 1e-1$ superior the other λ, but the most results are stored at a well level, e.g., $\geq$93.00%. Notably that, toward the H-S and PaviaU-C, the performance of ToMF-B is relatively poor when the $\lambda = 1e+1$, this is resulted by the overfitting. Specifically, task-related features dominate the contribution to classification but the land-covers distribution of datasets are simple. In term of the parameter K, when the $K \geq 5$, the sensitivity of ToMF-B is low on H-S and PaviaU-C datasets. Therefore, we select the $K = 10$ as the default setting, which can obtain a win-win result, suboptimal performance and faster training time. Towards the HyRANK dataset, the performance is relatively better if the K is set to larger, but increase the training time at same time. Comprehensively, on the H-S and PaviaU-C datasets, the ToMF-B with $\lambda = 1e-1$ and $K = 10$ can achieve the suboptimal performance, and is less sensitive to these two parameters. Aiming to the HyRANK dataset, ToMF-B with $\lambda = 1e+0$ and $K \geq 15$ will obtain better results.

IV. CONCLUSION

In this paper, a novel unsupervised domain adaptation (UDA) framework towards multi-level features and decision boundaries (ToMF-B) is proposed for the cross-scene HSIC. The ToMF-B encourages the model to align the task-related features and learn the task-specific decision boundaries in parallel. Firstly, the design of ToMF-B is based on maximum classifier discrepancy and a two-stage alignment scheme, which can reduce the interdomain gap in a gradual manner and learn the discriminative decision boundaries. Secondly, a multi-level features extractor (generator) hybrid the CNN and Transformer is developed to enrich the feature representation of two domains, thereby it will be easier to learn the task-related and domain-confused features. To the best of our knowledge, this is the first work to introduce the hybrid model as generator for the cross-scene HSIC task. Furthermore, a task-oriented features decomposition method is leveraged to enhance the task-related features while suppressing the task-irrelevant features, which can avoid the harmless of task-irrelevant features to the transfer process and enable the aligned domain-invariant features to explicitly serve the classification tasks of each classifier. Extensive experiments and analysis suggest that the proposed ToMF-B outperforms the state-of-the-art HSI UDA methods on three benchmark datasets. In future work, we will extend the ToMF-B to the cross-scene HSIs from the different sensors.

REFERENCES

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

[2] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, 2014.

[3] S. Feng, S. Tang, C. Zhao, and Y. Cui, "A hyperspectral anomaly detection method based on low-rank and sparse decomposition with density peak guided collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[4] D. Haboudane, "Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, 2004.

[5] D. R. A. d. Almeida, E. N. Broadbent, M. P. Ferreira, P. Meli, A. M. A. Zambrano, E. B. Gorgens, A. F. Resende, C. T. de Almeida, C. H. do Amaral, A. P. D. Corte, C. A. Silva, J. P. Romanelli, G. A. Prata, D. de Almeida Papa, S. C. Stark, R. Valbuena, B. W. Nelson, J. Guillemot, J.-B. Féret, R. Chazdon, and P. H. S. Brancalion, "Monitoring restored tropical forest diversity and structure through uav-borne hyperspectral and lidar fusion," *Remote Sens. Environ.*, vol. 264, 2021.

[6] Z. Shao, H. Fu, D. Li, O. Altan, and T. Cheng, "Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation," *Remote Sens. Environ.*, vol. 232, 2019.

[7] G. Sun, H. Fu, J. Ren, A. Zhang, J. Zabalza, X. Jia, and H. Zhao, "Spassa: Superpixelwise adaptive ssa for unsupervised spatial-spectral feature extraction in hyperspectral image," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6158–6169, 2022.

[8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[9] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pizurica, and W. Philips, "Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 14, no. 11, pp. 2142–2146, 2017.

[10] J. Peng, W. Sun, H.-C. Li, W. Li, X. Meng, C. Ge, and Q. Du, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 10–43, 2022.

[11] S. Yang, J. Hou, Y. Jia, S. Mei, and Q. Du, "Superpixel-guided discriminative low-rank representation of hyperspectral images for classification," *IEEE Trans. Image Process.*, vol. 30, pp. 8823–8835, 2021.

[12] H. Zhai, H. Zhang, P. Li, and L. Zhang, "Hyperspectral image clustering: Current achievements and future lines," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 35–67, 2021.

[13] S. Huang, H. Zhang, and A. Pizurica, "A structural subspace clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[14] X.-J. Tang, X. Liu, P.-F. Yan, B.-X. Li, H.-Y. Qi, and F. Huang, "An mlp network based on residual learning for rice hyperspectral data classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] M. Lin, W. Jing, D. Di, G. Chen, and H. Song, "Multi-scale u-shape mlp for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, 2019.

[17] C. Zhao, W. Zhu, and S. Feng, "Superpixel guided deformable convolution network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3838–3851, 2022.

[18] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "Cnn-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, 2021.

[19] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, 2021.

[20] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[21] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2021.

[22] R. Thoreau, V. ACHARD, L. Risser, B. Berthelot, and X. BRIOTTET, "Active learning for hyperspectral image classification: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–24, 2022.

[23] C. Zhao, B. Qin, S. Feng, and W. Zhu, "Multiple superpixel graphs learning based on adaptive multiscale segmentation for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 3, 2022.

[24] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE Inst. Electr. Electron. Eng.*, vol. 109, no. 1, pp. 43–76, 2021.

[25] W. Hao and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, 2018.

[26] H. Lee, S. Eum, and H. Kwon, "Exploring cross-domain pretrained model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. IEEE, 2009, pp. 248–255.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," pp. 740–755, 2014.

[29] S. Jia, X. Liu, M. Xu, Q. Yan, J. Zhou, X. Jia, and Q. Li, "Gradient feature-oriented 3-d domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[30] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[31] J. Bai, S. Huang, Z. Xiao, X. Li, Y. Zhu, A. C. Regan, and L. Jiao, "Few-shot hyperspectral image classification based on adaptive subspaces and feature transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[32] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2022.

[33] 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. (CVPR)*, 2020, pp. 12 452–12 461.

[34] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–57, 2006.

[35] J. Xia, N. Yokoya, and A. Iwasaki, "Ensemble of transfer component analysis for domain adaptation in hyperspectral remote sensing image classification," pp. 4762–4765, 2017.

[36] Z. Li, X. Tang, W. Li, C. Wang, C. Liu, and J. He, "A two-stage deep domain adaptation method for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, 2020.

[37] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, 2020.

[38] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," *Pattern Recognit.*, vol. 77, pp. 306–315, 2018.

[39] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J Mach. Learn. Res. (JMLR)*, vol. 17, no. 1, pp. 2096–2030, 2016.

[40] X. Ma, X. Mou, J. Wang, X. Liu, J. Geng, and H. Wang, "Cross-dataset hyperspectral image classification based on adversarial domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4179–4190, 2021.

[41] Z. Liu, L. Ma, and Q. Du, "Class-wise distribution adaptation for unsupervised classification of hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 508–521, 2021.

[42] H. Wang, Y. Cheng, C. L. Philip Chen, and X. Wang, "Hyperspectral image classification based on domain adversarial broad adaptation network," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2021.

[43] Y. Qu, R. K. Baghbaderani, W. Li, L. Gao, Y. Zhang, and H. Qi, "Physically constrained transfer learning through shared abundance space for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–18, 2021.

[44] J. Zheng, W. Wu, S. Yuan, Y. Zhao, W. Li, L. Zhang, R. Dong, and H. Fu, "A two-stage adaptation network (tsan) for remote sensing scene classification in single-source-mixed-multiple-target domain adaptation (smt da) scenarios," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[45] J. Zheng, H. Fu, W. Li, W. Wu, W. Li, Y. Zhao, R. Dong, and L. Yu, "Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network," *ISPRS-J. Photogramm. Remote Sens.*, vol. 167, pp. 154–177, 2020.

[46] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014.

[47] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9646–9660, 2021.

[48] M. Ye, J. Chen, F. Xiong, and Y. Qian, "Learning a deep structural subspace across hyperspectral scenes with cross-domain vae," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2022.

[49] Y. Qin, L. Bruzzone, and B. Li, "Tensor alignment based domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9290–9307, 2019.

[50] T. Xu, W. Chen, P. WANG, F. Wang, H. Li, and R. Jin, "CDTrans: Cross-domain transformer for unsupervised domain adaptation," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=XGzk5OKWFFc

[51] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2021.

[52] Z. Fang, Y. Yang, Z. Li, W. Li, Y. Chen, L. Ma, and Q. Du, "Confident learning-based domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[53] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2200–2207.

[54] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.

[55] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3723–3732.

[56] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1-2, pp. 151–175, 2009.

[57] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2019, pp. 7404–7413.

[58] Y. Zhang, B. Deng, H. Tang, L. Zhang, and K. Jia, "Unsupervised multi-class domain adaptation: Theory, algorithms, and practice," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[59] L. Zhou and L. Ma, "Extreme learning machine-based heterogeneous domain adaptation for classification of hyperspectral images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 11, pp. 1781–1785, 2019.

[60] G. Wei, C. Lan, W. Zeng, Z. Zhang, and Z. Chen, "Toalign: Task-oriented alignment for unsupervised domain adaptation," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.

[61] 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 *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[62] M. Ye, Y. Qian, J. Zhou, and Y. Y. Tang, "Dictionary learning-based feature-level domain adaptation for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1544–1562, 2017.

[63] K. C. K. Z. . A. G. Karantzalos, Konstantinos, "Hyrank hyperspectral satellite dataset i (version v001)," *Int. Soc. Photogramm. Remote Sens., Tech. Rep*, 2018, doi:10.5281/zenodo.1222202.

[64] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, 2020.

[65] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2502–2511.

Chunhui Zhao received the BS and MS degree from Harbin Engineering University, in 1986 and 1989, respectively, and his PhD degree in Department of Automatic Measure and Control at Harbin Institute of Technology in 1998. He was a postdoctoral research fellow in the College of Underwater Acoustical Engineering of Harbin Engineering University. At present, he is working in the College of Information and Communication Engineering of Harbin Engineering University as a professor and doctoral supervisor. Prof. Zhao is a senior member of Chinese Electronics Academy. His research interests include digital signal and image processing, mathematical morphology and hyperspectral remote sensing image processing.

Boao Qin (Student Member, IEEE) is a Ph.D. candidate of Harbin Engineering University, China.

His main research areas are hyperspectral image processing and etc.

Feng Shou (Member, IEEE) is an associate Professor of Harbin Engineering University, China. He received his Ph.D. degree in 2019 from Harbin Institute of Technology, China.

His main research interests include remote sensing image processing, data mining, machine learning and etc.

Wenxiang Zhu is currently pursuing the PhD degree with Harbin Engineering University, Harbin, China.

His research interests include remote sensing image processing and hyperspectral image processing.

Lifu Zhang is currently a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China.

His research interest is hyperspectral image processing and its applications.

Jinchang Ren received the B.Eng., M.Eng. and D.Eng. degrees from the Northwestern Polytechnical University, Xi'an, China in 1992, 1997 and 2000, respectively, and the Ph.D. degree from the University of Bradford, Bradford, U.K., in 2019. He is currently a Professor with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K.

His research interests include image processing, computer vision, machine learning, and big data analytics.