LiCa: Label-Indicate-Conditional-Alignment Domain Generalization for Pixel-Wise Hyperspectral Imagery Classification

Zhe Gao, Bin Pan, *Member, IEEE*, Xia Xu, Tao Li, and Zhenwei Shi, *Member, IEEE*

***Abstract*— One of the major difficulties for hyperspectral imagery (HSI) classification is the hyperspectral heterospectra, which refers to the same material presenting different spectra. Although joint spatial–spectral classification methods can relieve this problem, they may lead to falsely high accuracy because the test samples may be involved during the training process. How to address the hyperspectral-heterospectra problem remains a great challenge for pixel-wise HSI classification methods. Domain generalization is a promising technique that may contribute to the heterospectra problem, where the different spectra of the same material can be considered as several domains. In this article, inspired by the theory of domain generalization, we provide a formulaic expression for hyperspectral heterospectra. To be specific, we consider the spectra of one material as a conditional distribution and propose a domain-generalization-based method for pixel-wise HSI classification. The key of our proposed method is a new label-indicate-conditional-alignment (LiCa) block that focuses on aligning the spectral conditional distributions of different domains. In the LiCa block, we define two loss functions—cross-domain conditional alignment and cross-domain entropy (CdE)—to describe the heterogeneity of HSI. Moreover, we have provided the theoretical foundation for the newly proposed loss functions, by analyzing the upper bound of classification error in any target domains. Experiments on several public datasets indicate that the LiCa block has achieved better generalization performance when compared with other pixel-wise classification methods.**

***Index Terms*— Conditional domain generalization, hyperspectral classification, hyperspectral heterospectra.**

I. Introduction

HYPERSPECTRAL imagery (HSI) classification has extensive practical applications [1], [2], [3], [4]. However, this task suffers a serious challenge due to the phenomenon of *hyperspectral heterospectra* [5], [6], that is, the same material presents different spectra [6], [7], [8], [9], [10]. This phenomenon presents a major obstacle to HSI classification, as the vast differences in spectra within a single class can lead to misclassification and reduced accuracy.

Deep spatial–spectral methods are essential for HSI classification by leveraging both the spatial and spectral information present in the data. Deep methods are typically divided into the following categories [8]: convolutional neural network (CNN)-based, autoencoder-based, deep-belief-network-based, and spatial-RNN-based approaches. Various CNN-based methods [9], [11], [12], [13], [14] extract spatial and spectral features simultaneously with the convolution kernels. Autoencoder methods [15], [16], [17] are symmetrical networks that extract compressed representations of HSI data to overcome the high dimensionality hampering the classification tasks. Deep-belief-network-based [18], [19] approaches are hierarchical architectures that obtain features in a layer-by-layer way with the assistance of restricted Boltzmann machines. Spatial-RNN-based approaches [20], [21], [22] process the spectral information of HSI data as time sequences and jointly leverage the spatial information.

However, spatial–spectral methods may suffer the problem that test samples are unavoidably included in the training stage, which may result in false high accuracy [23]. While test pixels do not participate in the spectral extraction, they are included in the spatial feature extraction due to the shortage of most HSI. This shortage typically occurs because labeled pixels are only present in a small portion of the whole HSI captured by the remote sensors.

To avoid the ambiguity caused by introducing test samples, pixel-wise-based HSI classification approaches remain a focus of academic attention. Two mainstream approaches [24] can be distinguished as unsupervised feature extraction and supervised feature extraction algorithms. Unsupervised feature extraction methods [25], [26], [27], [28], [29] that lack of need for labeled training samples is of great importance in the case of HSI classification, where the labeled data is usually limited in size. Supervised algorithms [30], [31], [32], [33], [34] extract class-separable features more effectively with the auxiliary label information and do not rely on modeling multiple prior assumptions of hyperspectral data [24].

However, research on addressing hyperspectral heterospectra without spatial information is of significant importance.

Manuscript received 6 May 2023; revised 5 July 2023; accepted 20 July 2023. Date of publication 7 August 2023; date of current version 15 August 2023. This work was supported in part by the National Key Resource and Development Program of China under Grant 2022ZD0160401 and Grant 2022YFA1003803; in part by the National Natural Science Foundation of China under Grant 62001251, Grant 62001252, Grant 62125102, and Grant 62272248; and in part by the Beijing-Tianjin-Hebei Basic Research Cooperation Project under Grant F2021203109. *(Corresponding author: Bin Pan.)*

Zhe Gao and Bin Pan are with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin 300071, China (e-mail: gzben01@gmail.com; panbin@nankai.edu.cn).

Xia Xu and Tao Li are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: xuxia@nankai.edu.cn; litao@nankai.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3300688

Due to the lack of labeled samples in HSI, pixel-wise algorithms may have difficulty learning transferable features performing well on potential practical target regions by extending the training sets. Deep domain adaptation has emerged as a promising solution to address the distribution discrepancy between domains in hyperspectral heterospectra. These approaches are mainly categorized into the following types [35]: discrepancy-based approaches and adversarial-based approaches. Discrepancy-based methods [36], [37] aim to match cross-domain marginal or conditional distributions by adding adaptation blocks to baseline networks. Adversarial-based approaches [38], [39] learn domain-invariant features through adversarial learning to improve the obfuscation of adversarial generated samples. Many other deep methods [40], [41], [42] have also received significant attention from the academic community.

Although domain-adaptation-based methods have provided an interesting idea for the hyperspectral-heterospectra problem, they may suffer two challenges [43]. First, domain adaptation methods require retraining at test time, resulting in costly excess time in practice. Second, using samples from the test sets during the training stage has been questioned to some extent.

To further improve the generalization ability of features in unavailable data domains, domain generalization methods are developed [43]. Domain generalization is an extensive task in transfer learning that aims at extracting generalizable features by only source data to perform well on any other unseen target data. We aggregate this strategy to the task since an observation that transferable features perform better on unseen domains, thereby addressing the problem of hyperspectral heterospectra. Little work on domain-generalized HSI classification has been proposed until now. A single-source generalization algorithm [44] that seeks to extract domain-invariant features through a domain-generative architecture is proposed recently and brings domain generalization strategy to HSI tasks. However, it is designed for cross-scene HSI classification while such a domain-generative method seems unnecessary for other tasks where the real source domains are accessible, thereby no need for generating source domains.

In this article, we propose a domain-generalization-based algorithm for HSI classification, where a new label-indicate-conditional-alignment (LiCa) block is developed to address the problem of hyperspectral heterospectra. We construct a conditional domain generalization approach for the task, as we observe that hyperspectral heterospectra can lead to overfitting and neglect cross-domain shared features that contain critical classification information. To address this, the LiCa block primarily focuses on the cross-domain conditional distributions $P(X|Y)$ of spectra, which is a formulaic expression of hyperspectral heterospectra. The LiCa block performs conditional feature alignment by defining two loss functions, cross-domain conditional alignment, and cross-domain entropy (CdE). The aligned features are then fed into a classifier during training. Furthermore, we prove a tight theoretical binding of the two loss functions with an upper bound theorem on the error. Overall, our proposed loss functions provide general guidance for conditional domain generalization. Our contribution can be summarized as follows.

1) We propose a conditional domain-generalization-based method for HSI classification, which targets the hyperspectral-heterospectra problem by extracting potential cross-domain features.
2) We develop a new LiCa block to align the conditional distributions of the material's spectra, where two loss functions are defined to describe the conditional discrepancy among domains.
3) We provide theoretical proof for the effectiveness of the LiCa block and give an error upper-bound theorem for the target domains when the source domains have been conditionally aligned.

II. Backgrounds

A. Domain Generalization and Domain Adaptation

Domain generalization and domain adaptation are two subdirections of transfer learning. A brief definition can be summarized as follows.

Domain Generalization: Given M source domains with labels $\{X_i\}_{i=1}^{M}$, $X_i = \{x_j, y_j\}_{j=1}^{N_i}$. N_i is the number of samples in domain X_i, x_j is the data, and y_j is the label. The goal is to train a well-behaved model on any unseen target domain with $\{X_i\}_{i=1}^{M}$.

Multiple strategies have been developed to improve domain generalization efficacy [43]. Notably, current popular directions include domain alignment, data augmentation, disentangled representations learning, and meta-learning. Domain alignment comprises a set of algorithms that aim to minimize the divergence of distributions across different domains (such as $P(X, Y)$ and $P(X)$) based on multiple criteria (e.g., KL divergence and moments), utilizing various network structures (e.g., networks that minimize divergence directly and those that do so adversarially). Data augmentation encompasses various algorithms that generate new datasets (i.e., new domains) from existing samples to prevent overfitting and improve generalization. Disentangled representations learning seeks to reduce the dimensionality of features while maximizing their reconstruction capabilities. Finally, meta-learning involves dividing the training data into meta-train and meta-test sets, with the former used to train the model and improve its performance on the latter. Collectively, these approaches have the potential to advance the field of domain generalization and enhance the generalizability of machine-learning models.

Domain Adaptation: Given a source domain with labels $X_S = \{x_i, y_i\}_{i=1}^{N_S}$ and a target domain without labels $X_T = \{x_j\}_{j=1}^{N_T}$. x_i and x_j are the data and y_i is the label. N_T and N_S are the numbers of samples in domains X_T and X_S. The goal is to train a well-behaved model on the target domain with X_S, X_T.

The key distinction between them lies in data availability from the target domains. While domain generalization focuses on learning models that can generalize to new domains that are unseen during training, domain adaptation seeks to adapt models trained on a source domain to perform well on a data-accessible target domain. Specifically, the latter directions

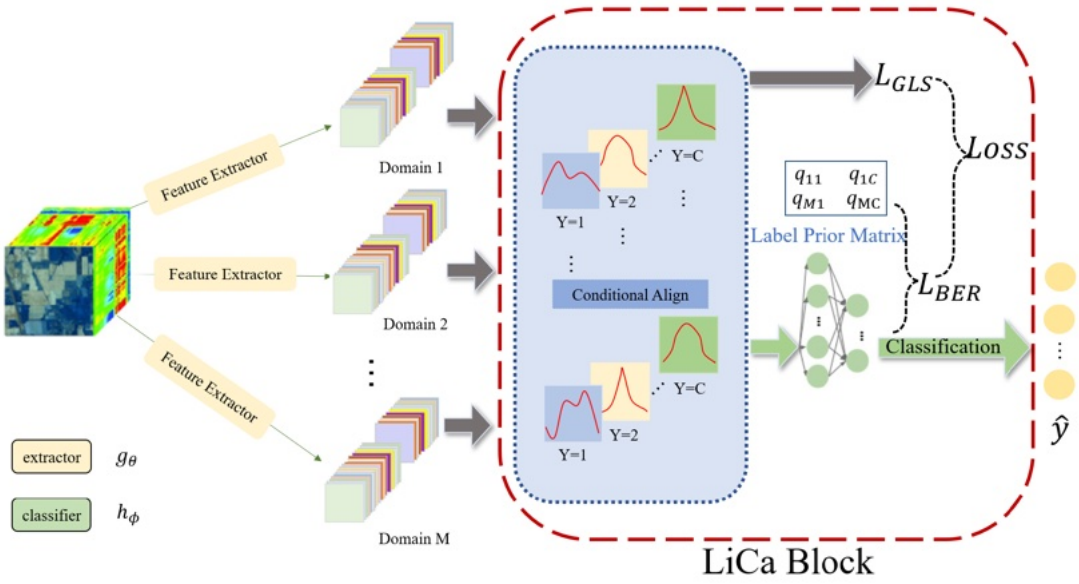


Fig. 1. Flowchart of the proposed algorithm. First, M source domains $\mathcal{X} = \{X_i\}_{i=1}^{M}$ with $X_i = \{x_j, y_j\}_{j=1}^{N_i}$ are input into the feature extractor $g_\theta(X)$ to perform initial feature extraction. Subsequently, the extracted features are aligned through the LiCa block, reducing divergence among distributions in different domains. Finally, the aligned features are put into a classifier for a vector classification result $\hat{Y}$.

leverage data from the target domain during training, whereas the former do not. Despite this difference, the research directions of the two directions share some similarities, particularly in the development of approaches to improve the generalizability and robustness of models in the face of domain shifts.

B. Importance-Weight Methods in Domain Adaptation

Importance-weight [45] methods are a collection of unsupervised domain adaptation techniques that have demonstrated notable effectiveness in addressing domain shift challenges. In particular, these methods seek to overcome the significant problem of large divergences between label distributions, which can severely hinder performance. To address this issue, the authors propose a suite of related methods, including IW-DANN, IW-CDAN, and IW-JAN, each built on a robust theoretical foundation. Compared to the base versions of these algorithms, the proposed methods yield significant improvements in performance. These promising results highlight the potential of these methods as a valuable approach to domain adaptation.

It is important to note that while importance-weight methods have demonstrated notable effectiveness in unsupervised tasks, they are not well-suited for domain generalization tasks. This limitation arises because these methods cannot fully leverage the label information across different domains simultaneously, leading to theoretical shortcomings as discussed in Section III. To address these challenges, we propose an improved architecture that is tailored specifically for domain generalization tasks.

III. Methodology

Our proposed framework is illustrated in Fig. 1 and comprises several key components. Specifically, it includes an HSI classification framework that is combined with a domain generalization strategy, as well as a LiCa block that aligns the conditional distribution and is integrated with a label prior matrix. We provide a detailed discussion of these two components in Sections III-A and III-B, including introducing several novel concepts based on reasonable observed assumptions. To highlight the effectiveness of our LiCa block, we employ an upper-bound theorem with rigorous proof that emphasizes the utility of the proposed concepts.

A. Overall Domain Generalization Classification Framework

Our method leverages a 1-D deep CNN (1D-DCNN) to extract pixel-wise features. Notably, the phenomenon of hyperspectral heterospectra, which refers to the diversity of spectral curves within the same class but across different domains, can be expressed as the difference between conditional distributions $P(X|Y)$. In our approach, we propose to address hyperspectral heterospectra by minimizing the divergence of $P(Z|Y)$, where Z denotes the extracted feature space. This is because the divergence of $P(X|Y)$ is inherent and cannot be eliminated. To achieve this, we introduce the LiCa block, which aligns the distributions of features from different domains by leveraging a label prior matrix. This approach enhances the robustness of the feature extractor by reducing the domain-specific information in the extracted features. The entire flowchart of our method is illustrated in Fig. 1. The steps of the LiCa algorithm are as follows.

1) The spectral-only dataset is separated into M source domains $\mathcal{X} = \{X_i\}_{i=1}^{M}$, $X_i = \{x_j, y_j\}_{j=1}^{N_i}$ and a label prior matrix is generated.
2) The data are input into a 1D-DCNN architecture, which serves as the extractor for obtaining the spectral features, denoted as $g_\theta(X)$.
3) The extracted features are then passed through the LiCa block, which aligns the distributions of features from different domains, and the resulting features are fed to a classifier $h_\phi(X)$ consisting of fully connected layers.

The output of the classifier is a vector $\hat{Y}$ representing the predicted labels.

B. LiCa Block

We aggregate the conditional alignment strategy, of which the following label prior matrix leads an instructive role, to the aforementioned 1D-DCNN. However, a blunt application of the domain adaptation method ignores the label information in the validation domain, which seriously hampers the performance of the extractor. Therefore, we utilize the label prior matrix for the entire label information from all source domains. As shown in Table I and Fig. 1, our proposed label prior matrix Q is a simple matrix that

$$Q = \begin{pmatrix} P_1(y=1) & P_1(y=2) & \cdots & P_1(y=C) \\ P_2(y=1) & P_2(y=2) & \cdots & P_2(y=C) \\ \vdots & \vdots & \ddots & \vdots \\ P_M(y=1) & P_M(y=2) & \cdots & P_M(y=C) \end{pmatrix} \quad (1)$$

where M is the number of source domains, C is the number of classes, and P_i is the mass probability in domain i. For the implementation of the proposed algorithm, we count the number of training samples in each class before the training stage to create the label prior matrix.

Many existing domain generalization methods belong to the fields of domain alignment, of which the central idea is to minimize the divergence between the source-invariant representations of learned domains. The alignment criteria are usually as follows:

$$L_{dA} = \sum_{i=1}^{M} \rho(D_i(Z), D_S(Z)) \quad (2)$$

where $Z = g_\theta(X)$ is the feature space and finally D_i denotes the distribution of the source domain i while D_S denotes the distribution of the public domain $\cup_{i=1}^{M} X_i$. ρ is a distance measure. However, our proposed LiCa block aligns the formulaic form $P(Z|X)$ of hyperspectral heterospectra. Guided by the label prior matrix, our proposed LiCa block achieves the condition of generalized label shift (GLS) through our designed cross-domain conditional alignment (CdCA) loss. The definition of these two terms is as follows.

Definition 1: A set of domains $\mathcal{X} = \{X_i\}_{i=1}^{M}$ is reaching the condition of GLS, when for any $i, j = 1, 2, \ldots M$, $c = 1, 2, \ldots C$, $D_i(X \mid Y = c) = D_j(X \mid Y = c)$ holds,

where D_i are the distributions in domain i.

Definition 2: CdCA loss function is defined as

$$L_{\text{CdCA}} = \sum_{i=1}^{M} \sum_{y=1}^{C} \text{MMD}(D_i(Z \mid y = i), D_S(Z \mid y = i)) \quad (3)$$

where MMD is the maximum mean discrepancy. M, C are defined the same as above. Theoretical analysis of the reason why extracted features can achieve GLS through our LiCa block is stated in Section III-C.

After the feature extractor, the classification layers follow, where a softmax function is utilized to obtain the vector-shaped prediction results. The commonly used cross-entropy loss function is typically applied to minimize the classification error

$$L_{\text{CE}} = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log\left(h_\phi(g_\theta(x))_y\right) \quad (4)$$

where y denotes the true label of data x, $|\mathcal{X}|$ is the number of training sets $\mathcal{X}$, and $h_\phi(g_\theta(x))_y$ is the yth number of vector $h_\phi(g_\theta(x))$.

However, theoretical analysis in the subsequent subsection reveals that the cross-entropy loss function is unsuitable for achieving conditional alignment. Consequently, we have introduced the CdE loss, which is integrated with the LiCa block during the training stage to fulfill a low balanced error rate (LBER) task. The definitions of these two terms are as follows.

Definition 3: For the classifier $\hat{Y}$ on D_i, the balanced error rate is

$$\text{BER}_{D_i}\left(\hat{Y}, Y\right) := \max_{c \in \{C\}} P_i\left(\hat{Y} \neq Y \mid Y = c\right). \quad (5)$$

Thus, the task of LBER is to

$$\min_{\phi 1, \theta} \max_{j \in \{M\}} \text{BER}_{D_j}\left(\hat{Y}, Y\right). \quad (6)$$

According to the definition, it is evident that BER represents the worst-case classification error across all classes. Therefore, if BER is bounded effectively (i.e., even the worst-case error rate for individual classes is well-contained), it naturally follows that not only the overall error rate will be constrained within a small upper bound, but also the error rates for each individual class will be balanced and limited within same level upper bounds.

Definition 4: The definition of the CdE loss function is as follows:

$$L_{\text{CdE}} = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{\log\left(h_\phi(g_\theta(x))_y\right)}{C \cdot P_{d_x}(Y = y)} \quad (7)$$

where d_x is the domain label of data x, y denotes the true label of x, $|\mathcal{X}|$ is the number of training sets $\mathcal{X}$, and $h_\phi(g_\theta(x))_y$ is the yth number of vector $h_\phi(g_\theta(x))$.

To summarize, the entire loss function in the training stage is

$$L = L_{\text{CdCA}} + L_{\text{CdE}}. \quad (8)$$

C. Theoretical Analysis

In comparison to other nonconditional similar approaches, conditional domain generalization is more theoretical. The fundamental concept behind this approach is to consider the difference between domains as the divergence between distributions derived from those domains. Zhang et al. [46] introduced the concept of conditional shift, where $P_i(X|Y) \neq P_j(X|Y)$, and proposed a reasonable and widely accepted assumption: as the number of participating source domains increases, the divergence between distributions of features in different domains after alignment will converge to zero for all potential target domains, which immediately leads to the following assumption.

Algorithm 1 Overall Training Procedure

Let $\mathcal{X} = \cup_{i=1}^{M} X_i$ be the *source domains*
T be the hyperparameter *Epoch*
λ be the hyperparameter *Learning Rate*
B be the number of batches.
INITIALIZE networks g_θ, h_ϕ randomly
LP Matrix $Q \leftarrow \mathbf{0}_{M\times C}$
for $i \leftarrow 1$ **to T Do** :
Generate LP Matrix $Q_{ij} \leftarrow P_i(y = j)$
// We create a matrix depending on the data already fed into the model, details in Eq.1.
for $j \leftarrow 1$ **to B Do**
Extract Feature $\mathcal{Z}_j = g_\theta(\mathcal{X}_j)$
// $\mathcal{X}_j$ is data in current batch j.
Input into LiCa Block $h(\mathcal{Z}_j)$
// After put into the LiCa block, not only will the features from different domains be aligned, a loss L^j_{CdCA} evaluating the divergence is also be computed.
Compute CdE Loss L^j_{CdE} with $h(\mathcal{Z}_j)$
end
Compute total loss $L = L_{CdCA} + L_{CdE}$
end
Update θ , ϕ by gradient descent

Assumption 1: Given M source domains $\mathcal{X} = \{X_i\}_{i=1}^{M}$ that satisfy the GLS condition. Then for any unseen target domain X_T, GLS will also hold for $\{\mathcal{X}, X_T\}$ as M grows large enough.

Based on the assumptions discussed above, we can derive a theorem with rigorous proof that provides an upper bound on the error in the target domains. Specifically, the theorem states that minimizing the term $\max_{i\le M} \mathrm{BER}_{D_i}(h(Z), Y)$ is sufficient to control the error in the unseen target domains. Now we are available to announce an upper-bound theorem.

Theorem 1: If $Z = g(\mathcal{X})$, then for any classifier h and any unseen target domain X_T, $\epsilon_S(h(Z)) + \epsilon_T(h(Z)) \le 2\max_{i\le M} \mathrm{BER}_{D_i}(h(Z), Y)$ holds when M grows great enough. Here, ϵ_S and D_S denote the classification error and the distribution of the common domain $X_S := \cup_{i=1}^{M} X_i$.

Proof: By assumption 1, when M grows enough, GLS holds for $\mathcal{X} \cup X_T$. Thus,

$$D_i(X \mid Y = c) = D_j(X \mid Y = c) \quad i, j = 1, 2, \ldots M, T,\ c = 1, 2, \ldots C. \tag{9}$$

By the law of total probability

$$\begin{aligned}
\epsilon_i(\hat{Y}) + \epsilon_T(\hat{Y}) &= P_i(Y \neq \hat{Y}) + P_T(Y \neq \hat{Y}) \\
&= \sum_{c=1}^{C} \sum_{j\neq c} P_i(\hat{Y} = j | Y = c) P_i(Y = c) \\
&\quad + P_T(\hat{Y} = j | Y = c) P_T(Y = c) \\
&= \sum_{c=1}^{C} \sum_{j\neq c} P_i(\hat{Y} = j | Y = c)(P_i(Y = c) \\
&\quad + P_T(Y = c)) \\
&= \sum_{c=1}^{C} P_j(\hat{Y} \neq |Y = c)(P_i(Y = c) \\
&\quad + P_T(Y = c)) \\
&\le \max_{c\le C} P_j(\hat{Y} \neq Y | Y = c) \sum_{c=1}^{C} (P_i(Y = c) \\
&\quad + P_T(Y = c)) \\
&= 2\mathrm{BER}_{D_i}(\hat{Y}, Y).
\end{aligned} \tag{10}$$

Let $q_1, q_2, \ldots, q_M$ be the contributing ratio of the error rates on different domains, thereby satisfying

$$\sum_{i=1}^{M} q_i = 1, \sum_{i=1}^{M} q_i \epsilon_i(\hat{Y}) = \epsilon_S(\hat{Y}). \tag{11}$$

Thus, we get

$$\begin{aligned}
\epsilon_S(\hat{Y}) + \epsilon_T(\hat{Y}) &= \epsilon_T(\hat{Y}) + \sum_{i=1}^{M} \epsilon_i q_i \\
&= \sum_{i=1}^{M} \epsilon_T q_i + \sum_{i=1}^{M} \epsilon_i q_i \\
&= \sum_{i=1}^{M} (\epsilon_i + \epsilon_T) q_i \\
&\le 2 \sum_{i=1}^{M} q_i \mathrm{BER}_{D_i}(\hat{Y}, Y) \\
&\le 2 \sum_{i=1}^{M} q_i \max_{i\le M} \mathrm{BER}_{D_i}(h(Z), Y) \\
&= 2 \max_{i\le M} \mathrm{BER}_{D_i}(h(Z), Y).
\end{aligned} \tag{12}$$

□

This theorem highlights the efficacy of our LiCa block in aligning the conditional distributions of features across diverse domains while introducing a theoretically grounded metric BER to handle it. To tackle the nontrivial challenge of minimizing BER, we introduce a practical application of the L_{CdE} loss function. To substantiate the effectiveness of this loss function, we present the following theorem.

Theorem 2: If $Z = g(\mathcal{X})$ and fix M as the total number of domains. Then, to minimize $\max_{i\in\{M\}} \mathrm{BER}_{D_i}(\hat{Y}, Y) = \max_{\substack{c\in\{C\} \\ i\in\{M\}}} D_i(\hat{Y} \neq Y \mid Y = c)$ is equal to minimize $L_{\mathrm{CdE}} = -(1/|\mathcal{X}|) \sum_{x\in\mathcal{X}} (\log(h_\phi(g_\theta(x))_y)/C \cdot D_{d_x}(Y = y))$.

Here, $\{C\} = \{1, 2, \ldots, C\}$ and $\{M\} = \{1, 2, \ldots, M\}$. C, M are the total numbers of classes and domains.

Proof: First, we simplify the target formula L_{CdE}

$$\begin{aligned}
L_{\mathrm{CdE}} &= -\frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \frac{\log\left(h_\phi(g_\theta(x))_y\right)}{C \cdot D_{d_x}(Y = y)} \\
&= -\frac{1}{|\mathcal{X}|} \sum_{i\in\{M\}} \sum_{c\in\{C\}} \sum_{\substack{x\in X_i \\ y=c}} \frac{\log\left(h_\phi(g_\theta(x))_c\right)}{C \cdot D_i(Y = c)} \\
&= -\frac{1}{C \cdot |\mathcal{X}|} \sum_{i\in\{M\}} \sum_{c\in\{C\}} \frac{\sum_{x\in X_i, y=c} \log\left(h_\phi(g_\theta(x))_c\right)}{D_i(Y = c)}.
\end{aligned} \tag{13}$$

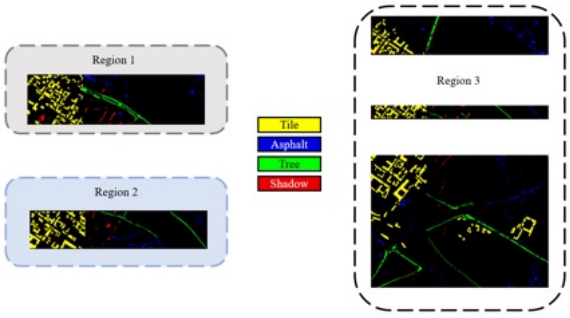


Fig. 2. Region separation of the Pavia dataset. We cut two regions horizontally from the original image as the source domain, and the remaining part as the target domain.

Therefore, we have an equivalent task of $\min_{\phi,\theta} L_{\text{CdE}}$

$$\begin{aligned}
&\min_{\phi,\theta} L_{\text{CdE}} \\
&\quad \Leftrightarrow \min_{\phi,\theta} -\frac{1}{C\cdot|\mathcal{X}|}\sum_{i\in\{M\}}\sum_{c\in\{C\}}\frac{\sum_{\substack{x\in X_i\\ y=c}}\log\left(h_\phi(g_\theta(x))_c\right)}{D_i(Y=c)} \\
&\quad \Leftrightarrow \min_{\phi,\theta} -\sum_{i\in\{M\}}\sum_{c\in\{C\}}\frac{\sum_{\substack{x\in X_i\\ y=c}}\log\left(h_\phi(g_\theta(x))_c\right)}{D_i(Y=c)}. \qquad (14)
\end{aligned}$$

The term $\sum_{\substack{x\in X_i\\ y=c}}\log(h_\phi(g_\theta(x))_c)$ of the above formula is just a cross-entropy loss that can measure error rate $D_i(\hat{Y}=Y, Y=c)$ of the method. Therefore, to maximize $D_i(\hat{Y}=Y, Y=c)$ is equal to minimize $-\sum_{\substack{x\in X_i\\ y=c}}\log(h_\phi(g_\theta(x))_c)$

$$\begin{aligned}
&\min_{\phi,\theta} L_{\text{CdE}} \\
&\quad \Leftrightarrow \max_{\phi,\theta}\sum_{i\in\{M\}}\sum_{c\in\{C\}}\frac{D_i\left(\hat{Y}=Y, Y=c\right)}{D_i(Y=c)} \\
&\quad \Rightarrow \max_{\phi,\theta}\max_{\substack{c\in\{C\}\\ i\in\{M\}}}\frac{D_i\left(\hat{Y}=Y, Y=c\right)}{D_i(Y=c)} \\
&\quad \Rightarrow \min_{\phi,\theta}\max_{\substack{c\in\{C\}\\ i\in\{M\}}} 1-\frac{D_i\left(\hat{Y}=Y, Y=c\right)}{D_i(Y=c)} \\
&\quad \Rightarrow \min_{\phi,\theta}\max_{\substack{c\in\{C\}\\ i\in\{M\}}}\frac{D_i(Y=c)-D_i\left(\hat{Y}=Y, Y=c\right)}{D_i(Y=c)} \\
&\quad \Rightarrow \min_{\phi,\theta}\max_{\substack{c\in\{C\}\\ i\in\{M\}}}\frac{D_i\left(\hat{Y}\neq Y, Y=c\right)}{D_i(Y=c)} \\
&\quad \Rightarrow \min_{\phi,\theta}\max_{\substack{c\in\{C\}\\ i\in\{M\}}} D_i\left(\hat{Y}\neq Y|Y=c\right). \qquad (15)
\end{aligned}$$

Notice that $\max_{i\in\{M\}}\text{BER}_{D_i}(\hat{Y}, Y) = \max_{\substack{c\in\{C\}\\ i\in\{M\}}} D_i(\hat{Y}\neq Y|Y=c)$, we have

$$\begin{aligned}
\min_{\phi,\theta} L_{\text{CdE}} &\Rightarrow \max_{\substack{c\in\{C\}\\ i\in\{M\}}} D_i\left(\hat{Y}\neq Y|Y=c\right) \\
&= \min_{\phi,\theta}\max_{i\in\{M\}}\text{BER}_{D_i}\left(\hat{Y}, Y\right). \qquad (16)
\end{aligned}$$

□

To further improve the conditional domain alignment in hyperspectral heterospectra, we propose the use of the CdE loss instead of the commonly used CE loss in typical classic conditional alignment domain generalization methods [47], [48]. In cross-domain datasets, the conditional distributions of features across different classes often exhibit diversity, which makes it challenging for the cross-entropy loss to achieve the necessary GLS when the divergence of distributions is pronounced. In contrast, the CdE loss incorporates domain knowledge to facilitate more effective conditional domain alignment. Neglecting this domain information can lead to a significant reduction in classification accuracy. Our upper-bound theorem demonstrates that this reduction can be overcome by replacing the cross-entropy with the proposed CdE loss, which has a solid theoretical foundation.

TABLE I
DOMAIN SPLITTING FOR EUROCROPS DATASETS

	Label	Items in Domain 1	Items in Domain 2	Items in Domain 3
Eurocrops	Arable Crops	oats tobacco soya barley hops wheat rye cotton	triticale flax millet hemp maize rice	potatoes sugar beat plants harvest green
	Permanent Crops	vineyard olive citrus	nuts berry	nurseries other permanent crops
Pavia City	C-1	Trees in region 1	Trees in region 2	Trees in region 3
	C-2	Asphalt in region 1	Asphalt in region 2	Asphalt in region 3
	C-3	Tile in region 1	Tile in region 2	Tile in region 3
	C-4	Shadows in region 1	Shadows in region 2	Shadows in region 3
Indian Pines	C-1	Corn-min-tiles in region 1	Corn-min-tiles in region 2	——
	C-2	Soybean-no-tile in region 1	Soybean-no-tile in region 2	Soybean-no-tile in region 3
	C-3	Soybean-min-tile in region 1	Soybean-min-tile in region 2	Soybean-min-tile in region 3
	C-4	Grass/Trees in region 1	Grass/Trees in region 2	Grass/Trees in region 3
	C-5	Grass/Pasture in region 1	Grass/Pasture in region 2	Grass/Pasture in region 3

IV. EXPERIMENTS

A. Dataset

1) Eurocrops: Eurocrops [49] is a dataset collection combining all publicly available self-declared crop reporting datasets from most countries of the European Union. All of the data are captured by Sentinel-II. The hyperspectral data is stripped of the spatial information and stored pixel by pixel. Every pixel has 13 channels. The criterion of labeling is named the Hierarchical Crop and Agriculture Taxonomy (HCAT) version 2. As shown in Table I, the split provides a feasible way of domain separation for the task. We endow label 1 to the arable crops and label 0 to the permanent crops.

2) Pavia: The Pavia dataset [50] has 1096 × 715 pixels and 102 bands and 1.3 m spatial resolution. Because hyperspectral heterospectra is of greater possibility to occur when the ground items are from the same categories but distributed discretely, we select all the vertical widely distributed categories. Therefore, we choose "trees," "asphalt," "tiles," and "shadows" from all nine categories to implement our experiments. 0.05 of the sorted dataset is selected as the training set and 0.95 of the dataset is the test set. The training set consists of 133 "trees," 157 "asphalts," 986 "tiles," and 84 "shadows." Then, the original cubic-shaped data are rearranged into a two-dimension matrix to simulate the lack of spatial information. The domains are produced through the separation of the Pavia dataset as shown in Fig. 2 and Table I, while the visible classification is shown in Fig. 3.

3) Indian Pines: The Indian Pines dataset is an HSI dataset collected by the AVIRIS sensor on a farm in Indiana, the United States. The size of the dataset is 145 × 145 and the original number of the spectral channel is 220. Since

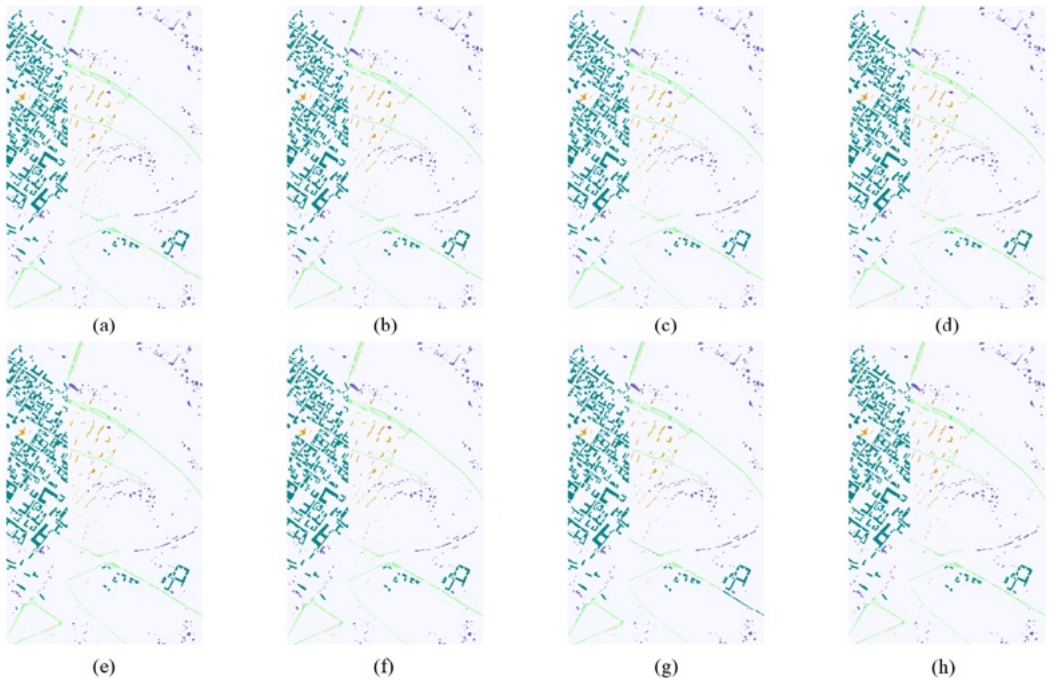


Fig. 3. Classification of Pavia dataset. (a) LiCa. (b) GLS [48]. (c) MMD. (d) CAD [52]. (e) VREx [54]. (f) ERM. (g) RF [51]. (h) SVM [31].

TABLE II
NUMBER OF EACH CLASS IN TRAINING SETS

Pavia City Dataset				Eurocrops Dataset				Indian Pines Dataset			
Class	No. in Domain 1	No. in Domain 2	Total No.	Class	No. in Domain 1	No. in Domain 2	Total No.	Class	No. in Domain 1	No. in Domain 2	Total No.
C-1	1011	914	1925	C-1	35343	4314	49046	C-1	282	127	409
C-2	151	126	277	C-2	13612	931	5245	C-2	6	198	204
C-3	196	123	329					C-3	127	121	248
C-4	109	78	187					C-4	27	58	85
								C-5	236	727	963

the 104th–108th, 220th, and 150th–163rd cannot be reflected by water, the corrected Indian Pines dataset removed these channels, and therefore 200 bands remain. Since the spatial resolution is at a very low rate of 20 m, the classification difficulty is quite large for the mixed pixels conducted by the low-resolution level. For the Indian Pines dataset, we split the HSI differently as denoted in Fig. 4 and Table I. All five classes of samples that are commonly distributed in both Region 1 and Region 2 are separated to construct the training set and validation set. Since samples with the label "Corn-min-tiles" do not exist, the test set which is constructed by the samples in region 3 only contains four classes despite the "Corn-min-tiles."

The training set consists of 409 "Corn-min-tiles," 204 "Soybean-no-tile," 248 "Soybean-min-tile," 85 "Grass/Trees," and 963 "Grass/Pasture." The original cubic-shaped data are also rearranged and the details of the training dataset are illustrated in Table II. And the visible classification on this dataset is shown in Fig. 5.

B. Comparative Experiments

Our experiments are implemented based on Domainbed [55], a popular test-bed for domain generalization to streamline reproducible and rigorous research.

The mainstream of this section can be divided into two parts. On the one hand, we compare the proposed method with several SOTA domain generalization strategies such as CAD [54], VREx [52], CausIRL [53], MMD (unconditional alignment strategy), and GLS (conditional alignment strategy with only loss L_{CdCA}). On the other hand, it is compared with multiple spectral-only HSI classification methods such as SVM [31], RF [51], and 1D-DCNN (called ERM in Tables III–V). The consequences are in Tables III–V, illustrating the efficiency of our proposed method.

1) Eurocrops: In hyperspectral classification tasks, the commonly used evaluation metrics include overall accuracy (OA), average accuracy (AA), accuracy for each class, and the kappa coefficient. In our experiments, we use these indicators to evaluate the performance of our proposed method. However, due to the nature of the data structure, the target domain for testing contains only one class of samples. As a result, all the aforementioned evaluation metrics share the same value.

The performance of our baseline, which is only a single three-layer 1D-CNN, is mightily worse than the RF and is merely comparable to the SVM. However, when the simple network is combined with the domain generalization strategy, the performance will be apparently improved. CausIRL [53], VREx [52], CAD [54], MMD, GLS, and our proposed LiCa

TABLE III

ACCURACY OF TEST SAMPLES IN THE TARGET DOMAIN OF THE PAVIA CITY DATASET (%)

Method		Pavia Dataset						
Type	**Name**	1	2	3	4	OA	AA	$\kappa \times 100$
Classic	**SVM** [32]	99.18	**100**	98.62	99.90	99.23	99.42	**98.49**
Methods	**RF** [52]	99.51	99.69	83.87	**100**	97.11	95.77	95.10
	ERM [34]	99.09	**100**	98.31	99.71	98.98	98.71	97.32
Nonaligned	**VREx** [53]	98.95	99.72	98.20	99.71	98.96	99.14	97.87
	CausIRL [54]	**99.82**	**100**	96.79	**100**	99.32	99.15	**98.71**
Strategies	**CAD** [55]	98.40	99.95	97.52	98.47	98.46	98.58	97.15
	CondCAD [55]	98.25	99.60	98.20	**100**	98.02	98.34	94.93
Aligned	**MMD**	98.19	**100**	98.38	99.31	98.34	98.97	97.01
Strategies	**GLS** [49]	98.72	**100**	99.10	99.81	99.01	99.41	98.03
	LiCa	99.57	**100**	**99.14**	99.71	**99.58**	**99.61**	98.21

TABLE IV

ACCURACY OF TEST SAMPLES IN THE TARGET DOMAIN OF THE EUROCROPS DATASET (%)

Method		EuroCrops Dataset			
Type	**Name**	1	OA	AA	$\kappa \times 100$
Classic	**SVM** [32]	67.84	67.84	67.84	67.84
Methods	**RF** [52]	80.31	80.31	80.31	80.31
	ERM [34]	66.57	66.57	66.57	66.57
Nonaligned	**VREx** [53]	69.02	69.02	69.02	69.02
	CausIRL [54]	70.86	70.86	70.86	70.86
Strategies	**CAD** [55]	70.02	70.02	70.02	70.02
	CondCAD [55]	70.70	70.70	70.70	70.70
Aligned	**MMD**	71.16	71.16	71.16	71.16
Strategies	**GLS** [49]	80.32	80.32	80.32	80.32
	LiCa	**82.81**	**82.81**	**82.81**	**82.81**

TABLE V

ACCURACY OF TEST SAMPLES IN THE TARGET DOMAIN OF THE INDIAN PINES DATASET (%)

Method		Indian Pines Dataset						
Type	**Name**	2	3	4	5	OA	AA	$\kappa \times 100$
Classic	**SVM** [32]	61.01	77.92	0.12	54.79	31.41	48.46	10.58
Methods	**RF** [52]	40.67	83.78	4.08	50.11	31.77	44.66	7.90
	ERM [34]	67.58	87.48	4.95	49.31	33.13	52.33	8.61
Nonaligned	**VREx** [53]	66.14	79.54	3.21	51.37	33.90	50.06	11.21
	CausIRL [54]	69.03	81.20	3.10	58.61	35.29	53.03	10.73
Strategies	**CAD** [55]	62.54	86.67	2.23	45.37	33.78	49.20	8.24
	CondCAD [55]	62.43	72.15	**7.66**	48.40	32.26	47.66	**12.81**
Aligned	**MMD**	66.03	**87.15**	5.36	52.29	34.62	52.70	8.76
Strategies	**GSL** [49]	56.57	68.88	3.09	**63.55**	35.23	48.02	7.62
	LiCa	**81.22**	83.31	7.44	40.51	**35.41**	**53.12**	9.78

all achieved a significant improvement while among them GLS and LiCa perform the best. We analyze that the better performance of the strategy is because of its improved generalizability. Moreover, the apparently better performance of conditional alignment methods (GLS and LiCa) compared with other (CausIRL [53], VREx [52], and CAD [54]) methods is due to the unique characteristics of the HSI dataset. The Eurocrops dataset has a giant intraclass diversity, and

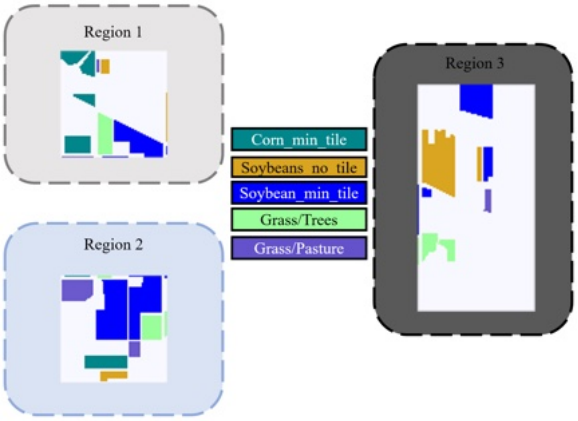


Fig. 4. Region separation of the Indian Pines dataset. We let the left side of the image be the two source domains and the right side as the target domain.

TABLE VI

ACCURACY OF TEST SAMPLES IN ABLATION EXPERIMENTS IN THE TARGET DOMAIN (%)

	ERM	MMD	CdCA	GLS	LiCa
Eurocrops	66.57	77.16	67.86	80.32	**82.81**
Pavia City	98.98	98.34	99.50	99.01	**99.58**
Indian Pines	33.13	34.62	33.51	35.23	**35.41**

TABLE VII

PHENOMENON OF HYPERSPECTRAL HETEROGENEITY IN PAVIA AND EUROCROPS

	EuroCrops		Pavia City		Indian Pines	
Domain	**test ratio**	**accuracy**	**test ratio**	**accuracy**	**test ratio**	**accuracy**
1(source)	0.2	94.33	0.95	99.58	0.5	95.13
2(source)	0.2	93.72	0.95	99.41	0.5	92.28
3(target)	–	66.57	–	98.98	–	33.13

therefore the conditional-alignment strategy that considers the cross-domain similarity of each class performs better than other domain generalization methods.

2) Pavia Dataset: As illustrated in Table III, our LiCa method achieves the best performance under indicators C-1 accuracy, C-2 accuracy, C-3 accuracy, OA, and AA while performing a second best rank under the indicator C-1 accuracy, C-3 accuracy, and the kappa, which have a mere difference between the performance of the first place.

The experimental results indicate that all three nonconditional-aligned strategies perform even worse than the standard ERM approach. Our analysis suggests that the reason why directly applying these three strategies leads to poor performance is that the hyperspectral-heterogeneity phenomenon is not significant among different regions of the Pavia city HSI dataset. Consequently, extracting potential domain-invariant features becomes challenging and less competitive compared to nondomain-generalization methods. Despite the adverse conditions of the HSI dataset, our proposed LiCa block still achieves promising results, demonstrating the efficiency of our approach.

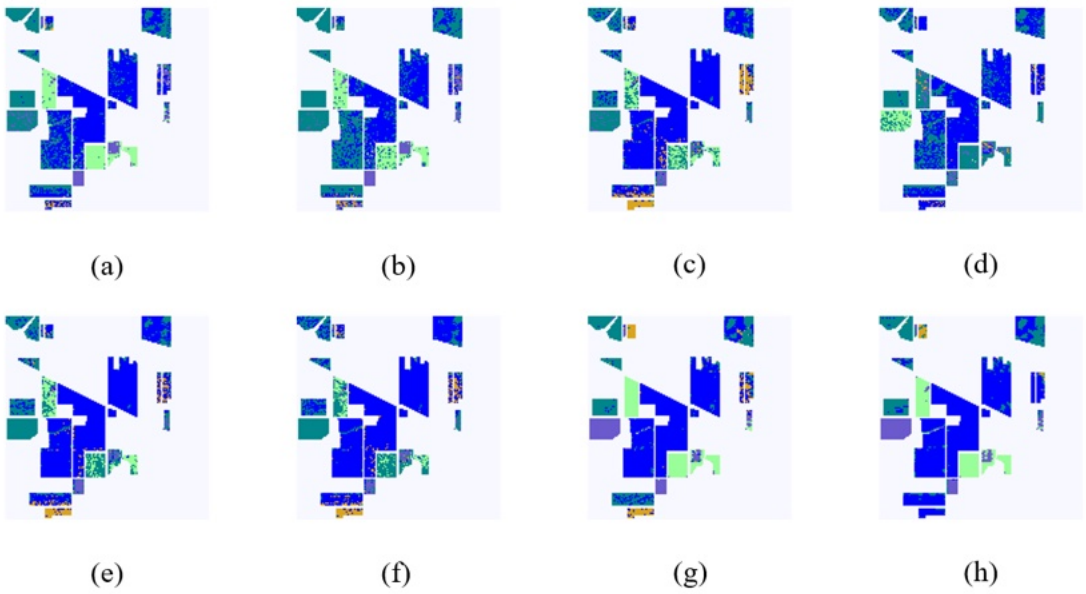


Fig. 5. Classification of Indian Pines. (a) LiCa. (b) GLS [48]. (c) MMD. (d) CAD [52]. (e) VREx [54]. (f) ERM. (g) RF [51]. (h) SVM [31].

3) Indian Pines: As illustrated in Table V, our LiCa method achieves the best performance under indicator C-2 accuracy, C-4 accuracy, OA, and AA, indicating the outstanding of our method. Experiment results show that all the domain generalization strategies perform better than ERM while all the methods perform not as well as in the Pavia City dataset. We analyze that the phenomenon of hyperspectral heterospectra is significant among different regions of Indian Pines, and therefore these algorithms that extract potential domain-invariant features become quite competitive against the other methods. However, the phenomenon of hyperspectral heterospectra is too apparent that leads to a serious decay of accuracy performance on the target domain. Moreover, our proposed LiCa still makes an achievement even for such an adverse HSI dataset.

In the three datasets discussed above, the performance of CausIRL [53], CAD [54], and VREx [52], which have achieved state-of-the-art performance in normal imagery classification tasks, is not exceptional for the Eurocrops dataset. We attribute this to the direct application of these methods without taking into account the unique characteristics of HSI, leading to a lack of performance. At the same time, we found that CondCAD, as a covariate shift version of CAD, performed worse than CAD. On the other hand, our proposed method, which combines the conditional alignment domain generalization strategy with pixel-wise HSI classification, achieves outstanding results, providing mutual verification with our proposed upper bound theorems and demonstrating the effectiveness of our approach.

C. Ablation Experiment

As shown in Tables IV and VI, MMD denotes the nonconditional alignment, GLS denotes the method with L_{CdCA}, CdE denotes the method with L_{CdE}, and LiCa denotes the proposed conditional alignment with losses L_{CdCA} and L_{CdE}. Our proposed LiCa block has an advantage compared with nonconditional alignment MMD and the intuitive conditional alignment GLS, which is in a mutual confirmation of the proposed theorem.

Though the LiCa method performs better than the rectified GLS method without L_{CdA} on these three datasets, the degree of such advantage is various. In the Pavia City dataset and the Indian Pines dataset, the advantage of the rectification is merely at a level of 0.5% and 0.2% against GLS while in the Eurocrops dataset, however, such advantage is at a level of 2.5%. We analyze that it is because the more serious problem is class imbalance. As denoted in Table II, the number of samples in C-1 (class with the largest sample size) is nearly ten times than all the other classes for the Eurocrops dataset, while for the Pavia City dataset and Indian Pines dataset, no class reaches a dominating rank like that.

In addition to the previous discussion, Table VII indicates the phenomenon of hyperspectral heterospectra in Pavia, Eurocrops, and Indian Pines datasets. In this experiment, all the results are obtained through the ERM method and test ratio is the ratio of the size of the test set to the entire set. In Eurocrops and Indian Pines datasets, the accuracies on validation sets in source domains are nearly 30% and 60% greater than in target domains while such divergence is still at a level of 0.51% in the Pavia City dataset. Apparent classification outfitting on test sets in source domains denotes that the aforementioned phenomenon of hyperspectral heterospectra is a formidable obstacle for the HSI classification task.

V. Conclusion

In this research, we introduce a novel deep neural network architecture designed for pixel-wise classification of HSI. Our approach incorporates a domain generalization strategy block

and a LiCa block, enabling effective alignment of conditional distributions of spectra using a label prior matrix. By integrating the LiCa block into a domain-generalization-based classification network, our method offers a promising solution to address the hyperspectral-heterospectra issue encountered in HSI classification tasks. Notably, our approach focuses exclusively on spectral features, reducing the risk of falsely high accuracy.

Additionally, our proposed algorithm introduces two loss functions specifically tailored for common conditional alignment strategies, supported by a solid theoretical foundation. Through comprehensive experiments conducted on three distinct HSI datasets, we demonstrate the superiority of our approach over classical pixel-wise algorithms and architectures based on nonalignment-based SOTA domain generalization strategies.

REFERENCES

[1] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7920–7930, Nov. 2020.

[2] Y. Zhou, W. Li, P. Ren, Z. Li, and R. Tao, "Sea-ice classification based on optical image using morphological profile features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 3055–3058.

[3] X. Kang, Z. Wang, P. Duan, and X. Wei, "The potential of hyperspectral image classification for oil spill mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5538415.

[4] X. Zheng, H. Cui, C. Xu, and X. Lu, "Dual teacher: A semisupervised cotraining framework for cross-domain ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613312.

[5] W. Lv and X. Wang, "Overview of hyperspectral image classification," *J. Sensors*, vol. 2020, pp. 1–13, Jul. 2020.

[6] B. Xie, Y. Zhang, S. Mei, G. Zhang, Y. Feng, and Q. Du, "Spectral variation augmented representation for hyperspectral imagery classification with few labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543212.

[7] G. Liu, Y. Yuan, Y. Zhang, Y. Dong, and X. Li, "Style transformation-based spatial–spectral feature learning for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401515.

[8] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.

[9] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[10] L. Fang, Z. Liu, and W. Song, "Deep hashing neural networks for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1412–1416, Sep. 2019.

[11] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2647–2650.

[12] L. Fang, G. Liu, S. Li, P. Ghamisi, and J. A. Benediktsson, "Hyperspectral image classification with squeeze multibias network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1291–1301, Mar. 2019.

[13] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1968–1972, Nov. 2020.

[14] B. Cui, X.-M. Dong, Q. Zhan, J. Peng, and W. Sun, "LiteDepthwiseNet: A lightweight network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502915.

[15] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[16] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[17] S. Paul and D. Nagesh Kumar, "Spectral–spatial classification of hyperspectral data with mutual information based segmented stacked autoencoder approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 265–280, Apr. 2018.

[18] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[19] A. Sellami and I. R. Farah, "Spectra-spatial graph-based deep restricted Boltzmann networks for hyperspectral image classification," in *Proc. Photon. Electromagn. Res. Symp.-Spring (PIERS-Spring)*, Jun. 2019, pp. 1055–1062.

[20] H. Wu 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, Mar. 2018.

[21] M. Seydgar, A. A. Naeini, M. Zhang, W. Li, and M. Satari, "3-D convolution-recurrent networks for spectral–spatial classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 7, p. 883, Apr. 2019.

[22] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Scalable recurrent neural network for hyperspectral image classification," *J. Supercomput.*, vol. 76, no. 11, pp. 8866–8882, Nov. 2020.

[23] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[24] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[25] X. Kang, S. Li, L. Fang, and J. A. Benediktsson, "Intrinsic image decomposition for feature extraction of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2241–2253, Apr. 2015.

[26] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou, and L. Guo, "KNN-based representation of superpixels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4032–4047, Nov. 2018.

[27] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5028–5039, Jun. 2021.

[28] G. Licciardi and J. Chanussot, "Spectral transformation based on nonlinear principal component analysis for dimensionality reduction of hyperspectral images," *Eur. J. Remote Sens.*, vol. 51, no. 1, pp. 375–390, Jan. 2018.

[29] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.

[30] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[31] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[32] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8827–8844, Jul. 2023.

[33] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jan. 2015.

[34] B. Kang, I. Park, C. Ok, and S. Kim, "ODPA-CNN: One dimensional parallel atrous convolution neural network for band-selective hyperspectral image classification," *Appl. Sci.*, vol. 12, no. 1, p. 174, Dec. 2021.

[35] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, "Domain adaptation in remote sensing image classification: A survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9842–9859, 2022.

[36] S. Zhu, B. Du, L. Zhang, and X. Li, "Attention-based multiscale residual adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5400715.

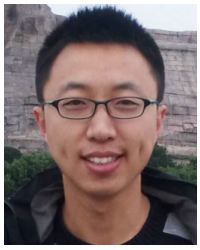
[37] J. Geng, X. Deng, X. Ma, and W. Jiang, "Transfer learning for SAR image classification via deep joint distribution adaptation networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5377–5392, Aug. 2020.

[38] 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, May 2021.

[39] S. Zhang, Z. Chen, D. Wang, and Z. J. Wang, "Cross-domain few-shot contrastive learning for hyperspectral images classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[40] J. Yang, Y.-Q. Zhao, and J. C. Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
[41] M. Chen, L. Ma, W. Wang, and Q. Du, "Augmented associative learning-based domain adaptation for classification of hyperspectral remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6236–6248, 2020.
[42] B. Lucas, C. Pelletier, D. Schmidt, G. I. Webb, and F. Petitjean, "A Bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping," *Mach. Learn.*, vol. 112, pp. 1941–1973, Mar. 2021.
[43] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.
[44] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.
[45] R. T. des Combes, H. Zhao, Y.-X. Wang, and G. J. Gordon, "Domain adaptation with conditional distribution matching and generalized label shift," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19276–19289.
[46] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 819–827.
[47] Y. Li et al., "Deep domain generalization via conditional invariant adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 624–639.
[48] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao, "Domain generalization via conditional invariant representations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–9.
[49] M. Schneider, C. Marchington, and M. Körner, "Challenges and opportunities of large transnational datasets: A case study on European administrative crop data," 2022, *arXiv:2210.07178*.
[50] X. Huang and L. Zhang, "A comparative study of spatial approaches for urban mapping using hyperspectral ROSIS images over Pavia city, Northern Italy," *Int. J. Remote Sens.*, vol. 30, no. 12, pp. 3205–3221, Jun. 2009.
[51] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2018.
[52] D. Krueger et al., "Out-of-distribution generalization via risk extrapolation (REx)," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5815–5826.
[53] M. Chevalley, C. Bunne, A. Krause, and S. Bauer, "Invariant causal mechanisms through distribution matching," 2022, *arXiv:2206.11646*.
[54] Y. Ruan, Y. Dubois, and C. J. Maddison, "Optimal representations for covariate shift," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–39.
[55] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," 2020, *arXiv:2007.01434*.

Zhe Gao received the B.S. degree in mathematical science from the School of Mathematical Sciences, Nankai University, Tianjin, China, in 2022, where he is currently pursuing the M.S. degree with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC.

His research interests include transfer learning, deep learning, and hyperspectral imagery classification.



Bin Pan (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2013 and 2019, respectively.

Since 2019, he has been an Associate Professor with the School of Statistics and Data Science, Nankai University, Tianjin, China. His research interests include hyperspectral image classification, unmixing and super-resolution, domain adaptation, and generalization for remote sensing.

Xia Xu received the B.S. and M.S. degrees from the School of Electrical Engineering, Yanshan University, Qinhuangdao, China, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019.

She is currently an Assistant Professor with the College of Computer Science, Nankai University, Tianjin, China. Her research interests include hyperspectral unmixing, multiobjective optimization, and remote-sensing image processing.

Tao Li received the Ph.D. degree in computer science from Nankai University, Tianjin, China, in 2007.

He is currently with the College of Computer Science, Nankai University, as a Professor. His main research interests include heterogeneous computing, machine learning, and the Internet of Things.

Dr. Li is a Distinguished Member of the China Computer Federation (CCF).

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 200 scientific articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). His research interests include remote-sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the *Pattern Recognition*, the *ISPRS Journal of Photogrammetry and Remote Sensing*, and the *Infrared Physics and Technology*. His personal website is http://levir.buaa.edu.cn/.