

# Dual-Intervention-Constrained Mask-Adversary Framework for Unsupervised Domain Adaptation of Hyperspectral Image Classification

Chunyan Yu<sup>ID</sup>, Senior Member, IEEE, Mingyang Xu, Qiang Zhang<sup>ID</sup>, Member, IEEE,  
and Xiaoqiang Lu<sup>ID</sup>, Senior Member, IEEE

**Abstract**—To mitigate the domain shift and enhance the alignment of the spatial–spectral features, this letter proposes a novel dual-intervention-constrained mask-adversary (DICMA) framework for unsupervised domain adaptation (UDA) of hyperspectral image classification (HSIC). Innovatively, DICMA integrates a generator, masker, and bi-classifier within an adversarial framework constrained by a dual intervention mechanism. Specifically, the correlation intervention module (CIM) ensures the preservation and independence of causal spatial–spectral variables, while the knowledge distillation intervention module completes the spatial–spectral generalization with constrained distillation information. Besides, with the collaborative adversarial training strategy, the proposed approach transfers effective knowledge for spatial–spectral feature alignment. Experimental results and analyses demonstrate the effectiveness of the proposed DICMA model, which yields an accuracy of 91.15% in the Pavia University (PaviaU) → Pavia Center (PaviaC). Our code will be released at <https://github.com/Chirsycy/DICMA>.

**Index Terms**—Adversarial training, hyperspectral image classification (HSIC), unsupervised domain adaptation (UDA).

## I. INTRODUCTION

HYPERSPECTRAL image classification (HSIC) assigns class labels to individual pixels and plays a crucial role in various fields, such as land cover mapping and environmental monitoring. With the development of deep learning models, remote sensing processing approaches have made remarkable advancements in recent years [1], [2], [3], [4], [5], [6], which rely on a substantial amount of labeled data for model training.

In recent years, the lack of labeled samples poses a challenge for the HSIC approach. Unsupervised domain adaptation (UDA) supplies an effective approach to solve the mentioned difficulties, which utilizes knowledge from a labeled source domain to improve performance on an unlabeled target domain. For UDA of HSIC, the approaches aim to minimize distribution discrepancies of different HSI domains [7]. Typically, generative adversarial networks (GANs) have provided

Manuscript received 31 May 2024; revised 5 July 2024; accepted 27 July 2024. Date of publication 30 August 2024; date of current version 9 September 2024. This work was supported by the National Nature Science Foundation of China under Grant 62471079 and Grant 62401095. (Corresponding author: Qiang Zhang.)

Chunyan Yu, Mingyang Xu, and Qiang Zhang are with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: yucy@dlmu.edu.cn; xumingyang1108@163.com; qzhang95@dlmu.edu.cn).

Xiaoqiang Lu is with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China.

Digital Object Identifier 10.1109/LGRS.2024.3451180

a popular pattern to implement feature alignment for domain adaptation in UDA [8], which implements feature alignment between source and target domains through an iterative adversarial training process. The UDA for HSIC encounters domain shifts caused by spectral variation and requires enhancement in feature alignment. TAADA [9] adopts a two-branch attention adversarial network, in which the generator extracts the attention-based spectral–spatial features for domain alignment. ADA-Net [10] leverages variational autoencoder and dual classifiers to optimize local class differences with different HSI domains. CLDA [11] employs confidence learning for HSIC to enhance classification accuracy. UDACA [12] builds an adversarial UDA with a presented contentwise alignment mechanism for HSIC. Although the existing methods have achieved significant progress, they primarily focus on adversarial framework construction, neglecting the potential benefits of leveraging causal relationships in improving feature alignment and generalization.

In this letter, we develop a dual-intervention-constrained mask-adversary UDA framework for HSIC denoted as DICMA. The presented model contains a generator, masker, and biclassifier within an adversarial framework. Besides, the dual-intervention mechanism is composed of the correlation intervention module (CIM) and distillation intervention module (DIM), which facilitates sufficient causal and distillation information to enhance the alignment of the spatial–spectral features for HSIC. The main contributions of this letter are as follows.

- 1) Unlike the generator–discriminator structure in the traditional GAN-based UDA method, we present a novel mask-adversary UDA framework for HSIC, which is structurally integrated with a masker into the adversarial framework for the first time. Besides, with collaborative adversarial training of the masker, biclassifier, and generator, the proposed network enhances the robustness of the spatial–spectral feature alignment between different HSIs.
- 2) Unlike the traditional way of the spatial–spectral feature alignment, we develop a dual intervention schema in the mask-adversary framework to constrain the UDA for HSIC. Specifically, CIM focuses on preserving and maintaining the independence of causal spatial–spectral variables, and DIM complements the alignment of spatial–spectral distributions by incorporating constrained distillation information.

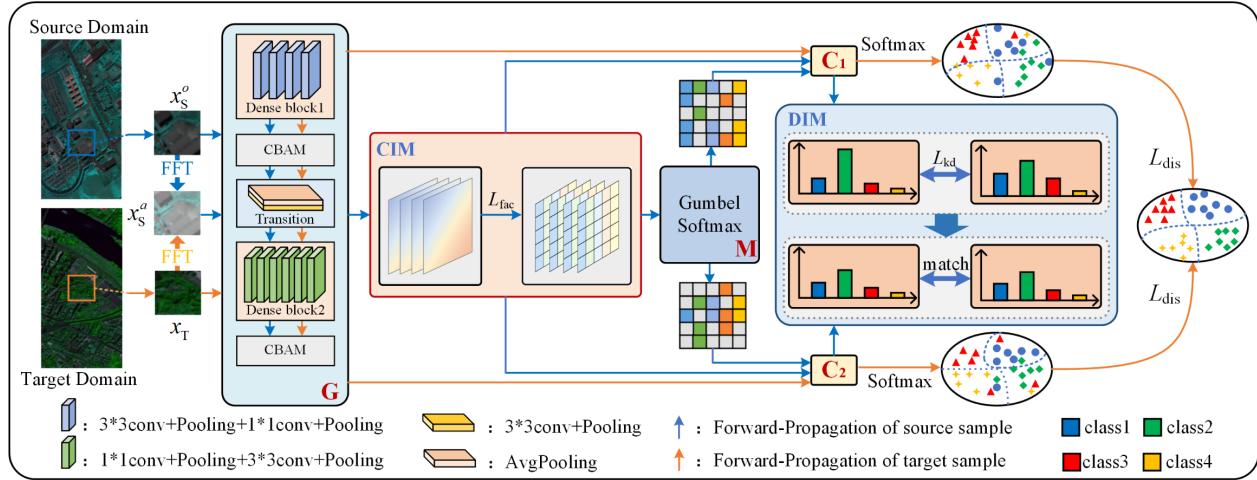


Fig. 1. Architecture of the proposed DICMA. Structurally, DICMA integrates a generator ( $G$ ), masker ( $M$ ), and biclassifier ( $C_1$  and  $C_2$ ) within an adversarial framework constrained by CIM and DIM. CIM enhances the spatial–spectral alignment with the preservation and independence of causal variables, and DIM decreases domain shift with constrained distillation information.

## II. PROPOSED APPROACH

Fig. 1 illustrates the adversarial architecture of DICMA. Specifically,  $G$  is responsible for the discriminative feature extraction with the original and the augmented samples processed by fast Fourier transform (FFT). CIM aims to decompose causal variables from mixed status to independent status.  $M$  separates the independent causal variables by a masker and imports them to  $C_1$  and  $C_2$  individually, which intends to obtain sufficient causal spatial–spectral features through the adversarial learning phase. Besides, DIM further constrains the biclassifiers with distillation information of the same class to improve the spatial–spectral feature adaptability. The detailed descriptions of components are provided as follows.

### A. Problem Definition

For two HSI datasets, assume the sample set of the source domain denoted as  $\{X_S^o, Y_S\} = \{(x_{S1}^o, y_{S1}), \dots, (x_{Sn}^o, y_{Sn})\}$ , where  $x_{Sn}^o$  denotes the  $n$ th sample and  $y_{Sn}$  is the related label. Similarly, the sample set of the target domain is denoted as  $\{X_T\} = \{x_{T1}, \dots, x_{Tm}\}$ , where  $x_{Tm}$  represents the  $m$ th sample. To mine the causal variables of  $\{X_S^o, Y_S\}$ , we employ random FFT [13], where the phase component is derived from  $x_{Sn}^o$  and the amplitude component is sourced from  $x_{Tm}$ , resulting in an augmented FFT-based sample set of the source domain denoted as  $\{X_S^a, Y_S\} = \{(x_{S1}^a, y_{S1}), \dots, (x_{Sn}^a, y_{Sn})\}$ . Notably, augmented samples inject causal information into the samples that are beneficial for the subsequent training. Our work intends to minimize the discrepancy between the two HSIs with the samples of  $\{\{X_S^o, Y_S\}, \{X_S^a, Y_S\}\}$ , and  $\{X_T\}$  to accomplish UDA for HSIC.

### B. Dual Intervention for Feature Alignment

In MADI-UDA, both original and augmented source HSI samples are fed into the dual-intervention module, which incorporates CIM and DIM to decrease spatial–spectral alignment bias.

1) **CIM:** Since the high correlation between causal variables of the same dimension increases the difficulty of the feature alignment, to eliminate the interference from spectral

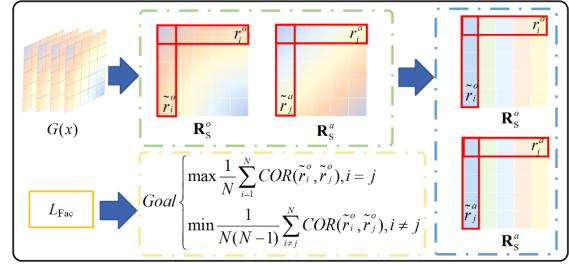


Fig. 2. Detail of CIM.  $R_S^o$  and  $R_S^a$  are the embeddings of  $G(x_S^o)$  and  $G(x_S^a)$ . Each square in  $r_S^o$  and  $r_S^a$  is considered a dimension, as shown in the green dotted box, after being constrained by  $L_{\text{fac}}$  with COR denoted as correlation calculation, and the dimensions are independent as shown in the blue dotted box.

variability, the UDA model for HSIC should guarantee that causal variables are independent. In our model, we design CIM to achieve the separation of causal spatial–spectral features from the others. Specifically, to guarantee each dimension of causal features contains unique information separate from the other dimensions, we employ the FAC loss [14] that minimizes the off-diagonal elements of the correlation matrix in CIM. As a result, CIM ensures that each dimension of the causal spatial–spectral features achieves independence from the others, which is illustrated in Fig. 2. The FAC loss is defined in the following equation:

$$L_{\text{fac}} = \frac{1}{2} \|\mathbf{C} - \mathbf{I}\|_F^2, \quad \mathbf{C}_{ij} = \frac{\langle \tilde{r}_i^o, \tilde{r}_j^a \rangle}{\|\tilde{r}_i^o\| \cdot \|\tilde{r}_j^a\|}, \quad i, j \in 1, 2, \dots, N \quad (1)$$

where  $\mathbf{I}$  denotes the unit vector, and  $\tilde{r}_i^o$  and  $\tilde{r}_j^a$  represent the Z-score normalization of the  $i$ th and  $j$ th column of  $R_S^o$  and  $R_S^a$  from  $G$ , respectively.

2) **DIM:** In the discriminator part, to effectively leverage the discriminative knowledge acquired from the HSI source domain, we design DIM to ensure feature representation consistency and stable classification predictions with dual classifiers. Specifically, DIM employs distillation to align the distributions of original and augmented samples to reduce prediction discrepancies. After performing the classification of the samples in the source domain, soft labels are generated

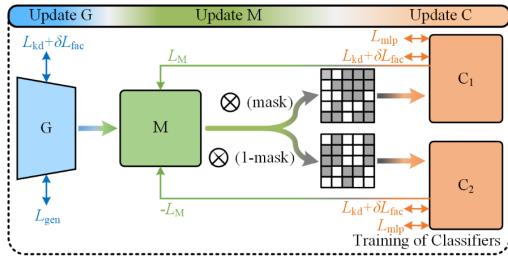


Fig. 3. Training steps of mask-adversary framework. The collaboration of adversarial learning accomplishes effective knowledge transferring through updates  $G$ ,  $M$ , and  $C$ , respectively.

with T-Softmax defined as follows:

$$P(y|x_S; \theta, T) = \frac{\exp(f_y(x_S; \theta)/T)}{\sum_{k=1}^K \exp(f_k(x_S; \theta)/T)} \quad (2)$$

where  $f_y$  is the real probability,  $f_k$  represents the predicted probability of the  $k$ th class,  $\theta$  denotes the parameters of the input network,  $T$  is the temperature hyperparameter, and  $K$  is the number of categories.

In the DIM, we employ the cross-entropy loss  $L_E(x_S^a, y_S; \theta_G, \theta_{C_i})$  and the Kullback–Leibler (KL) divergence to align the predictions of the source training samples. The loss in DIM denoted as  $L_{\text{kd}}$  is defined as follows:

$$\begin{aligned} L_i^{\text{cls}}(x_S^o, x_S^a; \theta_G, \theta_{C_i}, T) \\ = \text{KLF}(P(y|x_S^a; \theta_G, \theta_{C_i}, T) || P(y|x_S^o; \theta_G, \theta_{C_i}, T)) \end{aligned} \quad (3)$$

$$L_{\text{kd}} = \sum_i^2 (\alpha L_E(x_S^a, y_S; \theta_G, \theta_{C_i}) + (1 - \alpha) \cdot T^2 \cdot L_i^{\text{cls}}) \quad (4)$$

where  $\text{KLF}(\cdot, \cdot)$  represents KL divergence and  $i \in \{1, 2\}$  to measure the difference between the outputs of  $P(y|x_S^a; \theta_G, \theta_{C_i}, T)$  and  $P(y|x_S^o; \theta_G, \theta_{C_i}, T)$ .

### C. Mask-Adversary UDA Framework

As illustrated in Fig. 3, the mask-adversary framework consists of three adversarial training steps.

1) *Masker Training*: We facilitate the traditional adversarial framework by incorporating a masker module, which aims to supply sufficient causal features in the process of feature alignment. The principle behind this approach is that the separated causal spatial–spectral features encompass varying degrees of causality. Sufficient features for classification are obtained gradually through the masker and biclassifier learning in an adversarial way. Specifically, the mask is defined as (5) based on the Gumbel-Softmax function with  $\lambda \in U(0, 1)$

$$\text{mask} = \text{Gumbel-Softmax}(r, \lambda N). \quad (5)$$

With the  $G$ ,  $C_1$ , and  $C_2$  fixed, the loss of training the masker  $L_M$  is defined as

$$\begin{aligned} L_{\text{sup}} &= L_E(C_1((r^o \otimes \text{mask}^o) + (r^a \otimes \text{mask}^a)), y_S; \theta_m) \\ L_{\text{inf}} &= L_E(C_2((r^o \otimes (1 - \text{mask}^o)) \\ &\quad + (r^a \otimes (1 - \text{mask}^a))), y_S; \theta_m) \\ L_M &= L_{\text{sup}} - L_{\text{inf}} \end{aligned} \quad (6)$$

where  $r \otimes \text{mask}$  and  $r \otimes (1 - \text{mask})$  represent the two separate spatial–spectral causal features. In this manner,  $M$  effectively distinguishes the variables with sufficient expression or not.

### Algorithm 1 DICMA for HSIC

**Input:**  $\{X_S^o, Y_S\}$ ,  $\{X_T\}$ ,  $x_S^a \leftarrow \text{FFT}(x_S^o, x_T)$ , iteration number  $E$ , temperature parameters  $T$ , loss weight  $\delta$ .

Randomly initialized  $\theta_{C_1}$ ,  $\theta_{C_2}$ ,  $\theta_G$ ,  $\theta_M$ .

**Output:**  $\theta_{C_1}$ ,  $\theta_{C_2}$ ,  $\theta_G$ .

**For**  $i = 1$  **to**  $E$  **do**:

Step 1: Fixed  $\theta_M$  and optimize  $\theta_{C_1}$ ,  $\theta_{C_2}$ ,  $\theta_G$  with  $x_S^o$  and  $x_S^a$  according to (1) and (4) by:

$$\theta_G, \theta_{C_1}, \theta_{C_2} \leftarrow -\nabla_{\theta_G, \theta_{C_1}, \theta_{C_2}}(L_{\text{kd}} + \delta L_{\text{fac}});$$

Step 2: Fixed  $\theta_{C_1}$ ,  $\theta_{C_2}$ ,  $\theta_G$  and optimize  $\theta_M$  with  $x_S^o$  and  $x_S^a$  according to (6) by:

$$\theta_M \leftarrow -\nabla_{\theta_M}(L_M);$$

Step 3: Fixed  $\theta_M$ ,  $\theta_G$  and optimize  $\theta_{C_1}$ ,  $\theta_{C_2}$  with  $x_S^o$ ,  $x_S^a$  and  $x_T$  according to (9) by:

$$\theta_{C_1}, \theta_{C_2} \leftarrow -\nabla_{\theta_{C_1}, \theta_{C_2}}(L_{\text{mlp}});$$

Step 4: Fixed  $\theta_M$ ,  $\theta_{C_1}$ ,  $\theta_{C_2}$  and optimize  $\theta_G$  with  $x_S^o$ ,  $x_S^a$  and  $x_T$  according to (10) by:

$$\theta_G \leftarrow -\nabla_{\theta_G}(L_{\text{gen}});$$

**End.**

2) *Dual Classifier Training*: The biclassifiers that consist of  $C_1$  and  $C_2$  are designed with identical structures and randomly initialized with the same parameters. With the  $G$  and masker fixed, the loss is defined as

$$L_C = L_{\text{sup}} + L_{\text{inf}}. \quad (7)$$

Furthermore, dual classifiers align the source and target HSI domains by handling the prediction discrepancy, which is achieved by loss defined as (8) that quantifies the difference in predictions by  $C_1$  and  $C_2$  on the target domain

$$\begin{aligned} L_{\text{dis}}(x_T; \theta_G, \theta_{C_1}, \theta_{C_2}) = \frac{1}{K} \sum_{k=1}^K & \left| p_k^{C_1}(y|x_T; \theta_G, \theta_{C_1}) \right. \\ & \left. - p_k^{C_2}(y|x_T; \theta_G, \theta_{C_2}) \right|. \end{aligned} \quad (8)$$

Maximizing the prediction discrepancy is beneficial in detecting target domain samples that deviate significantly from the source domain. Thus, with the  $G$  and masker fixed, the loss  $L_{\text{mlp}}$  is defined as

$$L_{\text{mlp}} = L_C - L_{\text{dis}}. \quad (9)$$

3) *Generator Training*: Moreover, we fix the masker and biclassifier and train  $G$  to decrease the prediction discrepancy between the two classifiers, which promotes  $G$  to generate target features that are closer to the decision boundary of the biclassifier. The training loss  $L_{\text{gen}}$  is defined as

$$L_{\text{gen}} = L_{\text{dis}}. \quad (10)$$

### D. Algorithm of the Proposed Method

Briefly, the algorithm of DICMA is outlined as follows.

TABLE I  
STRUCTURE OF SPECIFIC NETWORK

Component	Structure
G	Convolutional, DenseBlock, Transition, CBAM [18]
M	Linear, BatchNorm, ReLU, Gumbel Softmax
C1 and C2	Linear, LeakyReLU

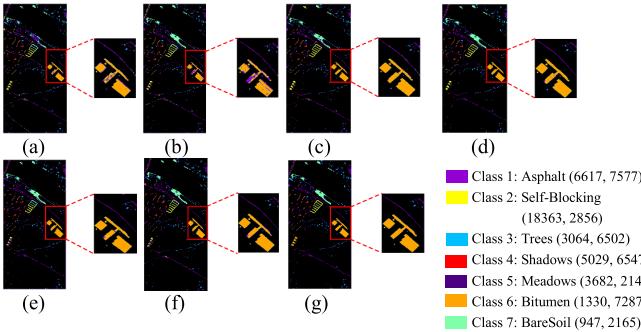


Fig. 4. Classification results with compared methods on the PaviaC. (a) DANN. (b) MCD. (c) CLDA. (d) UDACA. (e) SCLUDA. (f) DICMA. (g) Ground truth. The number after the category in the legend represents the number of PaviaU and PaviaC categories.

### III. EXPERIMENT AND RESULT ANALYSIS

#### A. Data Description

The first HSIC cross-scene includes the Pavia University (PaviaU) and Pavia Center (PaviaC) datasets. After removing the noise band, PaviaU has a resolution of  $610 \times 340$  with 103 bands, and PaviaC consists of  $1096 \times 492$  pixels and 102 bands.

The second HSIC cross-scene comprises the Houston 2013 (Hou13) and Houston 2018 (Hou18) datasets. The Hou13 has  $349 \times 1905$  pixels with 144 spectral bands, while the Hou18 has 48 bands and a resolution of  $209 \times 955$ . For the two scenes, seven classes were chosen for experimental analysis.

#### B. Experimental Settings

The experiments were conducted on Windows 10 equipped with an NVIDIA GeForce GTX 1650 with 4-GB GDDR6 memory. The experiments were performed with PyTorch 3.9 with Cuda11.3-cudnn8. Five UDA methods, including DANN [15], MCD [16], CLDA [11], UDACA [12], and SCLUDA [17], were employed for comparison. The sample size is  $11 \times 11$ . Besides, we set the batch size to 16 and the learning rate to 0.0001,  $E = 300$ ,  $T = 4$ , and  $\delta = 5$ . The evaluation criteria included overall accuracy (OA), average accuracy (AA), and kappa. The specific network structure of the  $G$ ,  $M$ ,  $C1$ , and  $C2$  is reported in Table I.

#### C. Results and Analysis

In the first experiment, the labeled PaviaU is utilized as the source domain, while the unlabeled PaviaC is adopted as the target domain. Table II and Fig. 4 present the compared classification results and maps. As observed, DICMA outperforms other methods on the PaviaC dataset, achieving the best OA, AA, and kappa of 91.15%, 91.22%, and 0.89, respectively. Besides, our method achieves satisfactory accuracies in most categories, particularly in classes 2, 4, and 7. In contrast, DANN, MCD, and SCLUDA have lower OA than DICMA,

TABLE II  
CLASSIFICATION PERFORMANCE FROM THE RESULTS WITH ALL THE COMPARED METHODS OF THE PAVIA C (PAVIAU → PAVIAC)

Class	DANN	MCD	CLDA	UDACA	SCLUDA	Ours
1	63.65	91.95	92.59	<b>96.39</b>	95.03	88.71
2	75.56	84.41	91.42	76.45	79.45	<b>96.26</b>
3	92.28	72.74	81.05	<b>93.31</b>	92.25	92.60
4	70.25	66.14	84.90	87.57	77.10	<b>90.48</b>
5	86.71	79.77	<b>96.25</b>	90.05	90.65	85.34
6	81.08	79.28	85.05	<b>86.10</b>	82.77	85.20
7	99.35	98.20	99.79	99.85	99.88	<b>99.93</b>
OA (%)	78.39	79.97	88.02	90.24	89.09	<b>91.15</b>
AA (%)	81.27	81.78	90.15	89.96	88.16	<b>91.22</b>
Kappa	0.74	0.76	0.85	0.88	0.87	<b>0.89</b>
Time(s)	901	1420	1627	1043	1152	2186

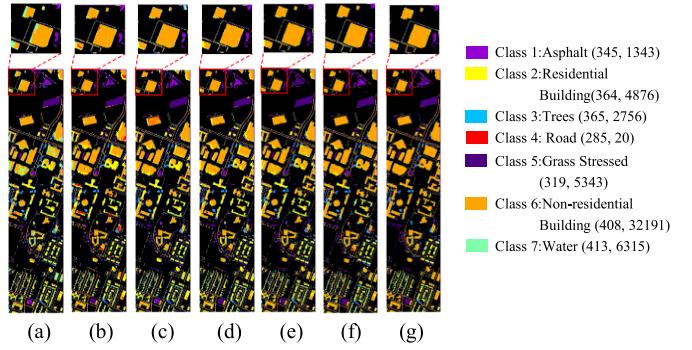


Fig. 5. Classification results with compared methods on Hou18. (a) DANN. (b) MCD. (c) CLDA. (d) UDACA. (e) SCLUDA. (f) DICMA. (g) Ground truth. The number after the category in the legend represents the number of Hou13 and Hou18 categories.

TABLE III

CLASSIFICATION PERFORMANCE FROM THE RESULTS WITH ALL THE COMPARED METHODS OF THE HOU18 (HOU13 → HOU18)

Class	DANN	MCD	CLDA	UDACA	SCLUDA	Ours
1	61.38	73.66	30.25	<b>73.05</b>	62.55	67.79
2	67.56	65.94	70.85	66.42	66.54	<b>73.09</b>
3	66.04	74.34	51.50	<b>74.17</b>	66.71	68.20
4	83.33	88.33	56.06	89.82	<b>90.00</b>	83.00
5	87.85	75.06	<b>96.18</b>	83.59	85.34	82.13
6	62.82	58.23	63.50	62.49	73.92	<b>78.10</b>
7	57.94	63.26	<b>74.85</b>	65.77	61.07	57.96
OA (%)	65.34	62.49	67.34	66.27	72.20	<b>74.67</b>
AA (%)	69.56	71.26	63.32	<b>73.62</b>	72.30	72.90
Kappa	0.50	0.48	0.53	0.52	0.58	<b>0.60</b>
Time(s)	783	1242	1536	995	930	1920

which indicates that the combination of dual-intervention constraints and adversarial learning in DICMA is more effective than mere adversarial approaches for UDA. The results of CLDA and UDACA show the benefits of dual-classifier and adversarial learning.

In the second experiment, we employ the labeled Hou13 as the source domain, while the unlabeled Hou18 serves as the target domain. Table III demonstrates the classification results comparison, and Fig. 5 illustrates the map results. It can be seen that DICMA achieves an OA of 74.67%, an AA of 72.90%, and a kappa of 0.60. The OA of UDACA is 66.27%, which indicates that dense networks with adversarial learning improve domain alignment performance. Notably, the performance of all methods on Hou13 → Hou18 is lower compared to the first scene, which indicates that the challenges of this cross-scene classification task are more complex than the Pavia scene.

Additionally, the computational efficiency of our method is lower than other approaches. In summary, with the aid of the

TABLE IV  
ABLATION STUDY

Loss Items	$L_{\text{fac}}$	$L_{\text{kd}}$	$\sqrt{\phantom{x}}$	$\sqrt{\phantom{x}}$
	$L_{\text{fac}}$	$L_{\text{kd}}$		
OA	PaviaU $\rightarrow$ PaviaC		75.97%	80.73%
	Hou13 $\rightarrow$ Hou18		62.49%	65.99%

Fig. 6. T-SNE visualization of feature maps. (a), (b), (e), and (f) represent the initial data distribution of PaviaU, PaviaC, Hou13, and Hou18, respectively. (c), (d), (g), and (h) represent the feature distribution after DICMA of PaviaU, PaviaC, Hou13, and Hou18, respectively.

constrained CIM and DIM, our proposed framework transfers effective knowledge for HSIC. However, the DICMA remains hindered by the problems of severe distribution gaps and high computational demands.

#### D. Ablation Study

To investigate the impact of the dual-intervention mechanism in the DICMA model, ablation experiments are conducted to analyze the roles of  $L_{\text{kd}}$  correlation loss  $L_{\text{fac}}$ . Table IV presents the comparison of the OA of the model under different loss combinations. The approach with both  $L_{\text{kd}}$  and  $L_{\text{fac}}$  yields optimal accuracy, which indicates the collaborative promotion of performance by the two losses. The approach with only  $L_{\text{kd}}$  generates the OA of 80.73% and 65.99% on the cross-scene of PaviaU  $\rightarrow$  PaviaC and Hou13  $\rightarrow$  Hou18, respectively, while the approach with  $L_{\text{fac}}$  improves the accuracy by 12.13% and 7.8% in the two situations. Overall, the ablation study demonstrates the effectiveness of both losses in the CIM and DIM modules.

#### E. Visualization of Feature Adaptation

In this section, we employ the t-distributed stochastic neighbor embedding (t-SNE) algorithm to visualize the intradomain features in the two scenes. The visualization results are shown in Fig. 6. As illustrated, during the initial stage, the samples within each domain exhibit a disordered arrangement. From Fig. 6(c) and (g), it is evident that the proposed model achieves impressive alignment results in the source domain for both scenes. Furthermore, Fig. 6(d) demonstrates that DICMA achieves favorable alignment effects in PaviaC with an OA of 91.15%. Besides, it is inferred from Fig. 6(h) that due to the relative complexity of Hou18, some instances of category mixing were observed even after domain alignment and the OA is 74.67%.

#### IV. CONCLUSION

In this letter, we presented a novel dual-intervention mask-adversary UDA framework that incorporates a generator, masker, and biclassifier for HSIC. The dual-intervention mechanism incorporates causal and distillation information to improve the spatial–spectral feature alignment. Importantly, the CIM accomplishes the independence of causal variables, and the DIM employs a biclassifier and distillation constraint to achieve the distribution alignment for HSIC. Experimental results demonstrate the superiority of the DICMA model. In the future, we plan to develop the few-shot and zero-shot UDA for HSIC with the proposed mask-adversary framework.

#### REFERENCES

- [1] R. Ji, K. Tan, X. Wang, C. Pan, and L. Xin, “PASSNet: A spatial–spectral feature extraction network with patch attention module for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [2] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, “A unified multiscale learning framework for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.
- [3] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, “Fusion of dual spatial information for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7726–7738, Sep. 2021.
- [4] C. Yu et al., “Distillation-constrained prototype representation network for hyperspectral image incremental classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 3359629.
- [5] P. Duan, X. Kang, P. Ghamisi, and S. Li, “Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509711.
- [6] X. Wang et al., “Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103456.
- [7] 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.
- [8] Y. Huang et al., “Two-branch attention adversarial domain adaptation network for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540813.
- [9] Y. Zhan, D. Hu, Y. Wang, and X. Yu, “Semisupervised hyperspectral image classification based on generative adversarial networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.
- [10] 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.
- [11] Z. Fang et al., “Confident learning-based domain adaptation for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5527116.
- [12] C. Yu, C. Liu, M. Song, and C.-I. Chang, “Unsupervised domain adaptation with content-wise alignment for hyperspectral imagery classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [13] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, “A Fourier-based framework for domain generalization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 14378–14387.
- [14] F. Lv et al., “Causality inspired representation learning for domain generalization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8036–8046.
- [15] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 2030–2096, 2016.
- [16] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3723–3732.
- [17] Z. Li et al., “Supervised contrastive learning-based unsupervised domain adaptation for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524017.
- [18] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis. (Lecture Notes in Computer Science)*, Munich, Germany: Springer, 2018, pp. 3–19.