BAN: A Universal Paradigm for Cross-Scene Classification Under Noisy Annotations From RGB and Hyperspectral Remote Sensing Images

Wentang Chen, Yibin Wen, Juepeng Zheng, *Member, IEEE*, Jianxi Huang, *Senior Member, IEEE*, and Haohuan Fu, *Senior Member, IEEE*

***Abstract*— While domain adaptation (DA) methods have made significant strides in remote sensing community, most current works assume that the source domain labels are accurate. However, limited emphasis has been placed on the scenario where source data are mislabeled with noisy annotations, which is more common in real applications and referred to as noisy DA (NDA). This article formulates remote sensing cross-scene classification on NDA scenarios and proposes a novel network called bilateral adaptation network (BAN), which consists of two parts: 1) forward learning (FL), which utilizes a model learning from the noisy source domain and transfers knowledge to target domain; and 2) backward learning (BL), which utilizes a dual model to acquire knowledge from the target domain and transfer it to source domain. We conduct two parts alternately and adopt a symmetrical Kullback–Leibler (KL) loss to align predictions of the model and its dual model in the same domain. This interactive strategy is able to explore bilateral relationships between domains, implicitly reducing label noise in the source domain. In addition, BAN could serve as a universal paradigm to not only improve the existing NDA methods but also enhance recent DA approaches. Comprehensive evaluations on three publicly available RGB-band remote sensing datasets and two hyperspectral datasets validate the superior effectiveness of our proposed BAN. BAN improves the average accuracy by 6.70%–15.70% on RGB datasets and overall accuracy (OA) by 1.36%–3.14% on hyperspectral datasets with flip-20% noise compared to other state-of-the-art DA and NDA approaches. Promising results indicate the potential of our approach in tackling more general and practical problems with noisy source domain.**

***Index Terms*— Deep learning, noisy domain adaptation (NDA), remote sensing, scene classification.**

I. Introduction

ALTHOUGH deep learning methods have demonstrated remarkable success across a diverse range of tasks within the remote sensing scene classification [1], [2], [3], [4], it demands sufficient annotations from the source data to train models from the scratch. But directly applying models trained on the source domain to target domain, the accuracy tends to deteriorate significantly because of dissimilar data distributions and variations resulting from lighting conditions, viewpoint shifts, and surface conditions. It is frequently observed that remote sensing images, although being categorized within the same class, exhibits significant spectral distinctions when sourced from different datasets. As a result, directly utilizing the model trained in the source domain will lead to a notable decrease in the accuracy of the target domain.

Domain adaptation (DA) approaches are designed to minimize the distribution discrepancy between domains [5]. Most of the existing DA researches assume the presence of clear source domains characterized by precise annotations. However, source domain labels may be corrupted with noise. For example, human annotations could lead to labeling errors due to cognitive differences from experts; Images directly collected from the Internet may be mistakenly labeled. For instance, we usually require transfer knowledge from one remote sensing dataset to another (i.e., NWPU-RESISC45 and UC Merced). As shown in Fig. 1, the samples in the source domain are correctly labeled in the standard DA scenarios, and the model will not acquire incorrect knowledge during the transferring. Nonetheless, in specific real-world scenarios, obtaining large-scale and precisely labeled datasets is often costly and time-consuming, and under certain circumstances, the labels in the source domain are corrupted by noise, and only unlabeled target domain data are available, which is termed noisy DA (NDA). As shown in Fig. 1, the NDA scenario (bottom) presents greater difficulties compared to the standard DA (top) due to incorrect labels (e.g., images of "Bridge" may be incorrectly labeled as "Overpass") in the source domain, which will lead to learning incorrect information and result in negative transfer knowledge.

Received 16 September 2024; revised 3 January 2025; accepted 4 February 2025. Date of publication 6 February 2025; date of current version 21 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant T2125006 and Grant 42401415 and in part by Jiangsu Innovation Capacity Building Program under Project BM2022028. *(Corresponding author: Juepeng Zheng.)*

Wentang Chen is with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China, and also with the State Key Laboratory of Fluid Power and Mechatronic Systems, the Engineering Research Center for Design Engineering and Digital Twin of Zhejiang Province, the School of Mechanical Engineering, and the Robotics Institute, Zhejiang University, Hangzhou 310027, China (e-mail: b241000657@hnu.edu.cn).

Yibin Wen and Juepeng Zheng are with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China (e-mail: wenyb5@mail2.sysu.edu.cn; zhengjp8@mail.sysu.edu.cn).

Jianxi Huang is with the Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 610032, China, also with the College of Land Science and Technology, China Agricultural University, Beijing 100107, China, and also with the Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100125, China (e-mail: jxhuang@cau.edu.cn).

Haohuan Fu is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China, also with the Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100190, China, and also with the National Supercomputer Center in Shenzhen, Shenzhen 518071, China (e-mail: haohuan@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3539499

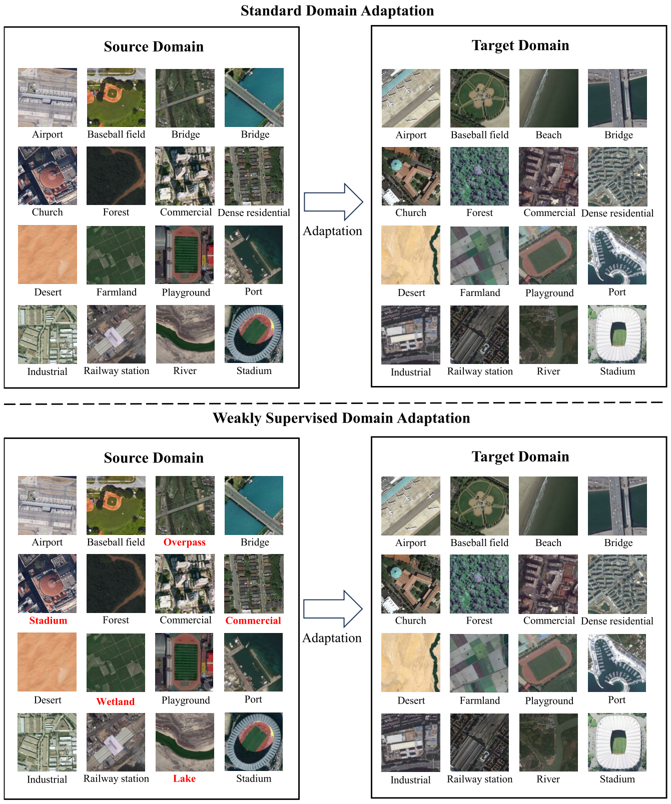

Fig. 1. (Top) Standard DA scenario. (Bottom) NDA scenario. The black labels are correct and the red labels are incorrect.

In practical remote sensing applications, NDA scenarios are quite common. For instance, when collecting a large amount of remote sensing data, it is possible to gather data with incorrect labels without manual supervision. When we directly apply DA methods to data with incorrect labels, the model may learn from these erroneous annotations, resulting in severe negative transfer effects. Some researchers may manually annotate the data, but it is time-consuming, labor-intensive, and demands substantial prior knowledge. Furthermore, even with manual annotation, data might still be labeled incorrectly due to limited knowledge. Especially for remote sensing community, two images may have the same spectral information but represent different categories (i.e., images of "Medium residential" and "Dense residential" may share similar spectral information because their aerial views are similar). Similarly, two images belonging to the same category may have significantly different spectral information (i.e., two images representing "Forest" were taken from different seasons, which brings a huge difference in the coverage of green vegetation). In addition, some images are easily annotated with wrong labels due to similarities in texture features, such as "Grassland" may be incorrectly labeled as "Farmland". Therefore, the labels of the source domain will inevitably be corrupted with noise in real-world scenarios. RAN [6] first tries to solve the problem of NDA in remote sensing and employs a curriculum learning strategy to filter out noisy samples and introduces class and public weighting factors to enhance cross-scene adaptation at the class level, thereby improving the model's discriminative capability and generalizability. RAN only uses the supervision from the source domain and ignore the potential supervision of the target domain, which in the following will be proved to play an important and positive supervisory role in the transfer process. Differently, our proposed method innovatively and fully considers potential supervision from the target domain, thereby effectively and implicitly reducing the label noise.

In this article, we address remote sensing cross-scene classification tasks under the NDA setting: labels in the source domain are partially corrupted with noises, and the samples in the target domain are entirely unlabeled. According to this scenario, we utilize NDA algorithms to mitigate the negative influence brought by noisy labels. By this means, we can greatly economize human interpretation and reduce the adverse transfer effects caused by standard DA methods. Our contributions can be summarized into three main parts.

1) We propose bilateral adaptation network (BAN) to tackle NDA scenarios in the remote sensing community. Our BAN innovatively and fully considers potential supervision from target domain by the mutual learning strategy.
2) BAN explores supervision knowledge derived from pseudo labels in the target domain and leverage bilateral knowledge between two domains, implicitly reducing label noise in the source domain. Moreover, BAN can serve as a universal paradigm to not only improve existing NDA approaches, but also enhance recent DA methods.
3) A comprehensive evaluation of both RGB and hyperspectral remote sensing datasets with different types and levels of noisy annotations validates the superior effectiveness of BAN. BAN improves the average accuracy by 6.70%–15.70% on RGB datasets and overall accuracy (OA) by 1.36%–3.17% on hyperspectral datasets with flip-20% noise compared to other SOTA NDA and DA approaches.

II. Related Work

A. Domain Adaptation

DA methods are designed to construct models capable of generalizing effectively across diverse domains with distinct data distributions. Many methods have been suggested to address this issue. One popular line of researches focus on aligning feature representations between the source and target domains, which aims to reduce the distribution discrepancy between the source and target domains [7], [8]. Besides, some studies exploit a domain discriminator to distinguish source domain and target domains while confusing the discriminator by learning the features adversarially [9], [10].

However, existing DA methods mentioned above typically concentrate on standard DA scenarios, assuming clean source domain data with accurate labels. In practical applications, particularly in extreme environments, the source domain labels are often corrupted by significant noise. Standard DA methods are becoming less suitable for a wide range of DA tasks, emphasizing the increasing importance of NDA scenarios.

B. Learning With Noisy Labels

Diverse algorithms manage to reduce the negative impact of label noise in the source domain by improving model

TABLE I

DETAILS OF OUR COLLECTED RGB DATASETS

Index	AID	UC Merced	NWPU-RESISC45
Year	2017	2010	2017
Classes	31	21	45
Images per class	220 ~ 420	100	700
Images	10,000	2,100	31,500
Resolution (m)	0.5 ~ 8	0.3	0.2 ~ 30
Size (pixel)	600 × 600	256 × 256	256 × 256
Source	Google Earth	USGS	Google Earth

robustness [11], [12], [13]. Early studies learn a robust model with a noise adaption layer [14] or a label transition matrix [15]. Recently, a line of works design robust loss functions, such as generalized cross entropy loss [16], information-theoretic loss [17], and Taylor cross entropy loss [13]. Another line of works focus on label correction using dual networks or contrastive learning [18], [19], [20]. Besides, some studies reweight samples to emphasize clean and reliable data during training [21], [22].

Although the aforementioned methods have shown promising results, they do not fully utilize information supervision from the target domain, which further combat label noise.

C. Noisy Domain Adaptation

In contrast to standard DA, NDA considers the source domain data with noisy labels, which is also referred to as weakly supervised DA. Numerous methods have been proposed to tackle NDA issues. A line of NDA approaches focus on reweighting the source samples in the training procedure. For example, transferable curriculum learning (TCL) [23] adopts curriculum learning strategy to selectively transfer clean source data to avoid negative transfer influence of noisy or irrelevant source data. Co-teaching [24] simultaneously trains two deep neural networks that mutually teach each other by selecting potentially clean data. RAN [6] adopts curriculum learning strategy to select clean data and designs two weighting factors to consider the class information among two domains. Another line of NDA approaches design regular terms to avoid overfitting the noisy samples [25], [26]. Other NDA methods reduce label noise by matching predictions of different models [27] and extracting invariant representations to construct a denoising MMD loss [28].

Although the aforementioned approaches have made notable progress in addressing NDA, they partly acquire and utilize supervision information from the source domain. In contrast, our proposed BAN explores potential supervision information in the target domain and leverages bilateral knowledge between two domains to implicitly reduce label noise.

D. DA in Remote Sensing

Though promising advancements have been made in remote sensing image classification [29], [30], there are two primary challenges that hinder the broader application of this technology [31]: 1) the difficulty of acquiring labeled data in sufficient quantities and 2) the models hardly meet the demand of sufficient generalization capabilities to handle data collected from different area, sensors, and environment. Domain adaption models have been introduced to tackle the challenges associated with long-time-series and large-scale applications that involve remote sensing images varying from different data distributions, which can significantly affect the model's transferable capacity [32]. Recently, DA methods significantly minimize the discrepancy between different images from various sensors and environments and demonstrate excellent performance across remote sensing applications varying from semantic segmentation [33], [34], [35], [36], [37], [38], classification [39], [40], [41], [42], [43], object detection [44], [45], [46], [47], and regression tasks [48], [49].

Existing DA methods in the remote sensing community primarily address standard DA problems. Many researchers have proposed various superior transfer learning strategies, such as partial DA [50], NDA [6], multisource DA [51], [52], multitarget DA [8], [53], and open-set DA [54], [55], [56]. In this article, we formulate the remote sensing cross-scene classification on NDA scenarios and propose a BAN to effectively leverage supervision knowledge between the source domain and target domain, thereby reducing adverse influence derived from noisy labels in the source domain.

III. METHODOLOGY

A. Preliminary and Overview

The standard DA scenario constitutes a labeled source domain ($\mathcal{D}_s = \{(\mathbf{x}_i^s, \hat{y}_i^s)\}_{i=1}^{n_s}$) of n_s labeled samples and an unlabeled target domain ($\mathcal{D}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$) of n_t unlabeled samples, where $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ denote an sample in $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively, and $\hat{y}_i^s$ denotes noisy labels in the source domain. Due to different distributions between source and target domains, the direct deployment of a model trained on source domain onto the target domain is likely to result in a substantial decrease in accuracy. In NDA, labels in source domain are corrupted from ground-truth labels, resulting in a mismatch between labels and samples. This is more common in applications especially dealing with large-scale data collection. Therefore, there are two main challenges in NDA scenarios: 1) the noisy labels in the source domain may lead to transfer of incorrect experiences during learning process; and 2) there is no way for us to identify which labels are incorrect since the target domain is fully unlabeled during training.

The overview of BAN is illustrated in Fig. 2, which consists of two models: $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$. The $\mathcal{M}_\theta$ is the backbone model and the $\tilde{\mathcal{M}}_\theta$ is the additional trained model from the previous training process. The training process contains two main processes: the forward learning (FL) and the backward learning (BL). The FL process primarily learns from the source domain and transfers knowledge to the target domain, while the BL process acquires knowledge from the target domain and transfers it to the source domain. During training, we conduct the FL process and the BL process iteratively to explore bilateral supervision information.

B. Forward Learning

During the first FL process, we need to annotate the samples in the target domain to utilize the bilateral supervision information. Therefore, we train $\mathcal{M}_\theta$ first on the source domain and generate pseudo labels for the target domain using the

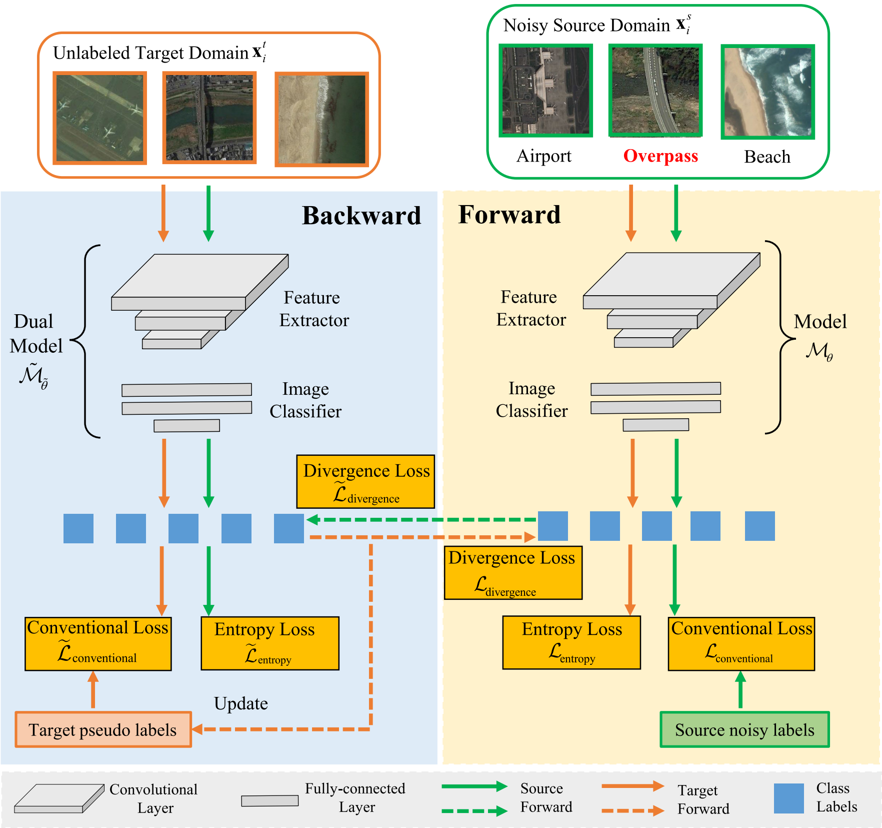
Unlabeled Target Domain $\mathbf{x}_i^t$
Noisy Source Domain $\mathbf{x}_i^s$
Airport
Overpass
Beach
Backward
Forward
Dual Model $\tilde{\mathcal{M}}_{\tilde{\theta}}$
Feature Extractor
Image Classifier
Feature Extractor
Image Classifier
Model $\mathcal{M}_\theta$
Divergence Loss $\tilde{\mathcal{L}}_{\text{divergence}}$
Divergence Loss $\mathcal{L}_{\text{divergence}}$
Conventional Loss $\tilde{\mathcal{L}}_{\text{conventional}}$
Entropy Loss $\tilde{\mathcal{L}}_{\text{entropy}}$
Entropy Loss $\mathcal{L}_{\text{entropy}}$
Conventional Loss $\mathcal{L}_{\text{conventional}}$
Update
Target pseudo labels
Source noisy labels
Convolutional Layer
Fully-connected Layer
Source Forward
Target Forward
Class Labels

Fig. 2. Overview of BAN, including two parts: The FL and BL process. The FL and BL processes are executed iteratively. Every time the FL process finishes, labels for the target domain are updated.

TABLE II

SIX TRANSFER TASKS BASED ON OUR COLLECTED RGB DATASETS. THE CLASS NAMES IN PARENTHESES DENOTE THE CLASS NAMES FOR THE TARGET DOMAIN

Source Domain	Target Domain	Shared Classes
AID	NWPU-RESISC45	Forest, Mountain, River, Industrial, Church, Airport,Port (Harbor), Beach, Medium residential, Sparse residential, Dense residential, Storage tank, Meadow, Stadium, Parking, Playground (Ground track field), Commercial, Bridge Farmland (Rectangular farmland & Circular farmland), Desert, Baseball field, Railway station, viaduct (Overpass)
AID	UC Merced	River, Parking, Commercial (Buildings), Storage tanks, Medium residential, Beach, Baseball field, Overpass (Viaduct), Forest, Viaduct (Overpass), Dense residential, Farmland (Agricultural), Port (Harbor)
UC Merced	AID	Medium residential, Storage tanks, Sparse residential, Agricultural (Farmland), Parking lot, Baseball diamond Buildings (Commercial), Dense residential, Beach, Harbor (Port), Sparse residential, Forest, River
UC Merced	NWPU-RESISC45	Tennis court, Baseball diamond, Agricultural (Circular farmland & Rectangle farmland), Chaparral, Freeway, Harbor, Forest, Sparse residential, Buildings (Commercial area), Medium residential, River, Mobile home park, Airplane, Intersection, Overpass, Runway, Beach, Storage tanks, Parking lot, Golf course
NWPU-RESISC45	AID	Bridge, Medium residential, Forest, Meadow, Beach, Industrial area, Church, Dense residential, Mountain, Parking lot, Desert, Storage tank, Stadium, Railway station, Baseball diamond, River, Ground track field (Playground), Circular farmland & Rectangular farmland (Farmland), Sparse residential, Overpass (Viaduct), Harbor (Port), Airport
NWPU-RESISC45	UC Merced	Overpass, Sparse residential, River, Storage tank, Golf course, Parking lot, Beach, Freeway, Chaparral, Forest, Intersection, Medium residential, Airplane, Baseball diamond, Tennis court, Dense residential,Runway, Mobile home park, Harbor, Circular farmland & Rectangular farmland (Farmland)

following equations:

$$\theta = \underset{\theta}{\operatorname{argmin}} \frac{1}{m_s}\sum_{i=1}^{m_s}\mathcal{L}\big(\hat{y}_i^s, \mathcal{M}\big(\mathbf{x}_i^s, \theta\big)\big) \tag{1}$$

$$\hat{y}_i^t = \underset{k}{\operatorname{argmax}}\mathcal{M}_k\big(\mathbf{x}_i^t, \theta\big) \quad \forall i = 1, 2, \ldots, n_t \tag{2}$$

where m_s denotes the number of samples chosen from the source domain. $\mathcal{L}$ is the loss of the backbone model for training $\mathcal{M}_\theta$. $\mathcal{M}_k$ is the output for the k'th label.

During the FL process, we train the model with a conventional supervised learning loss $\mathcal{L}_{\text{conventional}}$

$$\mathcal{L}_{\text{conventional}} = \frac{1}{m_s}\sum_{i=1}^{m_s}\mathcal{L}\big(\hat{y}_i^s, \mathcal{M}\big(\mathbf{x}_i^s, \theta\big)\big). \tag{3}$$

Although the model trained by minimizing the conventional supervised learning loss $\mathcal{L}_{\text{conventional}}$ demonstrates excellent performance in domains with supervision information [23], [24], [57], it often encounters a substantial decline in accuracy when tested on data from a different domain due to inherent domain shifts. These domain shifts arise from variations in data distribution, imaging conditions, or environmental factors, thereby challenging the model's generalization capability. Consequently, there is a pressing need to address the issue of domain shift to ensure robust and reliable performance across diverse domains.

In order to enhance the model's generalization and improve its performance on a different domain, we propose the incorporation of consistency regularization through the symmetric

Kullback–Leibler (KL) divergence loss. D_{KL} measures the difference between two probability distributions P and Q

$$D_{\text{KL}}(p\|q) = \sum_{i=1}^{n} p(\mathbf{x}_i)\log\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)} \tag{4}$$

where $\mathbf{x}$ is an element from the sample space, n is the number of samples, $p(\mathbf{x})$ is the probability of $\mathbf{x}$ according to distribution P, and $q(\mathbf{x})$ is the probability of $\mathbf{x}$ according to distribution Q. The FL process loss based on the target domain is

$$\mathcal{L}_{\text{divergence}} = D_{\text{KL}}\big(\boldsymbol{p}_1^t \big\| \boldsymbol{p}_2^t\big) + D_{\text{KL}}\big(\boldsymbol{p}_2^t \big\| \boldsymbol{p}_1^t\big) \tag{5}$$

where $\boldsymbol{p}_1^t$ and $\boldsymbol{p}_2^t$ denotes class labels distributions from $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$, respectively. This regularization technique encourages the model and its dual model to align predictions for each sample originating from different domains. By enforcing consistency in the predictions, the model learns to capture bilateral characteristics and reduce the impact of domain shifts. The symmetric KL divergence loss serves as a guidance signal, promoting cross-domain generalization.

However, the aforementioned $\mathcal{L}_{\text{conventional}}$ and $\mathcal{L}_{\text{divergence}}$ do not consider the discrepancies among samples in the target domain. To further decrease the uncertainty of the classifier predictions, we explore the entropy minimization principle for refining the classifiers adaptation, which minimizes the entropy of class-conditional distribution on target domain. Therefore, the entropy loss is as follows:

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{m_t}\sum_{\mathbf{x}_i\in\mathcal{D}_t}\sum_{c=1}^{C} p_c^t(\mathbf{x}_i)\log\big(p_c^t(\mathbf{x}_i)\big) \tag{6}$$

where m_t is the number of samples chosen from target domain, C is the number of classes, and p_c^t is the probability of predicting a sample $\mathbf{x}_i$ from target domain to class c. By minimizing the entropy loss of each sample in the target domain, the samples' predictions will become more confident and certain, thus improving the classifier's performance. Therefore, the overall loss function can be formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{conventional}} + \lambda\mathcal{L}_{\text{divergence}} + \mu\mathcal{L}_{\text{entropy}} \tag{7}$$

where the conventional supervised learning loss is used to optimize the model's performance on the domain with supervision information, and the symmetric KL divergence loss is incorporated as a consistency regularization term to encourage the model and its dual model to mimic predictions for each sample originating from different domains, and the entropy loss aims to enhance the certainty of target domain samples' predictions to improve the classifier's performance. The hyperparameter λ and μ control the importance of the consistency regularization term relative to the supervised learning loss and the entropy objective of classification.

To this end, the FL process which contains the conventional supervised learning loss, symmetric KL divergence loss, and entropy loss, offers three advantages: 1) the symmetric KL divergence loss as a consistency regularization term enables the model and its dual model to align predictions for each sample from different domains; 2) by enforcing consistency in the predictions, the model learns the bilateral information and reduce the impact of domain shifts; and 3) the entropy loss increases the confidence of target domain samples' predictions to improve the classifier's performance.

TABLE III

Details of Our Selected Hyperspectral Datasets

Class Name	Number of Samples	
	Houston 2013	Houston 2018
Grass healthy	345	1,353
Grass stressed	365	4,888
Trees	365	2,766
Water	285	22
Residential buildings	319	5,347
Non-residential buildings	408	32,459
Road	443	6,365
Total	2530	53,200

C. Backward Learning

Existing NDA methods will encounter the problem of error accumulation, where errors resulting from biased instance selection in the previous training iteration will be relearned in the following training. Consequently, the errors learned from the source domain will continuously amplify during iterative training, leading to a reduction in the model's accuracy. Several NDA works [23], [28] mainly focus on effectively leveraging information from the source domain and transferring it to the target domain to reduce domain discrepancy and mitigate the influence of noisy labels.

Inspired by mutual learning [58], in addition to obtaining supervision knowledge from the source domain, we can also extract valuable supervision information from the target domain. Therefore, we design a bilateral learning approach that effectively acquires and utilizes mutual supervision information between the source and target domains. This approach compensates for limitations of reliance on the supervision knowledge of the source domain and reduces adverse influence derived from noisy labels.

The bilateral learning approach consists of two models: $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$. The $\mathcal{M}_\theta$ is the backbone model and the $\tilde{\mathcal{M}}_\theta$ is the additional trained model from the previous training process. The training process comprises two main processes: the FL and the BL process. The FL process primarily aims to transfer learned knowledge from the source domain to target domain, whereas the BL process primarily focuses on transferring the learned knowledge from the target domain to source domain. FL process has been detailed in Section III-B. In the following, we will focus on elucidating the BL process.

Similar to the FL process, during the BL process, we train the model with a conventional supervised learning loss $\tilde{\mathcal{L}}_{\text{conventional}}$

$$\tilde{\mathcal{L}}_{\text{conventional}} = \frac{1}{m_t}\sum_{i=1}^{m_t}\mathcal{L}\big(\hat{y}_i^t, \tilde{\mathcal{M}}(\mathbf{x}_i^t, \theta)\big). \tag{8}$$

Therefore, BL process adopts symmetric KL divergence loss to promote cross-domain generalization

$$\tilde{\mathcal{L}}_{\text{divergence}} = D_{\text{KL}}\big(\boldsymbol{p}_1^s \big\| \boldsymbol{p}_2^s\big) + D_{\text{KL}}\big(\boldsymbol{p}_2^s \big\| \boldsymbol{p}_1^s\big) \tag{9}$$

where $\boldsymbol{p}_1^s$ and $\boldsymbol{p}_2^s$ denotes class labels distributions from $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$, respectively. Similarly, we adopt the entropy loss to

TABLE IV

ACCURACY (%) ON OUR COLLECTED RGB DATASET WITH UNIF-20% NOISE

Method	A → N	N → A	A → U	U → A	U → N	N → U	Avg
DAN [7]	68.33	76.66	73.67	63.20	46.41	73.99	67.04
CDAN [10]	68.16	74.92	65.94	53.90	45.18	68.70	62.80
TADA [24]	66.79	79.02	66.56	57.92	41.00	53.13	60.74
JoCoR [63]	66.86	72.98	63.75	55.23	44.30	65.50	61.44
GearNet [27]	73.33	85.84	81.88	75.40	56.11	81.92	75.75
UniDA [64]	67.32	73.53	64.38	54.80	45.56	68.46	62.34
ResNet-50 [65]	67.18	73.91	66.80	55.00	47.58	67.88	63.06
BAN+ResNet-50 [65]	74.09	85.24	70.88	77.18	48.87	78.61	72.48
DANN [9]	67.26	77.81	75.08	60.21	48.56	75.53	67.41
BAN+DANN [9]	72.89	86.51	87.76	85.70	56.01	85.34	79.04
Co-teaching [24]	63.98	76.95	64.38	59.49	48.44	60.38	62.27
BAN+Co-teaching [24]	85.50	85.80	76.25	77.71	64.38	64.95	75.77
TCL [23]	76.76	85.79	86.48	76.03	56.13	87.84	78.17
BAN+TCL [23]	77.49	87.89	87.42	86.72	60.70	84.62	80.65
SDAT [66]	75.52	88.8	91.95	87.72	65.23	85.10	82.39
BAN+SDAT [66]	78.91	92.08	**92.81**	91.44	**84.39**	87.88	**87.92**
MIC [67]	82.42	92.97	90.63	91.46	62.68	89.76	84.99
BAN+MIC [67]	**86.14**	**93.74**	90.00	**92.54**	70.27	**90.05**	87.12

enhance the certainty of samples in the source domain

$$\tilde{\mathcal{L}}_{\text{entropy}} = -\frac{1}{m_s}\sum_{\mathbf{x}_i\in\mathcal{D}_s}\sum_{c=1}^{C} p_c^s(\mathbf{x}_i)\log\left(p_c^s(\mathbf{x}_i)\right) \tag{10}$$

where C is the number of classes, and p_c^s is the probability of predicting a sample $\mathbf{x}_i$ from source domain to class c. We adopt the same overall loss as the FL process

$$\tilde{\mathcal{L}}_{\text{total}} = \tilde{\mathcal{L}}_{\text{conventional}} + \lambda\tilde{\mathcal{L}}_{\text{divergence}} + \mu\tilde{\mathcal{L}}_{\text{entropy}} \tag{11}$$

where hyperparameters λ and μ shares the same value of that in $\mathcal{L}_{\text{total}}$.

D. Optimization

The whole training process conducts FL process and BL process in an iterative manner as Algorithm 1 and Fig. 2. At the beginning of the whole training process, we generate pseudo labels for the target domain by (1) and (2). We first adopt BL process to train $\tilde{\mathcal{M}}_{\tilde{\theta}}$ with the overall loss $\tilde{\mathcal{L}}_{\text{total}}$ as (11), where the loss $\tilde{\mathcal{L}}_{\text{conventional}}$ is based on the samples in the target domain, the loss $\tilde{\mathcal{L}}_{\text{divergence}}$ is based on the $\tilde{\mathcal{M}}_{\tilde{\theta}}$ and pretrained $\mathcal{M}_\theta$ on the source domain and the loss $\tilde{\mathcal{L}}_{\text{entropy}}$ is based on the samples in the source domain. Then, we adopt the FL process to train $\mathcal{M}_\theta$ with the overall loss $\mathcal{L}_{\text{total}}$ as (7), where the loss $\mathcal{L}_{\text{conventional}}$ is based on the samples in the source domain, the loss $\mathcal{L}_{\text{divergence}}$ is based on the $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_{\tilde{\theta}}$ trained during the BL process on the target domain, and the loss $\mathcal{L}_{\text{entropy}}$ is based on the samples on the target domain. After a round of BL process and FL process, the pseudo labels of the target domain will be updated and utilized in the following training process. We repeat the BL process and the FL process iteratively until the whole algorithm stops. The future iterations of the models could benefit from mechanisms that selectively emphasize reliable information from pseudo labels during training.

E. Realization

In application, the BAN serves as a universal paradigm and is used to apply to backbone methods. With the help of BAN, the backbone methods are able to exploit bilateral information across domains and improve its performance. Before the FL process, we first use the backbone methods to annotate the samples. Then, we initialize the two models $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$ using the backbone methods. However, these backbone methods have different learning strategies, so their loss functions are generally formed as

$$\mathcal{L}_{\text{backbone}} = \mathbb{E}_{p^s(\mathbf{x}^s,\hat{y}^s),p^t(\mathbf{x}^t)}\left(\mathcal{L}\left(\mathbf{x}^s,\hat{y}^s,\mathbf{x}^t;\mathcal{M}_\theta\right)\right) \tag{12}$$

$$\tilde{\mathcal{L}}_{\text{backbone}} = \mathbb{E}_{p^t(\mathbf{x}^t,\hat{y}^t),p^s(\mathbf{x}^s)}\left(\mathcal{L}\left(\mathbf{x}^t,\hat{y}^t,\mathbf{x}^s;\tilde{\mathcal{M}}_{\tilde{\theta}}\right)\right) \tag{13}$$

where the $\mathcal{L}_{\text{bone}}$ and the $\tilde{\mathcal{L}}_{\text{bone}}$ denote the loss function of the backbone methods for training the $\mathcal{M}$ and the $\tilde{\mathcal{M}}$, respectively. $p^s(*)$ denotes the distribution of the source domain, while $p^t(*)$ denotes the distribution of target domain. In implementation, the two aforementioned loss functions replace the function in (3) and (8), assuming different meanings across different backbone methods. As for (5) and (9), in order to ensure mutual knowledge and information transfer between the two backbone models $\mathcal{M}$ and $\tilde{\mathcal{M}}$, the model is encouraged to leverage the corresponding class posteriors of its dual model to align its predicted probability distribution in the training process. For the task of scene classification in remote sensing, we only need to align classification predictions. Especially for the models that have multiclassifiers, each classifier of the dual model needs to compute the two losses to ensure comprehensive learning of the supervision information. Equation (6) and (10) encourages the two backbone models $\mathcal{M}$ and $\tilde{\mathcal{M}}$ to enhance the certainty of samples from the target domain and source domain, respectively.

IV. DATASETS

A. RGB-Band Remote Sensing Datasets

The details of three public remote sensing datasets (AID [59], UC Merced [60], and NWPU-RESISC45 [61]) are provided in Table I and Fig. 3. Images in these three datasets vary in resolution, size, and lighting conditions, thus making them suitable for evaluating DA approaches for remote sensing cross-scene classification [56], [62]. We set up six transfer tasks in AID, UC Merced, and NWPU-RESISC45

TABLE V

ACCURACY (%) ON OUR COLLECTED RGB DATASET WITH FLIP-20% NOISE

Method	**A → N**	**N → A**	**A → U**	**U → A**	**U → N**	**N → U**	**Avg**
DAN [7]	66.63	62.03	72.89	57.59	46.74	69.86	62.62
CDAN [10]	59.77	65.24	64.30	49.08	42.66	57.64	56.45
TADA [24]	61.03	72.82	70.31	45.86	34.81	47.31	55.36
JoCoR [63]	58.13	65.79	64.30	49.76	41.67	55.53	55.86
GearNet [27]	68.96	81.89	81.09	70.11	54.48	76.63	72.20
UniDA [64]	58.68	63.16	63.98	49.23	42.14	58.17	55.90
ResNet-50 [65]	58.24	65.11	65.46	48.39	41.20	59.95	56.39
BAN+ResNet-50 [65]	81.58	83.65	59.48	76.02	47.97	69.58	69.71
DANN [9]	63.64	71.32	72.89	48.72	45.21	69.81	61.93
BAN+DANN [9]	82.38	70.79	79.77	69.70	73.22	77.41	75.55
Co-teaching [24]	61.24	73.78	66.72	55.24	44.32	56.78	59.68
BAN+Co-teaching [24]	85.98	85.63	75.62	78.20	55.53	71.30	75.38
TCL [23]	78.42	84.80	87.03	67.63	57.87	82.36	76.35
BAN+TCL [23]	86.11	85.90	84.30	83.36	**79.62**	79.04	83.05
SDAT [66]	76.46	87.37	88.13	83.69	58.86	88.41	80.49
BAN+SDAT [66]	**87.15**	92.78	88.83	90.67	67.81	**92.64**	**86.65**
MIC [67]	80.96	93.52	89.61	89.94	64.98	82.45	83.58
BAN+MIC [67]	85.22	**94.22**	**91.25**	**92.78**	70.77	83.37	86.27

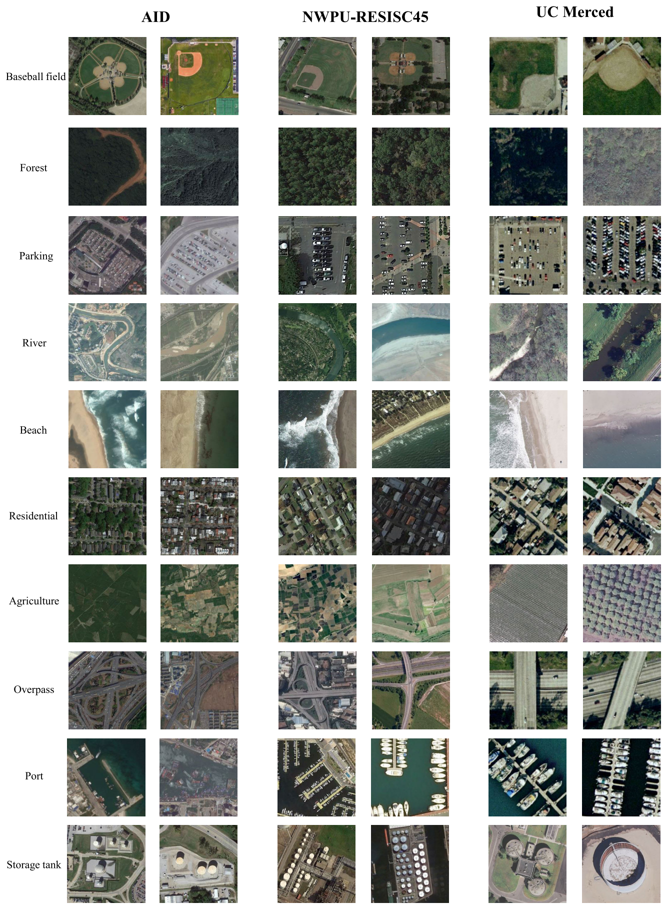

Fig. 3. Ten examples of classes (baseball field, forest, parking, river, beach, residential, agriculture, overpass, port, and storage tank) from three remote sensing datasets (AID, NWPU-RESISC45, and UC Merced).

by pairs and the labels are shown in Table II. We choose shared classes between two domains to ensure that the source and target domains have consistent classes. In contrast to standard DA datasets, we evaluate our proposed method on multiple remote sensing datasets. This allows deviating from the standard practice and allows for a more comprehensive assessment of our method's performance across diverse remote sensing scenarios. However, images from any two remote sensing datasets may be distinguished from diverse land use and land cover classes (see Tables I and II), which brings difficulty in standardizing the classes across different remote sensing datasets. During training and testing, input images are resized to 256 × 256, randomly cropped to 224 × 224, horizontally flipped with a probability of 0.5 for augmentation.

B. Hyperspectral Datasets

Houston dataset includes Houston 2013 [68] and Houston 2018 [69], and their classes and number of samples are detailed in Table III. These two datasets were collected over the urban area of the University of Houston campus and its neighborhoods over different years. Hyperspectral images in these two datasets are characterized by high spectral resolution, complex structure, and rich diversity, thus providing a more challenging and comprehensive transfer task for evaluating the robustness and effectiveness of DA approaches in remote sensing cross-scene classification. Fig 4 shows their pseudo-color and ground truth maps. The aforementioned two datasets are detailed as follows: We set up the transfer task from source dataset Houston 2013 to target dataset Houston 2018. Data augmentation strategies are employed in these two datasets. Specially, for hyperspectral data $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C denote the height, width, and number of spectral channels, we first apply $L2$-normalization across spectral bands after removing NaN values. We then extract spatial patches $\mathbf{P}ij \in \mathbb{R}^{k \times k \times C}$ centered at each pixel (i, j), where k is the patch size. Three data augmentation strategies are employed with 0.5 probability: 1) random flipping; 2) radioactive transformation ($\alpha \mathbf{P}ij + \beta \mathcal{N}(0, 1)$, where $\alpha \in [0.9, 1.1]$ and $\beta = 1/25$); and 3) mixture noise that combines two patches from the same class. Finally, patches are converted to channel-first tensors for network training.

C. Label Corruption

Since the collected datasets are meticulously annotated with only minor erroneous labels, we manually add noise to labels to simulate noisy annotations. A common way for modeling the label noise assumes that the corruption process

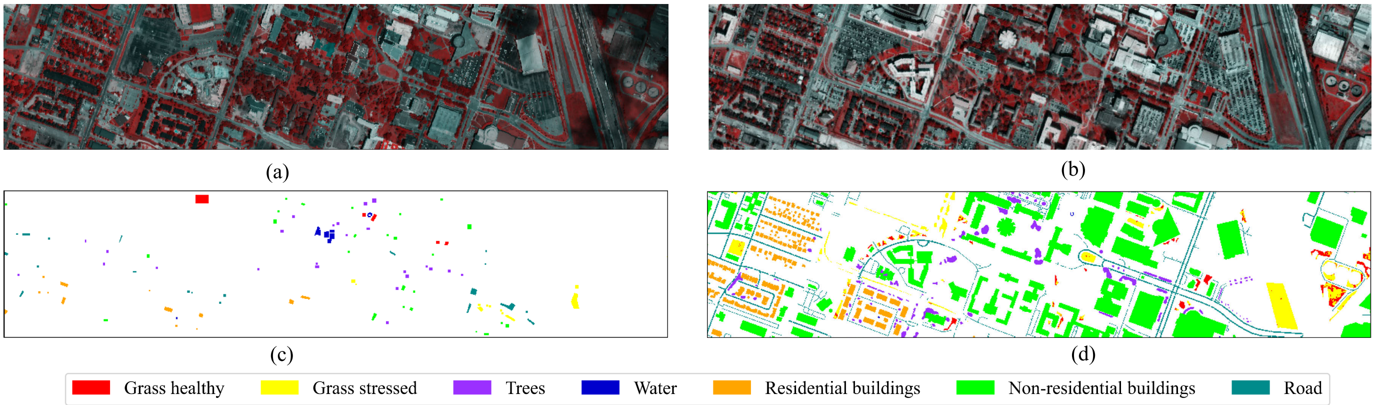

Fig. 4. Pseudo-color image and ground truth map of Houston datasets. (a) Pseudo-color image of Houston 2013. (b) Pseudo-color image of Houston 2018. (c) Ground truth map of Houston 2013. (d) Ground truth map of Houston 2018.

TABLE VI

CLASS-SPECIFIC AND OVERALL CLASSIFICATION ACCURACY (%) ON HOUSTON 2013 → HOUSTON 2018 WITH UNIF-20% NOISE

Class	1	2	3	4	5	6	7	OA
DAN [7]	85.37	57.06	**71.62**	81.82	40.25	48.03	18.32	46.71
CDAN [10]	28.31	79.75	65.76	81.82	35.96	60.69	71.69	60.72
TADA [71]	0.07	0.00	0.00	0.00	0.22	0.46	**99.91**	12.26
JoCoR [63]	12.20	**89.32**	67.03	95.45	33.21	78.87	43.28	68.68
GearNet [27]	**89.73**	33.10	69.81	72.73	69.31	59.39	63.39	59.77
VGG-16 [72]	26.02	85.19	70.68	**95.45**	43.20	59.24	62.03	60.11
BAN+VGG-16	82.63	53.13	69.92	4.55	57.53	73.82	65.26	69.25
DANN [9]	70.88	47.71	81.53	95.45	64.37	60.21	60.42	59.90
BAN+DANN	84.70	35.80	58.21	72.73	71.50	68.00	54.80	63.76
Co-teaching [24]	12.34	78.58	48.70	95.45	42.81	77.49	72.88	70.50
BAN+Co-teaching	9.24	81.12	40.17	86.36	**71.57**	**87.87**	33.28	**74.57**
TCL [23]	78.94	42.31	47.32	77.27	51.30	62.55	77.58	60.99
BAN+TCL	60.83	38.69	71.62	54.55	66.39	64.56	66.80	62.91

is conditionally independent of the data features when given correct labels. We generate label noise by corrupting true labels by label transition matrix, which defines the probability of the correct label being corrupted to noisy label. We partially inject two types of noise for the source domain labels: Uniform noise [70] and flip noise [15].

1) *Uniform noise:* changes the ground-truth label of each sample to wrong class label with probability of (r/D), while the label still has probability $1-[(1-D)/D]$ of being correct, where r denotes noise rate and D denotes the dimension of the label space. The label transition matrix is as

$$K_U = \begin{bmatrix} 1-\dfrac{1-D}{D}r & \dfrac{r}{D} & \cdots & \dfrac{r}{D} \\ \dfrac{r}{D} & 1-\dfrac{1-D}{D}r & \cdots & \dfrac{r}{D} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{r}{D} & \dfrac{r}{D} & \cdots & 1-\dfrac{1-D}{D}r \end{bmatrix}_{D\times D}. \tag{14}$$

2) *Flip noise:* changes the ground-truth label of each sample to a similar class with probability of (r/D), and the label transition matrix is as follows:

$$K_F = \begin{bmatrix} 1-r & 0 & \cdots & r \\ r & 1-r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r & 0 & \cdots & 1-r \end{bmatrix}_{D\times D}. \tag{15}$$

If source labels are corrupted by 20% flip noise, it means that 20% of labels in source domain are incorrect.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

The all experiments were conducted on Linux system with an Intel Core i9-12900K processor (16 cores, 3.7 GHz) and an NVIDIA RTX 3090 GPU (24 GB VRAM). The software environment included Ubuntu 20.04 LTS, Python 3.8, PyTorch 2.0, and CUDA 11.8. When training from scratch, we set a momentum as 0.9 and the initial learning rate as 0.003 for SGD [73]. The total epochs is set to 200 and the batch size is 32. The steps of the whole training process are set as 10. For each transfer task, we evaluate the average accuracy across all target domain samples under the setting of NDA on the collected datasets detailed in Section IV. The six transfer tasks of RGB datasets are **A** → **N**, **N** → **A**, **A** → **U**, **U** → **A**, **U** → **N**, and **N** → **U**, where **A**, **U**, and **N** denote AID, UC Merced, and NWPU-RESISC45, respectively. The transfer

Algorithm 1 Algorithm of BAN

```
Require: Source domain dataset $\mathcal{D}_s = \{(\mathbf{x}_i^s, \hat{y}_i^s)\}_{i=1}^{n_s}$ with
    noisy labels $\hat{y}_i^s$ and target domain dataset $\mathcal{D}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$
    unlabeled where $n_s$ and $n_t$ are the number of images in
    $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively, $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ are samples in $\mathcal{D}_s$ and
    $\mathcal{D}_t$, respectively, learning rate $\alpha$, max epochs $N$, and max
    steps $M$.
Ensure: The pretrained model $\mathcal{M}_\theta$ and its dual model $\tilde{\mathcal{M}}_\theta$.
Generate pseudo labels $\{\hat{y}_i^t\}_{i=1}^{n_t}$ on the target domain $\mathcal{D}_t$
    by $\mathcal{M}_\theta$
for $i = 0 : M$ do
    Initialize $\tilde{\mathcal{M}}_{\tilde{\theta}}$
    for $i = 0 : N$ do
        Fetch: $\{\mathbf{x}_i^t, \hat{y}_i^t\}_{i=1}^{m_t}$ from $\mathcal{D}_t$, $\{\mathbf{x}_i^s\}_{i=1}^{m_s}$ from $\mathcal{D}_s$
        Calculate: $\tilde{\mathcal{L}}_{\text{conventional}} = \frac{1}{m_t}\sum_{i=1}^{m_t}\mathcal{L}(\hat{y}_i^t, \tilde{\mathcal{M}}(\mathbf{x}_i^t, \tilde{\theta}))$;
        Forward: $\boldsymbol{p}_1^s = \mathcal{M}(\mathbf{x}_i^s, \theta)$, $\boldsymbol{p}_2^s = \tilde{\mathcal{M}}(\mathbf{x}_i^s, \tilde{\theta})$;
        Calculate: $\tilde{\mathcal{L}}_{\text{divergence}}$ by Eq. (9) using $\boldsymbol{p}_1^s$ and $\boldsymbol{p}_2^s$;
            $\tilde{\mathcal{L}}_{\text{entropy}}$ by Eq. (10) using $p_c^s(\mathbf{x}_i)$;
        Obtain: $\tilde{\mathcal{L}}_{\text{total}}$ by Eq. (11);
        Update: $\tilde{\theta} = \tilde{\theta} - \alpha\Delta\tilde{\mathcal{L}}_{\text{total}}$
    end for
    Initialize $\mathcal{M}_\theta$
    for $i = 0 : N$ do
        Fetch: $\{\mathbf{x}_i^s, \hat{y}_i^s\}_{i=1}^{m_s}$ from $\mathcal{D}_s$, $\{\mathbf{x}_i^t\}_{i=1}^{m_t}$ from $\mathcal{D}_t$
        Calculate: $\mathcal{L}_{\text{conventional}} = \frac{1}{m_s}\sum_{i=1}^{m_s}\mathcal{L}(\hat{y}_i^s, \mathcal{M}(\mathbf{x}_i^s, \theta))$;
        Forward: $\boldsymbol{p}_1^t = \mathcal{M}(\mathbf{x}_i^t, \theta)$, $\boldsymbol{p}_2^t = \tilde{\mathcal{M}}(\mathbf{x}_i^t, \tilde{\theta})$;
        Calculate: $\mathcal{L}_{\text{divergence}}$ by Eq. (5) using $\boldsymbol{p}_1^t$ and $\boldsymbol{p}_2^t$;
            $\mathcal{L}_{\text{entropy}}$ by Eq. (6) using $p_c^t(\mathbf{x}_i)$;
        Obtain: $\mathcal{L}_{\text{total}}$ by Eq. (7);
        Update: $\theta = \theta - \alpha\Delta\mathcal{L}_{\text{total}}$;
    end for
    Update: $\{\hat{y}_i^t\}_{i=1}^{n_t}$ by $\mathcal{M}_\theta$
end for
return $\mathcal{M}_\theta$ and $\tilde{\mathcal{M}}_\theta$
```

task of hyperspectral datasets is **Houston 2013 → Houston 2018**. All samples in the source domain dataset are used during training, and all samples in the target domain are used during testing to calculate the average classification accuracy.

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

We use ResNet-50 [65] for RGB-band dataset and VGG-16 [72] for hyperspectral dataset as the backbone. BAN is compared with 11 state-of-the-art baselines, including seven standard DA methods (DAN [7], DANN [9], CDAN [10], TADA [71], SDAT [66], UniDA [64], and MIC [67]) and four NDA methods (co-teaching [24], TCL [23], JoCoR [63], and GearNet [27]). The quantitative evaluation criterion is based on the classification accuracy of the model on target domain samples. Tables IV and V list the target domain accuracy on our collected RGB datasets, and Tables VI and VII list the class-specific accuracy and the OA on the target dataset Houston 2018 with different types and levels of label noise.

According to Tables IV–VII, the results show that under different types and levels of noise, BAN outperforms all the comparison methods by a large margin on all transfer tasks. It indicates that BAN not only improves the accuracy of existing methods but also maintains high accuracy under different noise conditions. Specifically, Tables IV and V show results on our collected RGB datasets, Among DA methods, DAN [7], DANN [9], CDAN [10], SDAT [66], and MIC [67] outperform ResNet-50 [65] and the reason is that all of them focus on discriminative information, which can slightly reduce the negative impact of source domain noise while decreasing the domain gap. However, TADA [71] and UniDA [64] exhibit comparatively lower performance compared to ResNet-50 [65]. Since TADA [71] requires global and local attention from transferable images and UniDA [64] generates data from noisy source domain, both of them struggle to eliminate the negative impact of noise from source data and fail in limited performance. Among NDA methods, co-teaching [24], JoCoR [63] and TCL [23] adopt a selective strategy to transfer knowledge from clean source domain data, and Gearnet [27] matches predictions of different models, thus reduce the affection of noisy samples. However, they only acquire knowledge from the source domain, which limits the model's performance. Our proposed BAN innovatively exploits potential supervision from the target domain and leverages bilateral information between domains, thus not only mitigating the negative influence from noisy samples but also maintaining superior and robust performance under different types and levels of noise across all transfer tasks.

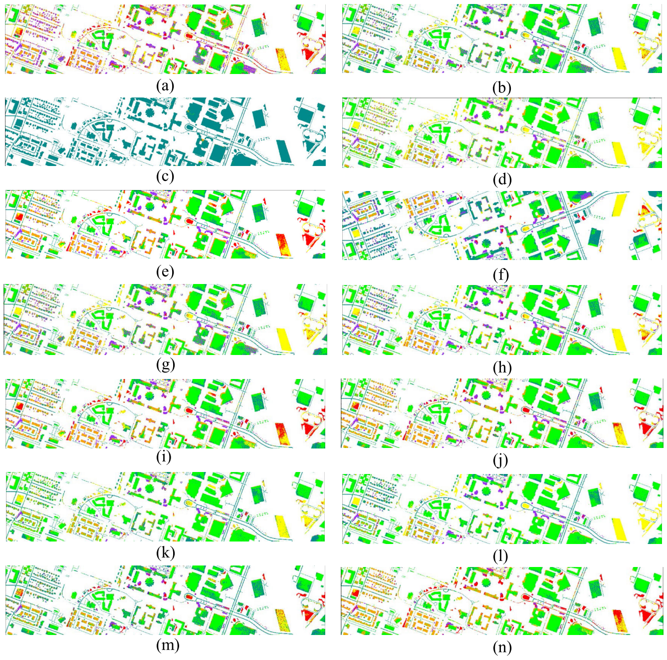
(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n)

Fig. 5. Data visualization and classification maps for the target hyperspectral dataset Houston 2018 with Unif-20% noise. Methods includes. (a) DAN (46.71%). (b) CDAN (60.72%). (c) TADA (12.26%). (d) JoCoR (68.68%). (e) GearNet (59.77%). (f) UniDA (58.02%). (g) ResNet-50 (60.11%). (h) BAN + ResNet-50 (69.25%). (i) DANN (59.90). (j) BAN + DANN (63.76%). (k) Co-teaching (70.50%). (l) BAN + co-teaching (74.57%). (m) TCL (60.99%). (n) BAN + TCL (62.91%).

Specifically, Tables VI and VII list results on the target hyperspectral dataset Houston 2018. Most of DA methods perform similar to their performance on the RGB dataset. However, DAN [7] shows inferior results compared to VGG-16 [72], since the overfitting and negative transfer caused

TABLE VII

Class-Specific and Overall Classification Accuracy (%) on Houston 2013 → Houston 2018 With Flip-20% Noise

Class	1	2	3	4	5	6	7	OA
DAN [7]	80.27	55.28	64.35	54.55	43.82	53.33	53.39	53.82
CDAN [10]	18.26	55.45	47.40	77.27	36.53	83.22	51.26	68.63
TADA [71]	0.22	2.03	0.18	36.36	5.35	**99.20**	0.41	61.33
JoCoR [63]	14.63	57.79	31.16	72.73	39.80	86.76	25.44	67.31
GearNet [27]	73.54	48.00	66.52	36.36	**80.74**	70.30	52.21	67.01
VGG-16 [72]	23.36	71.36	64.93	68.18	53.86	77.29	41.21	68.05
BAN+VGG-16	63.86	55.99	62.73	50.00	53.67	83.08	43.41	71.33
DANN [9]	67.70	70.43	61.93	68.18	68.06	68.68	56.70	65.18
BAN+DANN	78.86	37.52	57.45	9.09	53.67	73.99	**58.79**	66.01
Co-teaching [24]	22.47	72.07	60.20	72.73	54.33	85.60	40.27	72.86
BAN+Co-teaching	23.13	**78.99**	**66.59**	68.18	56.63	85.34	57.66	**75.99**
TCL [23]	65.93	44.54	64.82	22.73	57.15	77.50	57.17	69.02
BAN+TCL	**81.37**	28.56	49.86	31.82	66.32	84.11	42.64	70.38

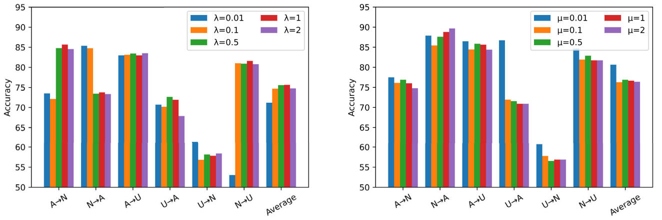

Fig. 6. Accuracy (%) sensitivity of BAN to hyperparameter (left) λ and (right) μ on our collected RGB dataset with Unif-20% noise.

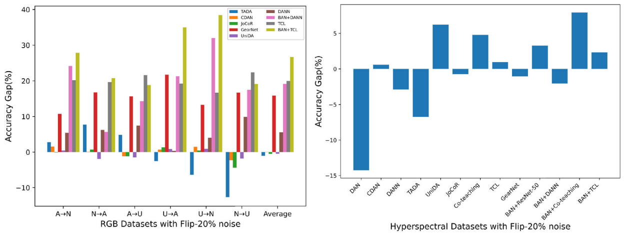

Fig. 7. Accuracy Gap (%) of different methods with flip-20% noise for NDA scenes.

by noisy labels will reduce the model's discriminative ability. Among NDA methods, JoCoR [63], GearNet [27], and TCL [23] perform similar to VGG-16 [72], while co-teaching [24] exhibits better performance. Particularly, BAN improves the OA by 1.36%–4.60% compared to other state-of-the-art NDA approaches, with the highest accuracy of 75.99%. For class-specific accuracy on Table VII, TADA [71] reaches the highest accuracy on class 6 while performing worse on other classes. That is because these methods assume all labels are accurate, leading them to overfit noisy labels, which distorts the learning process and hinders its ability to generalize to clean data. Differently, although BAN + co-teaching [24] does not achieve the highest accuracy across all classes, it achieves the highest OA, proving that BAN reduces impacts of noisy labels and improves the robustness. Fig. 5 shows data visualization and classification maps for target dataset Houston 2018 with Unif-20% noise obtained from different approaches.

C. Ablation Studies

As shown in Table VIII, we investigate the effectiveness of the consistency regularization and entropy minimization by ablation study: 1) w/o $\mathcal{L}_{\text{d}}$ & $\mathcal{L}_{\text{e}}$ removes $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$; 2) w/o $\mathcal{L}_{\text{d}}$ removes $\mathcal{L}_{\text{divergence}}$; 3) w/o $\mathcal{L}_{\text{e}}$ removes $\mathcal{L}_{\text{entropy}}$; and 4) w/ means the full BAN with $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$. Results of ablation study on our collected RGB datasets with uniform noise are shown in Table VIII. We choose ResNet-50 [65], co-teaching [24], and DANN [9] as the backbones. Without $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$, BAN performs very limited in all transfer task. With combination of $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$ separately, BAN demonstrates better performance, and the full combination of $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$ results in superior performance. This indicates that: 1) Both $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$ can individually enhance BAN's performance and 2) $\mathcal{L}_{\text{divergence}}$ and $\mathcal{L}_{\text{entropy}}$ are able to mutually complement each other to achieve the best performance of BAN. Therefore, the ablation study quantifies the importance of the consistency regularization and entropy minimization within BAN.

D. Sensitive Analysis

Fig. 6 presents ablation studies of BAN for our datasets in NDA scenarios. We conduct experiments using hyperparameters λ and μ in (7) and (11) for BAN + TCL. We evaluate λ and μ ranging from 0.01 to 2.0. It is evident that BAN performs best when λ and μ are set to 0.5 and 0.01.

VI. Discussion

A. Accuracy Gap

To further analyze the performance of standard DA methods, NDA methods and our proposed BAN, we compute their accuracy gap between them and ResNet-50 using the following equation:

$$\text{Accuracy Gap} = \text{ACC}_{\text{DA}} - \text{ACC}_{\text{ResNet-50}} \tag{16}$$

where ACC_{DA} and $\text{ACC}_{\text{ResNet-50}}$ denote the accuracy of DA algorithms and ResNet-50, respectively. Accuracy Gap $>$ 0 indicates that DA methods perform positively in NDA scenarios. Otherwise, Accuracy Gap $<$ 0 suggests that the performance of DA methods deteriorates in NDA scenarios.

As shown in Fig. 7, both DA and NDA methods exhibit a decline in accuracy in NDA scenarios. For instance, CDAN and TADA experience a decrease of 1.03%, UniDA shows a reduction of 0.50%, and JoCoR experiences a drop of 0.53%, all of which underscore the negative transfer effects rising

TABLE VIII

RESULTS OF ABLATION STUDIES ON OUR COLLECTED RGB DATASETS WITH UNIFORM NOISE. "W/O $\mathcal{L}_{\mathrm{d}}$" AND "W/O $\mathcal{L}_{\mathrm{e}}$" MEAN THAT BAN DOES NOT APPLY $\mathcal{L}_{\mathrm{divergence}}$ AND $\mathcal{L}_{\mathrm{entropy}}$, RESPECTIVELY. W/ MEANS THE FULL BAN WITH $\mathcal{L}_{\mathrm{divergence}}$ AND $\mathcal{L}_{\mathrm{entropy}}$

	BAN+ResNet-50 [65]				BAN+Co-teaching [24]				BAN+DANN [9]			
Tasks	w/o $\mathcal{L}_{\mathrm{d}}$ & $\mathcal{L}_{\mathrm{e}}$	w/o $\mathcal{L}_{\mathrm{d}}$	w/o $\mathcal{L}_{\mathrm{e}}$	w/	w/o $\mathcal{L}_{\mathrm{d}}$ & $\mathcal{L}_{\mathrm{e}}$	w/o $\mathcal{L}_{\mathrm{d}}$	w/o $\mathcal{L}_{\mathrm{e}}$	w/	w/o $\mathcal{L}_{\mathrm{d}}$ & $\mathcal{L}_{\mathrm{e}}$	w/o $\mathcal{L}_{\mathrm{d}}$	w/o $\mathcal{L}_{\mathrm{e}}$	w/
A-N Unif-20%	66.67	72.42	67.94	74.09	67.61	68.95	78.51	85.50	66.73	69.78	72.14	72.89
N-A Unif-20%	78.55	79.67	79.64	85.24	81.28	83.53	83.87	85.80	79.56	80.47	85.84	86.51
A-U Unif-20%	63.20	67.89	63.20	70.88	63.05	70.08	70.00	76.25	63.05	76.48	81.88	87.76
U-A Unif-20%	53.26	55.30	72.29	77.18	63.43	73.68	71.46	77.71	56.51	72.14	75.40	85.70
U-N Unif-20%	45.89	47.51	45.89	48.87	46.56	63.32	61.61	64.38	44.63	53.64	49.04	56.01
N-U Unif-20%	66.30	67.98	68.17	78.61	59.71	65.91	64.90	64.95	69.47	76.30	81.92	85.34
Avg	62.31	65.13	66.19	**72.48**	63.60	70.91	74.37	**76.57**	63.32	71.47	75.75	**79.04**

from noisy labels. While methods such as DANN, co-teaching, and TCL demonstrate positive performance, adopting BAN achieves a significantly superior outcome, with the highest accuracy gap reaching 26.66%. As for hyperspectral data, most DA methods face a severe transfer impact, with DAN exhibiting the most negative effect of -14.24%. Although NDA methods implement strategies to mitigate the negative effects of noisy labels, they still experience significant accuracy reductions, such as GearNet experiences an accuracy gap of -1.05%. Consequently, NDA algorithms that perform well in computer vision domain may not uniformly sustain equivalent performance when applied to the remote sensing domain.

B. Different Noise Levels

Fig. 8 lists the classification accuracy results of BAN with different noise on $\mathbf{A} \rightarrow \mathbf{U}$ transfer task. We use BAN + SDAT [66] as the main example for experimental analysis. The noises range in [0.2, 0.3, 0.4, 0.5, 0.6]. The results demonstrate the following findings.

1) BAN consistently performs best with all noise levels for both uniform and flip noise, indicating that BAN is able to handle most noisy environments in practice.
2) Flip noise is more disruptive than uniform noise since the accuracy of most methods declines significantly as the noise level increases, yet BAN still maintains a high level of accuracy, demonstrating its robustness.
3) Unlike methods such as UniDA [64], TCL [23], and MIC [67] where accuracy experiences significant fluctuations with increasing noise levels, BAN maintains stable performance, demonstrating that BAN can cope with changes in noise levels more robustly.

C. Potential of BAN for Remote Sensing Applications

For remote sensing image classification scenarios, NDA poses a more practical and challenging problem. In recent years, numerous studies have addressed and developed solutions for remote sensing scene classification on standard DA scenes. However, there has been limited research in addressing NDA issues. In the remote sensing community, source domain data collection may introduce label noise due to human errors and discrepancies in data sources. Moreover, domain shifts consistently manifest in the remote sensing data, which often arise due to the diverse and dynamic nature of data acquisition. These shifts can be broadly categorized into three main types. First, sensor specification differences frequently lead to variations in data characteristics. For instance, discrepancies in spatial resolution (e.g., high-resolution imagery versus lower resolution data) or spectral resolution (e.g., sensors capturing only visible light compared to those including thermal or multispectral bands) can significantly affect feature representation and information content. Second, environmental condition variations introduce additional challenges. Factors such as lighting conditions (e.g., the angle of sunlight), atmospheric effects (e.g., haze, cloud cover, or aerosol scattering), and other environmental dynamics can alter the appearance of the imagery and cause inconsistencies across datasets. Third, geographical and temporal contexts further contribute to domain shifts. The same land cover type, such as forests or urban areas, may exhibit substantial differences across regions (e.g., tropical versus temperate forests) or over time due to seasonal changes or long-term land use transformations. These variations underscore the dynamic and heterogeneous nature of remote sensing data, making domain shifts a critical consideration for practical applications.

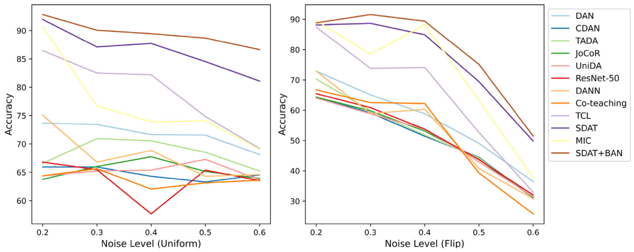

Fig. 8. Classification accuracy (%) on $\mathbf{A} \rightarrow \mathbf{U}$ transfer task with different noise levels. (Left) Uniform noise. (Right) Flip noise.

Hence, directly applying standard DA approaches will cause the model to learn the erroneous information caused by noisy labels in the source domain, thereby significantly reducing transferability and resulting in negative transfer effects. Furthermore, identifying and correcting incorrect labels during the collection of a large number of source domain samples is relatively time-consuming and labor-intensive. Therefore, our proposed BAN is the first attempt at remote sensing cross-scene classification under the setting of NDA. BAN outperforms standard DA methods, indicating that BAN mitigates the negative influence brought by noisy labels. Moreover, in comparison to existing state-of-the-art NDA methods, BAN still demonstrates superior performance, suggesting the superiority of our method and its great potential in practical applications.

D. Limitations and Future Works

Although BAN has demonstrated superior performance for transfer tasks with noisy labels, there are certain limitations that could be addressed for further enhancement. From the perspective of tasks, BAN can serve as a universal paradigm for existing DA and NDA approaches not only for image classification, but for other tasks (such as detection and segmentation) with other kinds of data (such as multispectral data and 3-D point data). From the perspective of methodology, the way utilizing all information from pseudo labels may implicitly transfer negative knowledge, potentially leading to biased DA for NDA. In the future, we will selectively emphasize reliable information from pseudo labels during the training process and make more improvements for NDA scenarios. From the perspective of applications, BAN has only been tested on our collected datasets. Though it has reached remarkable performance, these experiments were conducted under relatively ideal conditions. However, in more practical applications, acquired remote sensing images may vary in resolution, format, or quality. To this end, we will further exploit our BAN for more real-world scenarios.

VII. Conclusion

We propose a BAN to address the challenge of NDA for remote sensing cross-scene classification. BAN is designed with two main parts: FL and BL. FL utilizes a model learning from the noisy source domain and transfers knowledge to target domain. BL utilizes a dual model to acquire knowledge from the target domain and transfer it to source domain. Then, we conduct the FL and BL process alternately to finish the whole training process. By this means, the model is able to leverage supervision information and bilateral information between two domains, thereby reducing the adverse impact of noisy labels. BAN can serve as a universal paradigm for existing DA and NDA methods to boost their performance. To evaluate BAN, we conduct comprehensive experiments on our collected datasets. BAN improves existing NDA approaches with considerable improvements in average accuracy ranging from 6.70% to 15.70% on RGB datasets and OA from 1.36% to 3.14% on hyperspectral datasets with flip-20% noise. Experimental results demonstrate promising prospects of BAN in tackling more practical and general NDA scenes. We will manage to explore the potential of BAN in future work to address more practical NDA tasks in remote sensing scenarios.

References

[1] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[2] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[3] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[4] J. Zheng, S. Yuan, W. Li, H. Fu, L. Yu, and J. Huang, "A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future," *IEEE Geosci. Remote Sens. Mag.*, early access, Nov. 12, 2025, doi: 10.1109/MGRS.2024.3479871. [Online]. Available: https://ieeexplore.ieee.org/document/10750500

[5] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.

[6] S. Zhu, C. Wu, B. Du, and L. Zhang, "Robust remote sensing image cross-scene classification under noisy environment," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600411.

[7] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 97–105.

[8] J. Zheng et al., "Unsupervised mixed multi-target domain adaptation for remote sensing images classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 1381–1384.

[9] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2016.

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

[11] A. K. Menon, B. V. Rooyen, C. S. Ong, and B. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *Proc. ICML*, Jul. 2015, pp. 125–134.

[12] J. Han, P. Luo, and X. Wang, "Deep self-learning from noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5138–5147.

[13] Z. Fang, J. Lü, A. Liu, F. Liu, and G. Zhang, "Learning bounds for open-set learning," in *Proc. ICML*, Jan. 2021, pp. 3122–3132.

[14] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proc. ICLR*, Apr. 2017, pp. 1–9.

[15] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1944–1952.

[16] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. NeurIPS*, vol. 31, Dec. 2018, pp. 8792–8802.

[17] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "L_DMI: A novel information-theoretic loss function for training deep nets robust to label noise," in *Proc. NeurIPS*, vol. 32, Jan. 2019, pp. 6222–6233.

[18] D. Ortego et al., "Multi-objective interpolation training for robustness to label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6606–6615.

[19] E. Zheltonozhskii, C. Baskin, A. Mendelson, A. M. Bronstein, and O. Litany, "Contrast to divide: Self-supervised pre-training for learning with noisy labels," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 387–397.

[20] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 316–325.

[21] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.

[22] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.

[23] Y. Shu, Z. Cao, M. Long, and J. Wang, "Transferable curriculum for weakly-supervised domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 4951–4958.

[24] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. NeurIPS*, vol. 31, Jan. 2018, pp. 8527–8537.

[25] M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *Proc. ICML*, vol. 1, Jul. 2020, pp. 6448–6458.

[26] H. Wei, L. Tao, R. Xie, and B. An, "Open-set label noise can improve robustness against inherent label noise," in *Proc. NeurIPS*, Jan. 2021, pp. 7978–7992.

[27] R. Xie, H. Wei, L. Feng, and B. An, "GearNet: Stepwise dual learning for weakly supervised domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 8717–8725.

[28] X. Yu, T. Liu, M. Gong, K. Zhang, K. Batmanghelich, and D. Tao, "Label-noise robust domain adaptation," in *Proc. ICML*, vol. 119, Jul. 2020, pp. 10913–10924.

[29] J. Zheng et al., "Growing status observation for oil palm trees using unmanned aerial vehicle (UAV) images," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 95–121, Mar. 2021.

[30] J. Zheng et al., "Surveying coconut trees using high-resolution satellite imagery in remote atolls of the Pacific ocean," *Remote Sens. Environ.*, vol. 287, Mar. 2023, Art. no. 113485.

[31] D. Tuia, D. Marcos, and G. Camps-Valls, "Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization," *ISPRS J. Photogramm. Remote Sens.*, vol. 120, pp. 1–12, Oct. 2016.

[32] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 155, pp. 72–89, Sep. 2019.

[33] G. Mateo-García, V. Laparra, D. López-Puigdollers, and L. Gómez-Chova, "Transferring deep learning models for cloud detection between Landsat-8 and Proba-V," *ISPRS J. Photogramm. Remote Sens.*, vol. 160, pp. 1–17, Feb. 2020.

[34] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4673–4688, Jun. 2021.

[35] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[36] M. Luo and S. Ji, "Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 105–128, Sep. 2022.

[37] Y. Cai, Y. Yang, Y. Shang, Z. Chen, Z. Shen, and J. Yin, "IterDANet: Iterative intra-domain adaptation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629517.

[38] C. Liang, B. Cheng, B. Xiao, Y. Dong, and J. Chen, "Multilevel heterogeneous domain adaptation method for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601916.

[39] E. Alvarez-Vanhard, T. Houet, C. Mony, L. Lecoq, and T. Corpetti, "Can UAVs fill the gap between in situ surveys and satellites for habitat mapping?" *Remote Sens. Environ.*, vol. 243, Jun. 2020, Art. no. 111780.

[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, May 2021.

[41] Y. Hamrouni, E. Paillassa, V. Chéret, C. Monteil, and D. Sheeren, "From local to global: A transfer learning-based approach for mapping poplar plantations at national scale using Sentinel-2," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 76–100, Jan. 2021.

[42] Q. Li, Y. Wen, J. Zheng, Y. Zhang, and H. Fu, "HyUniDA: Breaking label set constraints for universal domain adaptation in cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5518415.

[43] J. Guo et al., "C^{3}DA: A universal domain adaptation method for scene classification from remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[44] X. Li, M. Luo, S. Ji, L. Zhang, and M. Lu, "Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection," *Int. J. Remote Sens.*, vol. 41, no. 19, pp. 7343–7367, Oct. 2020.

[45] Y. Koga, H. Miyazaki, and R. Shibasaki, "A method for vehicle detection in high-resolution satellite images that uses a region-based object detector and unsupervised domain adaptation," *Remote Sens.*, vol. 12, no. 3, p. 575, Feb. 2020.

[46] J. Zheng et al., "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, Sep. 2020.

[47] J. Zheng, W. Wu, S. Yuan, H. Fu, W. Li, and L. Yu, "Multisource-domain generalization-based oil palm tree detection using very-high-resolution (VHR) satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[48] C. Geiß, H. Schrade, P. A. Pelizari, and H. Taubenböck, "Multistrategy ensemble regression for mapping of built-up density and height with Sentinel-2 data," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 57–71, Dec. 2020.

[49] H. Zuo, J. Lu, and G. Zhang, "Multiple-source domain adaptation in rule-based neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[50] J. Zheng, Y. Zhao, W. Wu, M. Chen, W. Li, and H. Fu, "Partial domain adaptation for scene classification from remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601317.

[51] A. Elshamli, G. W. Taylor, and S. Areibi, "Multisource domain adaptation for remote sensing using deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3328–3340, May 2020.

[52] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, Apr. 2020.

[53] J. Zheng et al., "A two-stage adaptation network (TSAN) for remote sensing scene classification in single-source-mixed-multiple-target domain adaptation (S^2M^2T DA) scenarios," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609213.

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

[55] R. Adayel, Y. Bazi, H. Alhichri, and N. Alajlan, "Deep open-set domain adaptation for cross-scene classification based on adversarial learning and Pareto ranking," *Remote Sens.*, vol. 12, no. 11, p. 1716, May 2020.

[56] J. Zheng et al., "Open-set domain adaptation for scene classification using multi-adversarial learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 245–260, Feb. 2024.

[57] S. Tan, J. Jiao, and W.-S. Zheng, "Weakly supervised open-set domain adaptation by dual-domain collaboration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5389–5398.

[58] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

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

[60] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

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

[62] 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.

[63] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13723–13732.

[64] Q. Xu, Y. Shi, X. Yuan, and X. X. Zhu, "Universal domain adaptation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4700515.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[66] H. Rangwani, S. K. Aithal, M. Mishra, A. Jain, and V. B. Radhakrishnan, "A closer look at smoothness in domain adversarial training," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 18378–18399.

[67] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11721–11732.

[68] B. Le Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.

[69] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[70] Z. Zhang, H. Zhang, S. Ö. Arik, H. Lee, and T. Pfister, "Distilling effective supervision from severe label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9291–9300.

[71] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5345–5352.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[73] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.* Cham, Switzerland: Springer, Paris, France, 2010, pp. 177–186.