RFA-Net: Reconstructed Feature Alignment Network for Domain Adaptation Object Detection in Remote Sensing Imagery

Yangguang Zhu, Xian Sun, *Senior Member, IEEE*, Wenhui Diao, *Member, IEEE*, Hao Li, *Member, IEEE*, and Kun Fu, *Member, IEEE*

***Abstract*—With the development of deep learning, great progress has been made in object detection of remote sensing (RS) imagery. However, the object detector is hard to generalize well from one labeled dataset (source domain) to another unlabeled dataset (target domain) due to the discrepancy of data distribution (domain shift). Currently, adversarial-based domain adaptation methods align the semantic features of source and target domain features to alleviate the domain shift. But they fail to avoid the alignment of noisy background features and neglect the instance-level features, which are inappropriate for detection models that focus on instance location and classification. To mitigate domain shift existing in object detection, we propose a reconstructed feature alignment network (RFA-Net) for unsupervised cross-domain object detection in RS imagery. The RFA-Net includes one sequential data augmentation module deployed on data level for providing solid gains on unlabeled data, one sparse feature reconstruction module deployed on feature level to intensify instance feature for feature alignment, and one pseudo-label generation module deployed on label level for the supervision of the unlabeled target domain. Extensive experiments illustrate that our proposed RFA-Net is effective to alleviate the domain shift problem in domain adaptation object detection of RS imagery.**

***Index Terms*—Data augmentation, domain adaptation, feature reconstruction, object detection, pseudo-label filtering.**

I. Introduction

NOWADAYS, object detection is essential in remote sensing (RS) imagery interpretation and also has a widespread application in natural resource management, intelligent agriculture, building detection, etc. In the past decade, with the development of deep learning algorithms [3], [4], great progress has been made by deep convolutional neural networks (DCNNs) in RS object detection [5]–[13]. However, the outstanding performance highly relies on large quantities of training data with annotations. Specifically, the background and scale of the RS imagery vary greatly and are complicated. These render the annotation of RS imagery more computationally expensive and time-consuming. The model trained on a small number of the labeled dataset (source domain) will generalize poorly on a large number of the unlabeled dataset (target domain). Moreover, due to the diversity of RS imagery acquisition conditions, including varied geospatial regions, weather conditions, ground sampling distances, and arbitrary shooting angles [14], the discrepancy of data distribution between two datasets, treated as the domain shift, is inevitable. And, it disastrously restraints the improvement of model generalization further. As shown in Fig. 1, the discrepancy between the frequency domain histograms of the source and target domains is significant.

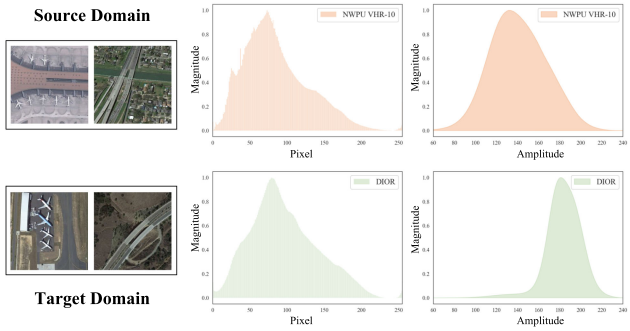


Fig. 1. Difference between the NWPU VHR-10 [1] (source domain) and DIOR [2] (target domain) datasets. From left- to right-hand side: Original image, the image histogram after normalization, and the frequency domain magnitude histogram after normalization.

To improve the generalization performance of deep learning models on the target domain, one naive way is to train the model on the source domain and fine-tune the model on the target domain. Such an approach requires manual annotations of the data in the target domain, which is computationally expensive and time-consuming as described previously. Therefore, an effective method is to transfer the knowledge learned on the source domain data into the unlabeled target domain in an

Manuscript received 16 March 2022; revised 17 May 2022 and 1 July 2022; accepted 10 July 2022. Date of publication 13 July 2022; date of current version 22 July 2022. This work was supported by the National Natural Science Foundation of China under Grant 61725105. *(Corresponding author: Xian Sun.)*

Yangguang Zhu, Xian Sun, and Kun Fu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, with the Key Laboratory of Network Information System Technology (NIST) and Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhuyangguang19@mails.ucas.ac.cn; sunxian@aircas.ac.cn; fukun@mail.ie.ac.cn).

Wenhui Diao and Hao Li are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Key Laboratory of Network Information System Technology (NIST) and Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: diaowh@aircas.ac.cn; lihaoaircas@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3190699

unsupervised way, which means the data in the target domain will be utilized for training in an unsupervised manner without ground truth. Following this idea, the unsupervised domain adaptation methods [15]–[23] have appeared and been widely utilized in classification and segmentation tasks of RS imagery. These methods are devoted to alleviating the discrepancy by utilizing the semantic feature alignment between the source and target domains. In the initial application of classification and segmentation in RS imagery, the most popular approaches [15]–[19] project features in the source and target domains into a subspace and design a metric loss to minimize the discrepancy. Maximum mean discrepancy (MMD) [24] has been used for preserving the main statistical property in domains and minimizing the distance of the distribution between the source and target domains. However, the specific distance metric varies between different domains and requires to be designed manually. As the gradient reversal layer (GRL) [25] has been proposed, the adversarial-based domain adaptation methods appear and have been deployed in domain adaptation, classification, and segmentation in RS imagery. The adversarial-based domain adaptation methods mainly utilize a classifier treated as a domain discriminator and achieve domain confusion between the source and target domains based on adversarial training.

While many domain adaptation algorithms have been widely applied in the classification and segmentation of RS imagery, to the best of our knowledge, there are few algorithms specifically designed for multicategories object detection in RS imagery. Xu *et al.* [26] proposed the FADA with a single-stage detector to align the cross-domain features. Besides, there are also some methods [27]–[29] exploring the domain adaptation object detection, which only focus on the detection of only one category. For example, Chen *et al.* [27] proposed the rotation-invariant and relation-aware cross-domain adaptation object detection network based on a relation-aware graph to align the feature distribution and a rotation-invariant regularizer to deal with the rotation diversity. However, all of them conduct the feature alignment directly, as most algorithms in classification and segmentation do without considering the redundant features in the RS imagery. Specifically, different from classification or segmentation of RS imagery, which mainly deals with semantic-level features to conduct a classification of image or pixel, object detection focuses more on local instance-level features for regressing bounding box and object classification. Therefore, domain adaptation object detection in the RS imagery requires the design of specialized algorithms for alleviating the domain shift between the source and target domains. To relieve the limitations of the domain adaptation algorithm with the methods in the classification and segmentation of RS imagery, we propose a reconstructed feature alignment network (RFA-Net) for domain adaptation object detection in RS imagery. The RFA-Net improves the performance from one dataset with a small quantity of labeled data to one dataset with a large quantity of unlabeled data. The RFA-Net shown in Fig. 3 includes one sequential data augmentation (SDA) module, one sparse feature reconstruction (SFR) module, and one pseudo-label generation (PLG) module. All of these modules have been deployed on multistages, respectively.

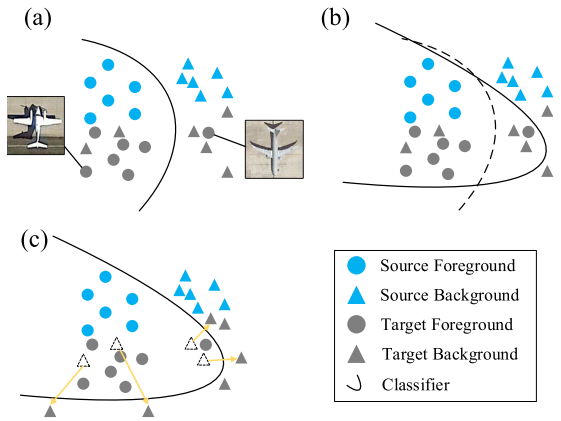


Fig. 2. Models' performance visualization with the feature space representation. (a) Model is trained with the source domain data. It can perform well in the source domain data, while some foregrounds and backgrounds in the target domain data are misclassified. (b) Adversarial-based domain adaptation methods of aligning the semantic features between different domains directly. Rough alignment tends to suffer from misclassification between some foregrounds and background. (c) Our proposed method RFA-Net rectifies the misclassification.

In the data preprocessing stage, it is difficult for the model trained on the source domain to perform well on the target domain, since the domain shift between different domains. Moreover, the volume gap in data quantity exacerbates the difficulties. Therefore, we propose an SDA module, which not only expands the data volume, but improves the robustness of the model trained with the target data with noisy labels [30]. The label generation for target domain data is described in Section III-D. In the feature extraction stage, excessive semantic alignment in the popular approaches of classification and segmentation leads to the alignment of noisy features and implicitly weakens instance-level features. Inspired by that, the matrix reduction [31], [32] can denoise the redundant features utilizing low-rank matrix reconstruction, the SFR module is proposed to design a selection of representative features with feature reconstruction. The SFR reconstructs the low-level features not only to denoise the redundant features, but to implicitly intensify the specific instance features for alignment in their respective domains. In the label supervision stage of the detection head, only the source domain can provide label supervision for the output, since the target domain data are unlabeled. The model tends to learn biased information in the source domain, which does harm to the generalization ability. To expand the number of labeled data, we propose one PLG module to generate the labels for the unlabeled target domain by exploiting the knowledge learned from the source domain. The high-quality labels are used as the pseudo labels for the unlabeled target domain, implicitly conducting information alignment between different domains in the detection head.

Fig. 2. shows the visualization of the detection model's performance in RS imagery. The detection model only trained on the labeled source domain data is difficult to perform well on the unlabeled target domain data. Meanwhile, the adversarial-based domain adaptation methods, with aligning the semantic features between different domains directly, can only perform slightly better in the target domain compared to the source-only detection

model. There are still false alarms and objects missing during the testing phase because of the excessive alignment. Our proposed method can effectively rectify the false alignment and solve these problems with the consideration of the characteristics of object detection in RS imagery. And, extensive experiments prove the effectiveness of our proposed RFA-Net. The main contributions of this article are described as follows.

1) We argue that the domain shift exists in RS imagery object detection and propose an RFA-Net to alleviate the domain shift. The RFA-Net includes one SDA module, one SFR module, and one PLG module.
2) The SDA module is deployed on data preprocessing at the data level to expand the data volume and improve the robustness, the SFR module is deployed on feature extraction at feature level to intensify instance feature and feature alignment, and the PLG module is deployed on label supervision in label level as the pseudo labels of the unlabeled target domain.
3) We achieve a significant improvement in the experimental results with the increase of a small number of parameters, and all of the modules are not introduced in the inference phase to avoid the reduction of the inference speed.

The rest of this article is organized as follows. In Section II, we briefly discuss some related methods. In Section III, we describe our implementation in detail. In Section IV, we explained the experiments and results. Finally, Section V concludes this article.

II. Related Work

In this section, we provide a brief overview of some related work, which covers object detection, domain adaptation in RS imagery, and some comparable domain adaptation object detection methods in natural scene images.

A. Object Detection

With the progress of DCNNs [3], [4], [33], the DCNNs-based object detection methods [34]–[38] can automatically extract the input image features. They detect the objects with higher accuracy and more robustly than the previous methods that use manual methods [39], [40] for feature extraction. In the field of RS, many object detection methods [6], [7], [9]–[13], [41] have followed the DCNNs-based methods and achieved excellent performance. Cheng *et al.* [9] proposed a new rotation-invariant layer, named rotation-invariant CNN model, to lift the performance of object detection in RS imagery. Yang *et al.* [7] tended to solve the problem caused by the narrow width of the ship and proposed a rotation dense feature pyramid networks framework compared with the feature pyramid network [42]. Fu *et al.* [41] constructed a unified framework for arbitrary-oriented and multiscale object detection in RS imagery by combining features with different levels. The model can get a robust feature representation and augment the anchors with multiple default angles to get powerful performance on multiscale-oriented objects. Lin *et al.* [10] employed a novel geometric transformation to represent the oriented object in angle prediction and an enhanced intersection over union (IOU) loss for oriented bounding boxes (OBBs) detection. The detection model predicts the OBBs in a per-pixel fashion. Shi *et al.* [12] designed the centerness-aware network with a new centerness-aware model, which can utilize the symmetrical shape of objects in RS imagery. However, the object detection methods rely on the test dataset with similar data distribution of the train dataset, and these methods do not consider the domain shift and fail to perform well on an unlabeled dataset with discrepancy.

B. Domain Adaptation

As for a source domain with labeled training data and a target domain with an inconsistent distribution, domain adaptation concentrates on how to design a model that generalizes well on the unlabeled target domain data [43], [44]. General approaches [15]–[19], [45]–[48] are proposed to formalize the domain gap and minimize it. Ghifary *et al.* [45] deployed the MMD metric in the supervised domain adaptation to diminish the mismatch in the features subspace between the cross-domains. Since the GRL [25] has been proposed, the adversarial-based methods [20]–[23], [49]–[52] have become increasingly popular. Zhu *et al.* [20] designed a semisupervised center-based discriminative adversarial learning framework for cross-domain classification, which is based on filtering out easy triplets, proposed hard triplet loss, and the adversarial learning with center loss. Yan *et al.* [21] proposed to distinguish two segmentation maps in the same domain from the two maps in the different domains with a triplet adversarial domain adaptation method based on adversarial training. The algorithm explicitly narrows the distribution gap across domains with the consideration of both domains' information. Iqbal and Ali [22] proposed an adaptive method with a strategy of weakly supervision, where they introduced the image-level labels for the unlabeled target domain data. Li *et al.* [23] proposed a semantic segmentation network utilizing a novel objective function with multiple weakly supervised constraints. They introduced the DualGAN to employ unsupervised style transfer between the cross-domains. However, these methods were proposed for classification and segmentation of RS imagery excessively align semantic features. They are not suitable to be implemented in object detection. Object detection focuses more on local instance-level features for bounding box regression and object classification. Therefore, domain adaptation object detection algorithms of RS imagery should be explored specifically.

C. Domain Adaptation for Object Detection

Object detection domain adaptation methods have been wildly explored in the field of computer vision because of their importance in the wild. The object detection domain adaptation methods can be mainly classified as discrepancy-, adversarial-, and reconstruction-based methods. Discrepancy-based methods [53]–[55] mitigate the domain shift by refining the network with labeled and unlabeled data in the target domain. Cai *et al.* [54] focused on cross-domain object detection from synthetic images to real images with a mean teacher paradigm. Cao *et al.* [55] presented an autoannotation framework and

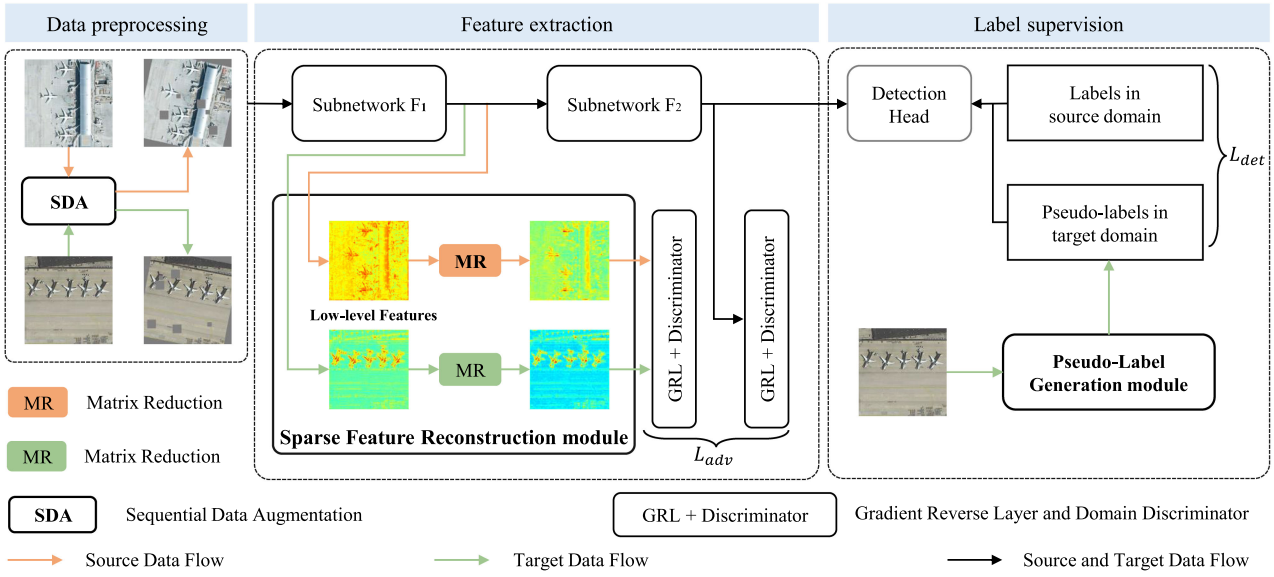


Fig. 3. Architecture of our RFA-Net. The SDA module is deployed on both image and instance levels to expand the data volume and improve the robustness of the model. The SFR sends the source and target domains low-level features to their respective corresponding flows and deploys matrix reduction for feature reconstruction. The PLG is utilized to generate the labels for the unlabeled target domain by exploiting the knowledge learned from the source domain data, lifting the generalization ability of the detection model.

used a region proposal network with two-streams, which extracted the multispectral features for robust pedestrian detection. Adversarial-based methods [56]–[61] conduct adversarial training with GRL, which is accompanied by the domain discriminator to expect domain confusion between the cross-domains. Chen *et al.* [56] proposed the domain adaptive faster RCNN to tackle the domain shift raised from the image and instance levels, which was the first work to carry out the domain adaptation object detection problem. Saito *et al.* [58] proposed an unsupervised domain adaptation object detection method that considers the global and local features alignment, respectively. Following this method, many incremental domain adaptation object detection methods [59]–[61] have been proposed. Reconstruction-based methods [62], [63] employ the image reconstruction in the source or target domains to enhance the performance of the detection model. Arruda *et al.* [62] explored the CycleGAN to generate a synthetic dataset by translating data style from the daytime domain to the nighttime domain. The synthetic dataset is then used to train the detection model, which achieves significant improvements. Although these methods have achieved promising performance on natural images, they are still difficult to solve the domain adaptation object detection in RS imagery, since the structure of RS imagery is more diverse and complex. Some methods [27]–[29] were proposed that only focus on the detection of only one category. Xu *et al.* [26] proposed the FADA with a single-stage detector to align the cross-domain features, which also conducts on detection of a few categories (e.g., 1–3). However, all of them are not suitable for the detection of complex RS scenes with multicategories. We propose an object detection domain adaptation method specifically for RS imagery considering the characteristics of the RS imagery. Extensive experiments are conducted to demonstrate that our approach enables excellent performance in RS.

III. Domain Adaptation for RS Imagery

In this section, we elaborate on the RFA-Net. Fig. 3 shows the overview of the structure of our proposed RFA-Net.

A. Preliminaries

For the source domain images $\{X^s, Y^s\}$ and the target domain images $\{X^t\}$ without labels, we take an image x^s and ground truth y^s in source domain, as well as an image x^t without labels in target domain. Specially, the popular adversarial-based domain adaptation methods [58], [59], [61] divide the backbone into subnetworks F_1 and F_2. The subnetwork F_1 extracts the low-level features P_1^s and P_1^t of x^s and x^t for local features alignment, respectively. Subsequently, F_2 takes P_1^s and P_1^t to extract the high-level features P_2^s and P_2^t for global features alignment. Specifically, for the input image x, F_1 extracts the low-level features with a shape of $C \times H \times W$ from x. The domain discriminator D_1 is used to predict a domain probability map for low-level features with a shape of $H \times W$. Each pixel is used to estimate the probability of the domain label d, which is 0 if the input features are from the source domain, otherwise 1. The loss function of low-level feature alignment can be formulated as follows:

$$L_{\text{low}}^s = -\frac{1}{n_s HW} \sum_{i=1}^{n_s} \sum_{w=1}^{W} \sum_{h=1}^{H} D_1 \left(F_1\left(x_i^s\right)\right)_{wh}^2 \tag{1}$$

$$L_{\text{low}}^t = -\frac{1}{n_t HW} \sum_{i=1}^{n_t} \sum_{w=1}^{W} \sum_{h=1}^{H} \left(1 - D_1\left(F_1\left(x_i^t\right)\right)_{wh}\right)^2 \tag{2}$$

$$L_{\text{low}}\left(F, D_1\right) = \frac{1}{2}\left(L_{\text{low}}^s + L_{\text{low}}^t\right). \tag{3}$$

Similarly, F_2 extracts the high-level features of the input image x, and the domain discriminator D_2 is a binary classifier used to predict a domain label of high-level features for alignment. The loss function of high-level feature alignment can be formulated as follows:

$$L_{\text{high}}^{s} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\text{FL}\left(D_2\left(F\left(x_i^s\right)\right)\right) \tag{4}$$

$$L_{\text{high}}^{t} = -\frac{1}{n_t}\sum_{i=1}^{n_t}\text{FL}\left(D_2\left(F\left(x_i^t\right)\right)\right) \tag{5}$$

$$L_{\text{high}}\left(F, D_2\right) = \frac{1}{2}\left(L_{\text{high}}^{s} + L_{\text{high}}^{t}\right) \tag{6}$$

where $\text{FL} = (1-p_t)^{\gamma}\log(p_t)$ is the focal loss [36] and p_t is $1-p$ if domain label is 0, otherwise p. Therefore, the adversarial loss in (1) can be formulated as follows:

$$L_{\text{adv}} = L_{\text{low}}\left(F, D_1\right) + L_{\text{high}}\left(F, D_2\right). \tag{7}$$

Finally, only P_2^s extracted from source domain images is used as input to the detection head R. The detection head R consists of the RPN and ROI heads in faster RCNN [34], and outputs the bounding boxes with the specific class labels. To perform the domain adaptation on the object detection, the overall objective function is summarized as follows:

$$L\left(X^s, Y^s, X^t\right) = L_{\text{det}}\left(X^s, Y^s\right) - \lambda L_{\text{adv}}\left(X^s, X^t\right) \tag{8}$$

where L_{det} is the detection loss, L_{adv} is the adversarial loss to minimize the domain shift between the different domains, and λ is the weight to balance the L_{det} and L_{adv}.

Based on such an adversarial-based approach, we propose the RFA-Net, which is specially designed for domain adaptation object detection of RS imagery. The overall workflow of RFA-Net is shown in Fig. 3, which is composed of one SDA module, one SFR module, and one PLG module. The detailed formulation is described in the following sections.

B. SDA Module

In the RS imagery, the object position angle varies greatly and the discrepancy of pixel distribution exists. Moreover, while a gap exists in the amount of data between different domains, it is difficult for the model trained on the source domain to perform well on the target domain. Therefore, the input data are viewed as an important factor needs to be promoted in the data preprocessing. We proposed an SDA module to conduct the augmentation on both image and instance levels. The SDA not only considers the volume gap in data quantity, but improves the robustness of the model when it is trained with the target data with noisy labels [30]. At first, we simply flip the labeled data in the source domain to expand the data quantity. After that, the image- and instance-level data transformations are combined in a sequential way for the input images. The sequential data transformation is viewed as a strong data transformation to deteriorate the data distribution discrepancy and to execute augmentation at image and instance levels. In addition, we believe that SDA can improve the robustness of adversarial training as proposed by Rebuffi *et al.* [64].

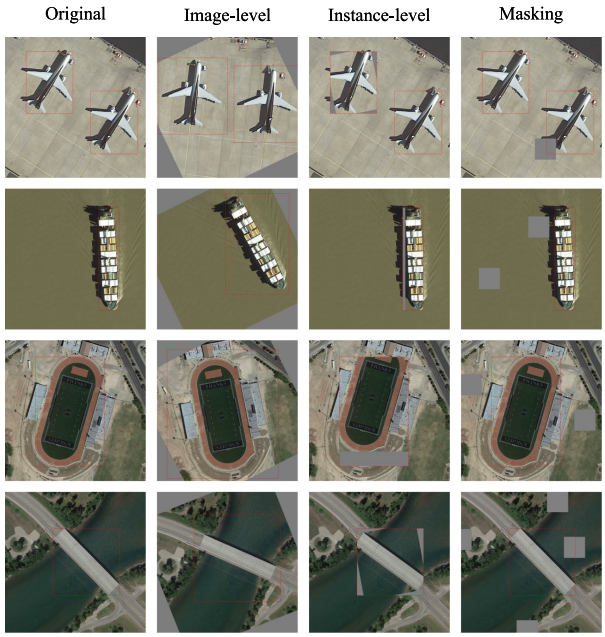


Fig. 4. Visualization of sequential data transformation with different strategies. From left- to right-hand side: Original image, image-level data transformation, instance-level data transformation, and masking.

As for the image-level data transformation, global geometric transformation [65] is applied in the input images, which includes *x*–*y* translation, rotation, and shear. As for the instance-level data transformation, we simply deploy the global geometric transformation in the bounding boxes of the images randomly. The transformation at instance-level has smaller magnitude ranges compared with that at image level. Especially, as for the unlabeled target domain images, we apply the instance-level data transformation on the pseudo labels generated by PLG (as described in Section III-D). Furthermore, the complex structure of RS imagery and redundant information of the background exist in each domain. Such redundant information distributes the detection model to extract efficient features for bounding box regression and object classification. We randomly mask the input images with a low rate to compress the input images for alleviating the domain shift caused by the redundant information in each domain. For the abovementioned data transformations, we sequentially construct them and view such mixed transformation as a strong data augmentation [30].

In the training phase, we equip the SDA module for the input images in the source and target domains. All the data transformations of SDA are shown, respectively, in Fig. 4. Unlike the traditional data transformation, which selects the certain data transformation by a probability randomly, we apply the abovementioned data transformations to the input data with a probability of 1 in turn. We visualize the adopted data transformation methods in Fig. 4.

C. SFR Module

The adversarial-based object detection domain adaptation methods align the semantic features directly to alleviate the

domain shift. Specifically, The subnetworks F_1 and F_2 extract low- and high-level features of the input image, respectively. The low-level features are deployed on strong local alignment, and the high-level features are deployed on weak global alignment. Such alignment intends to align the distribution of the features in the source and target domains and alleviates the domain shift. The paradigm is also a general method adopted by domain adaptation classification and segmentation in RS imagery. But it is not appropriate for domain adaptation object detection in RS imagery, as the structure of RS imagery is complicated. We argue that the low-level features extracted from subnetwork F_1 are strongly aligned pixel-by-pixel, ignoring the fact that complex background features exist in RS imagery.

To align the low-level features effectively, the proposed SFR module reconstructs the low-level features before deploying strong alignment. Specifically, the matrix reduction, as described in the method proposed by Geng *et al.* [31], projects it into a low-dimensional space, thus removing the interference of background in the features. For the low-level feature $F_{ll} = F_1(x)$ $(F_{ll} \in R^{(C\times H\times W)})$ extracted from the low-level network, we deploy matrix reduction on F_{ll}. Regarding the low-level feature F_{ll} extracted from the subnetwork F_1, we first flatten it by channel to obtain F_{ll}^f with the shape of $C \times HW$. F_{ll}^f can be represented by a dictionary matrix $D_{\text{dict}} = [d_1, \ldots, d_K] \in R^{C\times K}$ and corresponding codes $C = [c_1, \ldots, c_{HW}] \in R^{K\times HW}$ in the following way:

$$F_{ll}^f = \bar{F_{ll}^f} + E = D_{\text{dict}}C + E \tag{9}$$

where $\bar{F_{ll}^f} \in R^{C\times HW}$ is the low-rank reconstruction of F_{ll}^f and E is the noisy matrix, which we want to ignore in the alignment.

We can specify a very small value K to restrict the ranks of D_{dict} and C, because the rank in matrix product has the following property:

$$\begin{aligned}\text{rank}\left(\bar{F_{ll}^f}\right) &\le \min\left(\text{rank}\left(D_{\text{dict}}\right), \text{rank}(C)\right)\\ &\le K \ll \min(C, HW).\end{aligned} \tag{10}$$

Hence a low-rank matrix $\bar{F_{ll}^f}$ is obtained to represent F_{ll}^f with minimizing the SFR loss L_{SFR}, which is summarized as

$$L_{\text{SFR}}\left(D_{\text{dict}}, C\right) = L\left(D_{\text{dict}}, C\right) + R_1\left(D_{\text{dict}}\right) + R_2(C) \tag{11}$$

where $L(D_{\text{dict}}, C) = \|F_{ll}^f - D_{\text{dict}}C\|_F$, $\|\cdot\|_F$ in $L(D_{\text{dict}}, C)$ is the Frobenius norm, and R_1 and R_2 are the regularization terms for the dictionary matrix D and the codes C, respectively. We reshape the $\bar{F_{ll}^f}$ to represent the low-rank reconstruction of input feature F_{ll}.

Shekhar *et al.* [32] mapped features into a low-dimensional feature subspace with the implementation of sharing the dictionary matrix D_{dict} to create different projections for the features in the cross-domains, so that the source and target domains can retain specific semantic features in their respective domain. Our proposed SFR allocates split paces for the source and target domains features, respectively. The low-level feature alignment is deployed on the reconstructed low-dimensional features of the source and target domains from different paths subsequently. The SFR module with low-level alignment is shown in Fig. 3.

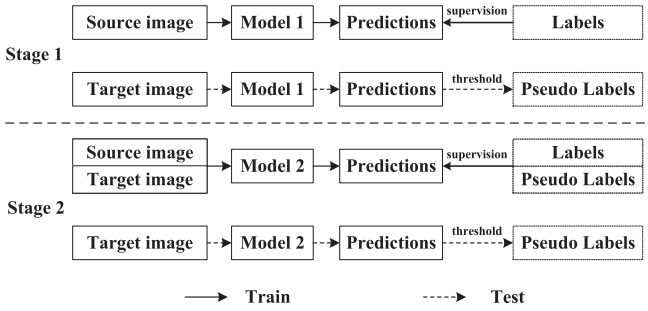


Fig. 5. PLG. In stage 1, we train a detection model with the labeled source data and inference the unlabeled target data for label supervision generation. We also set a threshold to filter out the predictions whose confidence scores are below the threshold during the inference phase. After that, we train a new model with labeled source data and target data with pseudo labels, and the new model predicts the target domain data with the same threshold for final label supervision generation.

D. PLG Module

In the training phase of adversarial-based object detection domain adaptation, knowledge learned on the source domain data is transferred into the unlabeled target domain in an unsupervised way. Only a small number of labeled data in the source domain are utilized for supervision, which renders the detection model to learn biased information in the source domain. The benefits of the unlabeled data, proposed by Carmon *et al.* [66], demonstrate that the unlabeled data can improve the robustness of adversarial training. Therefore, we propose the PLG module to generate the labels for the unlabeled target domain. As shown in Fig. 5, PLG generates the labels for the unlabeled data utilizing self-training with multistages.

As for the PLG of the target domain, for stage 1, we simply train the first detection model with source domain data until it converges. Such a model inferences the unlabeled target domain data to get the predictions. The pseudo labels in stage 1 are generated by filtering the predictions with a threshold τ. After that, the labeled source domain data and the unlabeled target domain data with pseudo labels are merged for training the second detection model from scratch in stage 2. And, the second detection model inferences the unlabeled target domain data to generate the final pseudo labels for domain adaptation object detection in RS imagery, in which the threshold τ is the same as that in stage 1. Specifically, we believe that some noisy labels also exist in the pseudo labels inevitably even though the threshold is applied. As for the second detection model, the labels in the source domain are the supervisions for the RPN and ROI heads, while the pseudo labels in the target domain are only for the RPN. The pseudo labels in the target domain are generated with self-training and only these in stage 2 are utilized as supervision for domain adaptation.

Finally, for the ground truth of the source domain and the pseudo labels generated by PLG of the target domain, we consider them as supervision during training. As we believed above, the pseudo labels of unlabeled target data may still accompany by noisy labels inevitably. Therefore, the pseudo labels in the target domain are only utilized as supervision of RPN, expecting the RPN to produce higher quality proposals than before. The

objective function of detection in (8) is described as follows:

$$L_{\text{det}} = L_{\text{rpn}}(x^s, y^s) + L_{\text{roihead}}(x^s, y^s) + \alpha L_{\text{rpn}}(x^t, y^t_{pl}). \quad (12)$$

The hyperparameter α controls the tradeoff between the total RPN losses of source and target domains.

E. Overall Objective

Considering that the SFR is inserted before the low-level feature alignment, we set the hyperparameter β in adversarial loss to control the tradeoff between the feature alignment losses in low and high levels. The objective function of adversarial losses in (7) is described as follows:

$$L_{\text{adv}} = L_{\text{low}}(F, D_1) + \beta L_{\text{high}}(F, D_2). \quad (13)$$

The total loss consists of the detection loss and the adversarial loss, where the detection loss is described as in (12) and the adversarial loss is described as in (13), so the total loss is described as follows:

$$L_{\text{total}} = L_{\text{det}}(F, D) + L_{MD}(D_{\text{dict}}, C) - \lambda L_{\text{adv}}(F, D, D_{\text{dict}}, C). \quad (14)$$

And, λ is usually set to 1 in the total loss, and α is considered as a hyperparameter in the detection loss as well as β in the adversarial loss.

IV. Experiments and Results

A. Datatsets

1) NWPU VHR-10 Dataset [1]: The dataset collects 800 images, in which 715 high-spatial-resolution color images are from Google Earth and 85 very-high-spatial-resolution pansharpened color infrared (CIR) images are from the Vaihingen dataset [67]. In NWPU VHR-10, the spatial resolution of the images from Google Earth ranges from 0.5 to 2 m and the resolution of the images from CIR is 0.08 m. Out of all 800 images, 650 images containing a total of ten categories of objects are used as a positive image dataset and the other 150 images without any objects are used as the negative image set, and the ratio of the train, validation, and test sets in the positive image set is 14:6:5.

2) DIOR Dataset [2]: The DIOR dataset contains 23 463 optical RS images and 192 472 objects with 20 common object categories. All images in the dataset are got from Google Earth (Google Inc.) with the spatial resolution ranges from 0.5 to 30 m and the size of 800×800 pixels. By contrast to the NWPU VHR-10 dataset, the DIOR dataset has richer image variations in geospatial regions weather conditions, scales, image quality, etc., which contains the RS images covering more than 80 countries. Consequently, the DIOR dataset has more variations for each object class in viewpoint, occlusion, appearance, background, object pose, etc. We draw the same ten categories in the NWPU VHR-10 dataset from the DIOR dataset, formulating the DIOR* dataset with a total of 6997 images and 50 410 objects for RS images object detection domain adaptation.

TABLE I
Statistics of the NWPU VHR-10, DIOR, and HRRSD Datasets With Ten Categories

Dataset	**NWPU VHR-10**	**DIOR***	**HRRSD***
Airplane	757	1712	4901
Baseball diamond	390	2083	4042
Basketball court	159	909	4064
Bridge	124	1122	4651
Ground track field	163	998	4033
Harbor	224	2041	3902
Ship	302	23 319	3975
Storage tank	655	2502	4424
Tennis court	524	4142	4402
Vehicle	598	11 582	4756
Total objects number	3896	50 410	43 168
Total images number	650	6997	17 016

3) HRRSD Dataset [68]: The HRRSD is an RS imagery object detection dataset with multi categories, in which an image may contain one object or multiple objects with multiple categories. This dataset includes 13 categories with a total of 26 722 RS images, of which 21 761 images with a spatial resolution of 0.15–2 m are from Google Earth and another 4961 images with the spatial resolution of 0.6–2 m are from Baidu Maps. An algorithm was designed in [68] for the division of the train, validation, and test sets, which makes the object distribution in HRRSD more balanced compared to the NWPU VHR-10 dataset. We also draw the same ten categories of the NWPU VHR-10 dataset from the HRRSD dataset. These data formulate the HRRSD* dataset with a total of 17 016 images and 43 168 objects for RS images object detection domain adaptation. The distribution of the number of objects in each category in the NWPU VHR-10 and DIOR* datasets is given in Table I.

B. Implementation Details

We employ faster RCNN [34], a two-stage object detection network with RoIAlign, for object detection domain adaptation. ResNet101 [69] is used as the backbone for input image feature extraction. The low-level features extracted from the conv2_x layer in the ResNet101 are used for strong local alignment, and the features extracted from the conv4_x layers are used for weak local alignment. To better illustrate the performance of our proposed method for RS imagery domain adaptation object detection, we also adopt the same settings as Saito *et al.* [58] to implement the method in our experiments. Besides, we also experimented with the backbone of VGG16 [70]. The backbone is initialized with the pretrained model on ImageNet. In the training phase, we train the model with stochastic gradient descent for 20 epochs. The initial learning rate is 0.001 and reduced by ten times every 10 epochs. The weight decay and momentum are configured to 0.0001 and 0.9, respectively. All experiments are conducted on the PyTorch framework and carried out on a Tesla P100 GPU with 16 GB of memory.

We adopt the average precision (AP) to evaluate the performance of the detection model, and the AP can be calculated by

TABLE II
QUANTITATIVE COMPARISON WITH THE SOURCE-ONLY AND OTHER DOAMIN ADAPTATION OBJECT DETECTION METHODS IN THE EXPERIMENTS OF THE NWPU VHR-10 TO DIOR*

Method	Backbone	Params (MB)	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
Source only	ResNet101	47.37	0.5368	0.1067	0.1653	0.6147	0.6541	0.3805	0.4409	0.1154	0.0391	0.2343	0.3288
SWDA		52.94	0.6239	0.1330	0.4598	0.8651	0.8162	0.5732	0.5870	0.2131	0.0442	0.2929	0.4608
HTCN		56.16	0.5876	0.1297	0.4659	0.8613	0.8138	0.5401	0.5549	0.2143	0.1050	0.2315	0.4504
SCL		63.99	0.6810	**0.1564**	0.4661	0.8645	0.8242	0.5163	0.6301	**0.2345**	0.0745	0.2884	0.4736
Proposed		53.06	**0.7755**	0.1371	**0.4826**	**0.8865**	**0.8965**	**0.6676**	**0.6900**	0.2006	**0.1354**	**0.3046**	**0.5159**
Source only	VGG16	136.88	0.5130	0.1104	0.2342	0.8129	0.7709	0.4253	0.5752	0.1135	0.0314	0.2570	0.3844
SWDA		140.09	0.5578	0.1269	0.3543	0.8133	0.7999	0.5871	0.4356	0.1920	0.0361	0.2733	0.4176
HTCN		143.41	0.5918	**0.1336**	0.3800	0.8150	0.8062	0.5785	0.4762	0.2024	0.1154	0.2816	0.4381
SCL		152.33	0.5817	0.1335	**0.3813**	0.8314	0.8335	0.5932	0.5713	**0.2251**	0.1082	0.2843	0.4544
Proposed		140.21	**0.7237**	0.1282	0.3033	**0.8845**	**0.8752**	**0.6385**	**0.6713**	0.2190	**0.1374**	**0.2983**	**0.4879**

*Values of AP and mAP in bold are the best.

TABLE III
QUANTITATIVE COMPARISON WITH THE SOURCE-ONLY AND OTHER DOAMIN ADAPTATION OBJECT DETECTION METHODS IN THE EXPERIMENTS OF THE NWPU VHR-10 TO HRRSD*

Method	Backbone	Params (MB)	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
Source only	ResNet101	47.37	0.7452	0.3682	0.1981	0.2366	0.5735	0.2210	0.5683	0.4300	0.2600	0.5454	0.4146
SWDA		52.94	0.7367	0.2281	0.2594	**0.3613**	0.6338	**0.3483**	0.3442	0.4571	**0.5894**	**0.6802**	0.4639
HTCN		56.16	0.7335	0.3084	0.6089	0.1322	0.6054	0.2235	0.6280	0.3912	0.3420	0.5985	0.4572
SCL		63.99	0.7302	0.3075	0.6174	0.1965	0.5972	0.2204	0.6683	0.4231	0.4131	0.5566	0.4730
Proposed		53.06	**0.8555**	**0.4164**	**0.6371**	0.2250	**0.6814**	0.2555	**0.7687**	**0.5829**	0.4559	0.6490	**0.5528**
Source only	VGG16	136.88	0.5524	0.1810	0.2461	0.1667	0.5471	0.1783	0.3474	0.1103	0.0930	0.4647	0.2887
SWDA		140.09	0.4648	0.1915	0.1770	0.1283	0.6112	0.1857	0.3701	0.1454	0.1701	0.4587	0.2903
HTCN		143.41	0.6503	0.2041	0.2251	**0.2435**	0.6575	0.1738	0.4458	0.2039	0.1603	0.4432	0.3408
SCL		152.33	0.6801	0.2431	0.2412	0.1952	0.6057	0.2282	0.5069	0.3182	0.2490	0.5808	0.3848
Proposed		140.21	**0.7862**	**0.3512**	**0.3675**	0.2130	**0.6762**	**0.3038**	**0.6205**	**0.5280**	**0.3533**	**0.6238**	**0.4824**

*Values of AP and mAP in bold are the best.

precision and recall, which are defined as

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{15}$$

$$\text{Recall} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{16}$$

where N_{TP}, N_{FP}, and N_{FN} stand for the number of true positives (TPs), false positives (FPs), and false negatives (FNs), respectively. TPs are the objects detected correctly in the test set, FPs are the objects detected that are incorrectly, and FNs are the missed objects in the detection results. In general, when calculating AP, a predicted bounding box is viewed as TP if its IOU with the ground truth is greater than 0.5. Otherwise, it is treated as FP. Also, when multiple objects match the ground truth with high IOU, the object that has the highest detection confidence is usually selected as TP. And, the AP is calculated as the mean precision with a set of equally spaced recall rates S. In our experiments, we use $S = \{0, 0.1, \ldots, 1\}$, and the AP defined as

$$\text{AP} = \frac{1}{11} \sum_{r \in S} \text{Precision}|_{\text{Recall}=r}. \tag{17}$$

And, we calculate the mean of AP with all categories as mAP to evaluate the detection model for multiclass object detection.

C. Experimental Results and Analyses

In this section, we evaluate our domain adaptation method RFA-Net for domain adaptation object detection of RS imagery qualitatively and quantitatively. As described in Section IV-A and Table I, we collect three RS imagery datasets for conducting experiments. In all experiments, the NWPU VHR-10 dataset serves as the source domain, while all others are viewed for the target domain. Tables II and III give the results of our experiments. Source only indicates that the detection model is trained only on the source domain data, after which it is tested directly on the target domain. SWDA denotes the strong–weak alignment domain adaptation method [58]. Besides, we also compare the HTCN [61] and SCL [59] methods deployed with different backbones in the DIOR* and HRRSD* datasets.

1) NWPU VHR → DIOR:* As given in Table II, we have presented the AP and mAP for RS imagery object detection on the DIOR* test set. The proposed RFA-Net outperforms the source-only method and other domain adaptation object detection methods, besides achieving the highest AP for almost all categories and the highest mAP compared to other methods. In our experiments, we set α in (12) and β in (13) to 0.5 and 1.5, respectively. To generate stable pseudo labels for the unlabeled target domain data as supervision, the threshold τ for filtering predictions is set to 0.9.

Compared with the performance of the source-only model, all of the methods can improve the performance of the detection model. The experimental results indicate that a domain shift certainly exists between the source domain dataset NWPU VHR-10 and the target domain dataset DIOR* because of the diversity in the acquisition conditions of RS imagery in the two datasets. Meanwhile, the experimental results of our proposed RFA-Net are much higher than those of the SWDA, HTCN, and

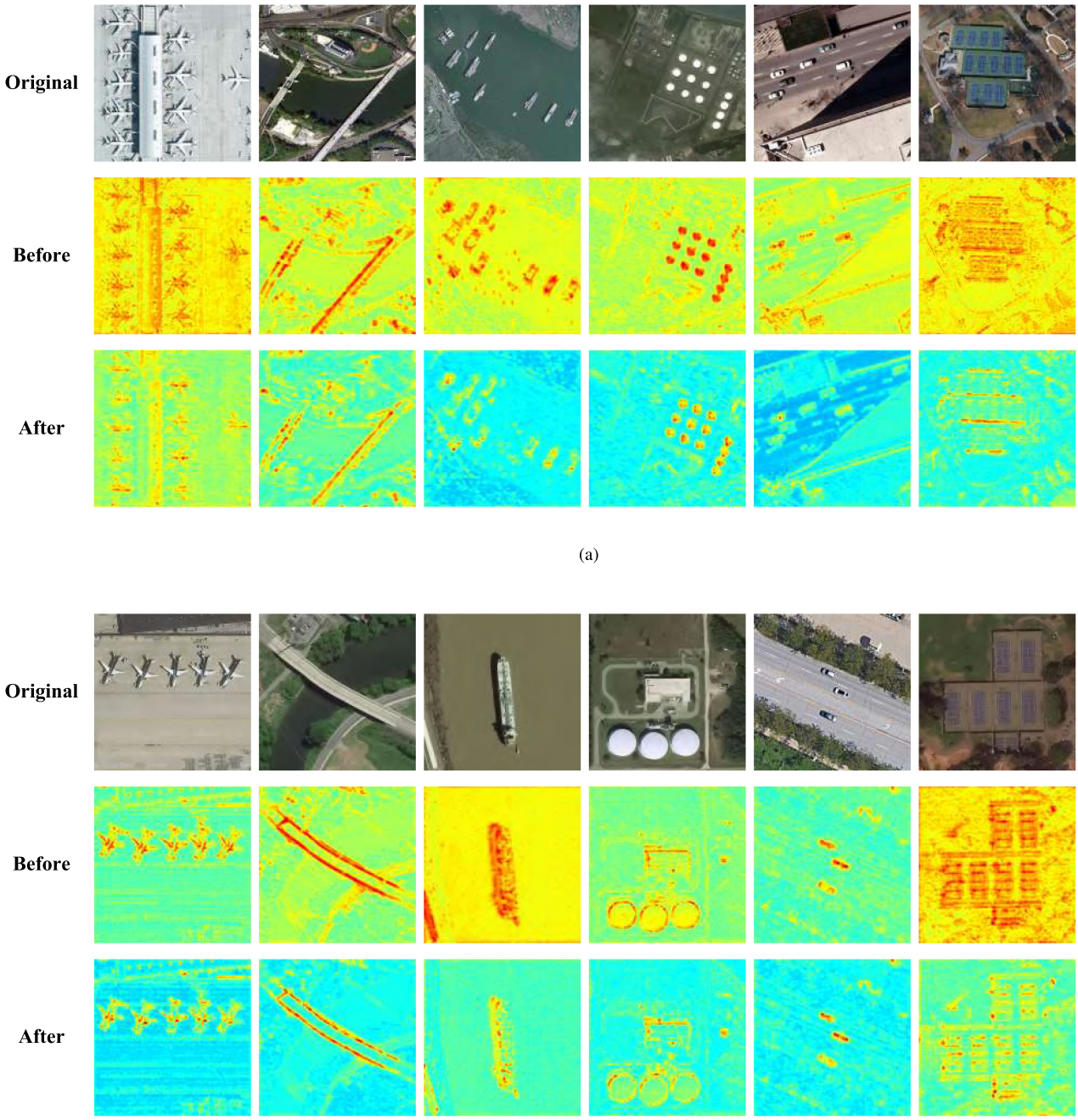
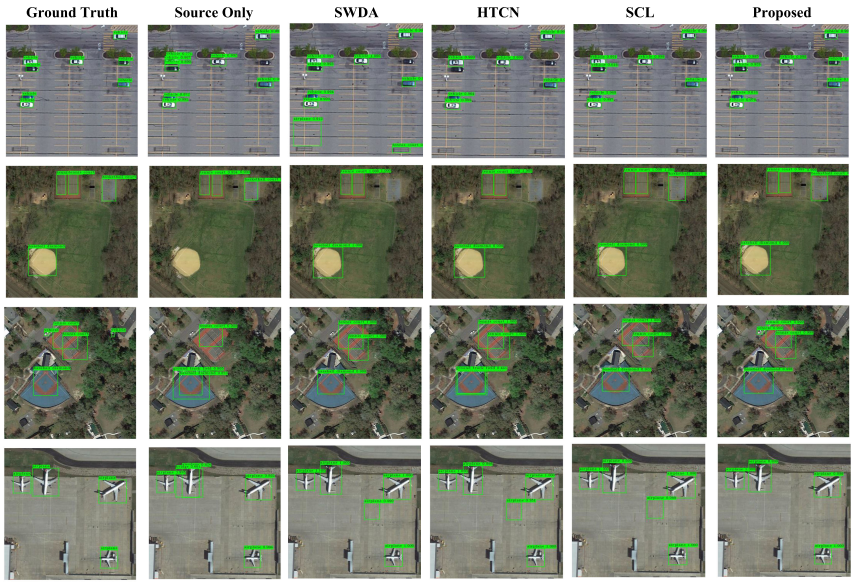


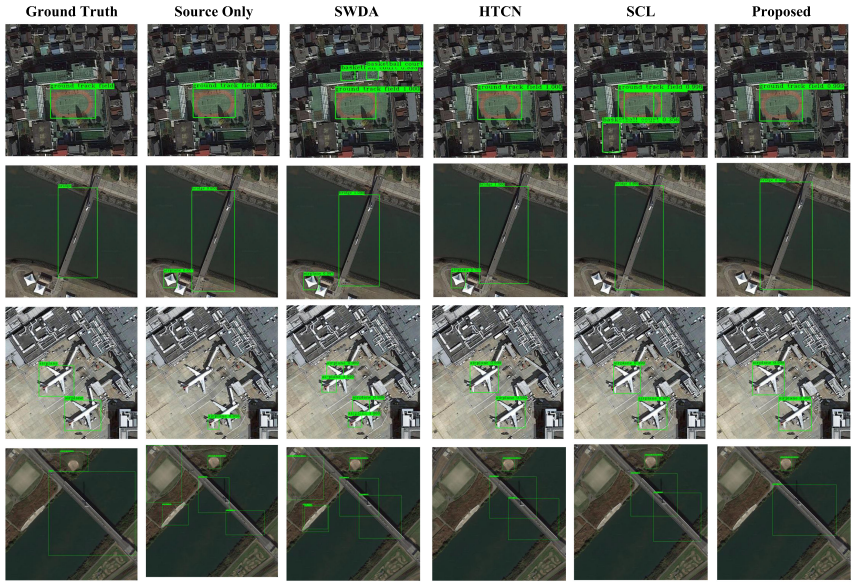
Fig. 6. In (a) and (b), from top to down, the original images in the NWPU VHR and DIOR* datasets, the visualization of the features in the low-level network before SFR, and the visualization of the features in the low-level network after matrix reduction. In the visualization, orange color represents more information embedded in the feature and blue color indicates less information embedded. The redundant features are suppressed significantly after SFR. (a) Visualization of images and features in the NWPU VHR dataset. (b) Visualization of images and features in the DIOR* dataset.

SCL methods conducting the alignment of the semantic features directly. Since the structure of RS imagery is more complex and instance-level features are crucial to be concerned during conducting alignment. Such methods aligning the alignment of the semantic features directly is not effective for domain adaptation object detection of RS imagery, and the proposed algorithms need to take the properties of object detection in RS imagery into account.

Addressing the complexity of RS imagery, our method deploys matrix reduction on the low-level features extracted from the subnetwork F_1 in the source and target domains first, and then deploys low-level alignment on the features after suppressing the redundant information. Fig. 6 shows the visualized heatmap of low-level features in the source and target domains extracted from the subnetwork before and after the SFR. In Fig. 6, the orange color represents more information embedded in the feature



(a)



(b)

Fig. 7. (a) and (b) shows the qualitative detection results from the NWPU VHR dataset to the DIOR* and HRRSD* datasets, respectively. In (a) and (b), columns from left- to right-hand side represent the visualization detection results, i.e., ground truth, source only, SWDA, HTCN, SCL, and our proposed RFA-Net. (a) Visualization of detection results in the DIOR* dataset. (b) Visualization of detection results in the HRRSD* dataset.

and the blue color indicates less information embedded. We can see that the redundant features are suppressed significantly after the SFR, and the instance-level features are more significant, which facilitates the subsequent low-level alignment of the features to alleviate the domain shift. Moreover, by comparing the experimental results of our proposed method and others in Table II, our proposed SDA, SFR, and PLG work together to perform the optimal results of the experiments. Comparisons of the AP results for each category in Table II show that we can achieve the best performance in almost all categories.

We also show the visualization of the detection results on the DIOR* dataset in Fig. 7(a). While the SWDA, HTCN, and SCL methods can do better than the source-only method on some categories, we can find that there are still many false alarms in the detection results, as it fails to align the instance-level features effectively. All of them adopt the same pixel-by-pixel feature alignment on the low-level feature, which is difficult to obtain good performance. Due to the overalignment of features at low-level, it is easy to cause some inaccurate information to be aligned. The excessive alignment is accompanied by the biased information in the source domain, which leads to more FPs in target domain data and decreases the detection model's performance.

2) NWPU VHR → HRRSD:* In the experiments from the NWPU VHR-10 dataset to the HRRSD* dataset, α in (12) and β in (13) are also set to 0.5 and 1.5, respectively. τ, used to filter predictions, is set to 0.6 to enable the model to perform best.

TABLE IV
INFERENCE SPEED COMPARISON WITH OTHER METHODS ON THE DIOR* DATASET BY A 2080TI GPU

Method	Source only	SWDA	HTCN	SCL	Ours
FPS	18	16	15	14	18

The Backbone is ResNet101 for all the models in this table.

Compared with the task from the NWPU VHR-10 to DIOR* datasets, the task from the NWPU VHR-10 to HRRSD* datasets is more challenging. The quantity of HRRSD* is approximately 26 times larger than the NWPU dataset. Not only there is a domain shift between the different domains, but a large amount of unlabeled target domain data also makes it difficult to have a good performance of the model. Considering the huge volume gap of data quantity between the different domains, the pseudo labels generated by the PLG work well to supplement supervision information of the unlabeled target domain data and improve the robustness of adversarial training.

As given in Table III, we can conclude that our RFA-Net is considerably better than the source-only method and other domain adaptation object detection methods in the experiments of NWPU VHR-10 to HRRSD*. Moreover, in the experiments with the backbone of ResNet101, our proposed method achieves a great improvement on some categories, such as ship, storage tank, and ground track field. In particular, our method achieves an AP improvement of 0.3778 on the storage tank compared with other methods.

Despite the noisy pseudo labels in target domain cannot be filtered by the PLG, the SDA treated as one strong data augmentation is robust for the model training with noisy labels. Furthermore, the redundant features in low-level features are removed before the alignment. The instance-level features can be aligned more precisely and the FPs during inference time can be avoided by our RFA-Net. Fig. 7(b) shows the visualization of detection results in the HRRSD* dataset. Our method not only regresses to the bounding boxes accurately, but also has fewer false alarms in the inference results.

3) Network Computational Burden Comparison: We also calculate the number of parameters of the model within each method and give the detailed results in Tables II and III. Compared with SWDA, we add the SFD module, but remove the corresponding context vector computation in the pipeline. Therefore, the parameters of our proposed method are only 0.21 MB more than SWDA. However, the final experimental results are much more significant compared to SWDA due to the effectiveness of our proposed RFA-Net. In addition, we conducted the inference speed test using one NVIDIA 2080 Ti GPU on the DIOR* dataset with ResNet101 as the backbone. We yield the final inference speed results, as given in Table IV. Since our SFR module is only employed in the training process, the entire model of RFA-Net is a pure faster RCNN during testing, but SWDA, HTCN, and SCL still rely on the computation of other modules in their inference stages, such as the context vector or attention. Therefore, their inference speed is reduced compared to our RFA-Net.

TABLE V
ABLATION STUDY OF OUR PROPOSED RFA-NET

	Backbone	SDA	SFR	PLG	mAP
Baseline	ResNet101				0.4608
w/ SDA		✓			0.4712
w/ SFR			✓		0.4641
w/ PLG				✓	0.4677
w/o SDA			✓	✓	0.4693
w/o SFR		✓		✓	0.4900
w/o PLG		✓	✓		0.4756
RFA-Net		✓	✓	✓	**0.5159**

The significance of boldface number means the best value of mAP.

D. Ablation Study

In this section, we first carry out an ablation study to demonstrate the effectiveness of each component (one SDA module, one SFR module, and one PLG module) in RFA-Net in the experiments from the NWPU VHR-10 to the DIOR* datasets. Furthermore, we investigate the effect of different thresholds τ used for filtering pseudo labels and hyperparameters α in the detection loss in (12). In the end, we also display the performance of the source only, SWDA, and our proposed method in the DIOR* dataset varies with different IOU thresholds.

1) Effectiveness of Each Component: To verify the effectiveness of each component in RFA-Net, we equip individual modules of SDA, SFR, and PLG, respectively. Besides, we also remove SDA, SFR, and PLG in our RFA-Net, respectively. The results of these experiments are given in Table V. It is worth noting that in the ablation experiment of the single SDA module, the instance-level annotations of the data in the target domain are not available. Therefore, we can only apply the image-level transformation and randomly mask the image when we add the SDA module. And, the situation in the ablation experiment of removing PLG is the same.

We can observe that the model does not yield many benefits in the end when equipped with only a single module. In terms of SDA, the promotion of single SDA is limited due to the large data volume gap between the source and target domains. Similarly, the model with a single SFR module can denoise the redundant features in the source domain with instance supervision, but has difficulty performing effectively in the unlabeled target domain data during training. PLG not only provides supervision for the target domain, but also improves the robustness of adversarial training. However, noisy labels are also difficult to be completely filtered out by the strategy in PLG. So, it is still challenging to acquire high performance when using PLG only.

The model can have a better performance when the SDA and PLG cooperate, gaining a 2.91% improvement over the baseline model, as given in Table V. It not only introduces supervised information in the source and target domains to the model, but also makes the model more robust to training with noisy labels. Although the pseudo labels generated by PLG contain noisy labels inevitably, the training of the model is more robust due to the promotion of SDA, and hence SFR can better develop the role of denoising the redundant features for source and target

TABLE VI
PERFORMANCE OF OUR PROPOSED METHOD IN THE DIOR* DATASET WITH DIFFERENT α

α	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
0.50	0.7765	0.1368	**0.4909**	0.8676	0.8495	0.6285	0.6863	**0.2130**	0.1236	0.2991	0.5072
0.75	0.7358	**0.1663**	0.4081	0.8728	0.8540	0.6292	0.6835	0.1493	0.0857	0.3277	0.4912
1.00	0.7635	0.1452	0.4487	0.8754	0.8965	**0.6738**	0.6381	0.2039	0.1358	0.3111	0.5092
1.25	0.7364	0.1548	0.4604	0.8773	0.8839	0.6118	0.6296	0.1317	0.1286	**0.3327**	0.4947
1.50	0.7755	0.1371	0.4826	0.8865	0.8965	0.6676	**0.6909**	0.2006	0.1354	0.3046	**0.5177**
1.75	0.7334	0.1374	0.4196	**0.8912**	0.8769	0.6055	0.6482	0.1543	0.1189	0.3164	0.4902
2.00	**0.7868**	0.1450	0.4278	0.8834	**0.9033**	0.6679	0.6481	0.1761	**0.1382**	0.3113	0.5088

*Values of AP and mAP in bold are the best.

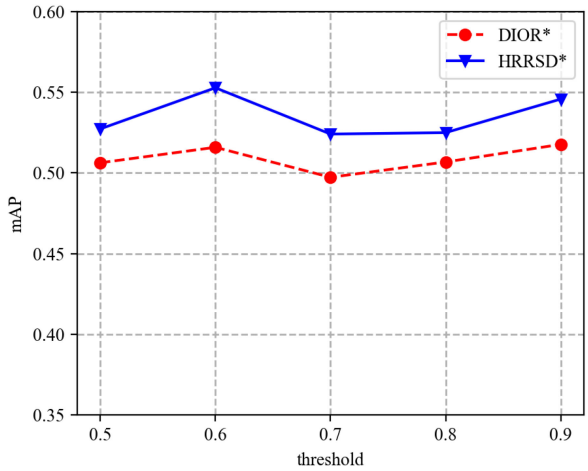


Fig. 8. Performance of our proposed method in the DIOR* and HRRSD* datasets varies with different thresholds τ.

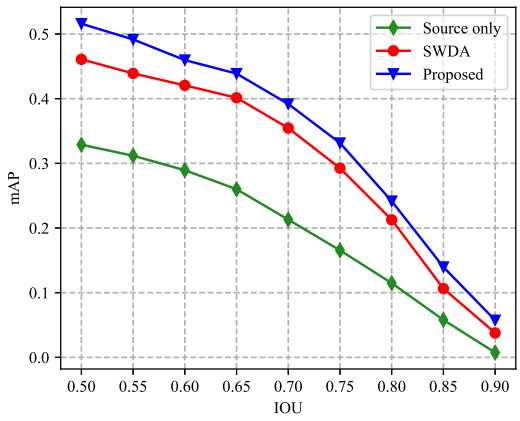


Fig. 9. Performance of source only, SWDA, and our proposed method in the DIOR* varies with different IOU thresholds.

domains. Finally, the model with SDA, SFR, and PLG obtains a 5.1% improvement compared to the baseline.

2) Effectiveness of Threshold τ: Fig. 8 shows the test results of the detection model on the DIOR* and HRRSD* datasets with different thresholds when α is fixed to 1.5. For the DIOR* dataset, the detection model achieves the highest mAP when the threshold is 0.9. However, for the HRRSD* dataset, the detection model is optimal when the threshold is 0.6.

The threshold enables the predictions with high quality to be selected. A large threshold can filter out most of the low confidence proposals, but it can easily lead to a low recall of the detection model. It is because the pseudo labels are hard to cover all the real objects. Therefore, the model can neither learn those hard-to-learn instance features in the target domain dataset from the source domain nor can the pseudo labels be used as reliable proposals. While a small threshold tends to cause false alarms with low confidence to be added to the pseudo labels, rendering large quantities of false alarms in the prediction results with low precision. Therefore, for different object detection domain adaptation tasks, it requires us to set the appropriate threshold for the generation of pseudo labels for the target domain data.

3) Effectiveness of Hyperparameters α: We further conduct the ablation study on the difference α of (12), and the experimental results are given in Table VI. The hyperparameter α controls the tradeoff between the total RPN losses of source and target domains.

We can conclude from observing the experimental results in Table VI that different α affects the performance of the model in different categories. When the selected pseudo labels of the target domain are utilized as supervision, it is necessary for us to make a tradeoff between the RPN losses of different domains because of the discrepancy between the pseudo label and ground truth. Thereby, better proposals can be selected for subsequent bounding box regression and classification.

When α is set as 1.5, the detection model achieves the best performance of the overall mAP compared to the other values of α. Even though the detection model does not perform optimally in each category, it performs uniformly on each category, with no results where the AP of one category is much lower than under the other values of α.

4) Influence of Different IOU Thresholds: To compare the accuracy and robustness with other models, we conducted experiments utilizing different IOU thresholds in source only, SWDA, and our proposed RFA-Net. Fig. 9 shows the experimental results that our proposed method consistently outperforms the source-only and SWDA methods with the variation of IOU thresholds. Simultaneously, it demonstrates that our RFA-Net has better accuracy and more robust bounding boxes regression for domain adaptation object detection of RS imagery.

V. CONCLUSION

In this article, we proposed an RFA-Net for unsupervised domain adaptation object detection in RS imagery, to alleviate

the domain shift between different domains in RS imagery. The RFA-Net includes one SDA module, one SFR module, and one PLG module. All of these modules were deployed on multistages of the pipeline separately. In the stage of data preprocessing, due to the domain shift and volume gap in data quantity between different domains, we introduced the SDA to expand the data volume and improve the robustness of adversarial training. Moreover, we proposed the SFR and inserted it before the low-level feature alignment to relieve the excessive alignment of semantic features. Such a module reconstructs the features for denoising the redundant features and implicitly intensifying the specific instance features for alignment in their respective domains. Furthermore, the detection model tends to learn biased information in the source domain with only labeled data of the source domain available. We conducted one PLG to generate the labels for the unlabeled target domain by exploiting the knowledge learned from the source domain data. Our experiments demonstrated that our method can promote the generalization of the RS imagery object detection model effectively. We hope our work can inspire future exploration in alleviating the domain shift between different domains of object detection in RS imagery. In the future, we will also further pursue the problem of category imbalance in domain adaptation object detection to improve performance. Besides, when source domain data are unavailable and only the model trained in the source domain is available, source-free domain adaptation object detection will also be explored to protect data privacy and save data transmission expenses.

REFERENCES

[1] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.

[2] 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. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[5] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8231–8240.

[6] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for detecting oriented objects in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2844–2853, doi: 10.1109/CVPR.2019.00296.

[7] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 132.

[8] X. Sun *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.

[9] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[10] Y. Lin, P. Feng, J. Guan, W. Wang, and J. Chambers, "IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," 2019, *arXiv:1912.00969*.

[11] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 268–279, 2020.

[12] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5603613.

[13] X. Zhang, G. Wang, P. Zhu, T. Zhang, C. Li, and L. Jiao, "GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3518–3531, Apr. 2021.

[14] 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.

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

[16] C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2615–2626, May 2016.

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

[18] L. Ma, M. M. Crawford, L. Zhu, and Y. Liu, "Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2305–2323, Apr. 2019.

[19] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.

[20] 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. Photogrammetry Remote Sens.*, vol. 155, pp. 72–89, 2019.

[21] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3558–3573, May 2020.

[22] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 263–275, 2020.

[23] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 20–33, 2021.

[24] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 25, pp. 723–773, 2012.

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

[26] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "Fada: Feature aligned domain adaptive object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617916.

[27] Y. Chen, Q. Liu, T. Wang, B. Wang, and X. Meng, "Rotation-invariant and relation-aware cross-domain adaptation object detection network for optical remote sensing images," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4386.

[28] 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, 2020.

[29] 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, 2020, Art. no. 575.

[30] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," 2020, *arXiv:2005.04757*.

[31] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is attention better than matrix decomposition," in *Proc. Int. Conf. Learn. Representations*, 2021.

[32] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 361–368.

[33] Y. Feng, X. Sun, W. Diao, J. Li, X. Gao, and K. Fu, "Continual learning with structured inheritance for semantic segmentation in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607017.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
[35] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[38] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9626–9635.
[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.
[40] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[41] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020.
[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[43] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1, pp. 151–175, 2010.
[44] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
[45] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2014, pp. 898–904.
[46] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.
[47] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[48] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn*, 2017, pp. 2208–2217.
[49] L. Du *et al.*, "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 982–991.
[50] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.
[51] H. Yang, X. Xu, Y. Ma, Y. Xu, and S. Liu, "CraterdaNet: A convolutional neural network for small-scale crater detection via synthetic-to-real domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4600712.
[52] Y. Shi, J. Li, Y. Li, and Q. Du, "Sensor-independent hyperspectral target detection with semisupervised domain adaptive few-shot learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6894–6906, Aug. 2021.
[53] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 480–490.
[54] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11449–11458.
[55] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Inf. Fusion*, vol. 46, pp. 206–217, 2019.
[56] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.
[57] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 687–696.
[58] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong–weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6949–6958.
[59] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "SCL: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses," 2019, *arXiv:1911.02559*.
[60] C. Zhuang, X. Han, W. Huang, and M. Scott, "iFAN: Image-instance full alignment networks for adaptive object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020 pp. 13122–13129.
[61] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8866–8875.
[62] V. F. Arruda *et al.*, "Cross-domain car detection using unsupervised image-to-image translation: From day to night," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[63] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1029–1038.
[64] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29935–29948.
[65] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3008–3017.
[66] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11192–11203.
[67] M. Cramer, "The DGPF-test on digital airborne camera evaluation overview and test design," *Photogrammetrie, Fernerkundung, Geoinformation*, vol. 2010, pp. 73–82, May 2010.
[68] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

Yangguang Zhu received the B.Sc. degree in automation from the University of Science and Technology of China, Hefei, China, in 2019. He is currently working toward the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

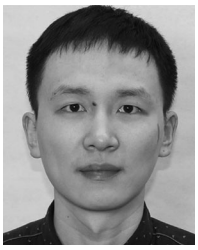
His research interests include computer vision and remote sensing interpretation, with a focus on object detection domain adaptation.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Wenhui Diao (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016.

He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.



Hao Li (Member, IEEE) received the B.E. degree from Jilin University, Changchun, China, in 2014, and the M.Sc. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2017, all in information and communication engineering.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and remote sensing image processing.

Kun Fu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, and geospatial data mining and visualization.