IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2020, accepted March 13, 2020, date of publication March 20, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982034

Multi-Adversarial Partial Transfer Learning With Object-Level Attention Mechanism for Unsupervised Remote Sensing Scene Classification

PENG LI[1,2], DEZHENG ZHANG[1,2], PENG CHEN[3,4], XIN LIU[1,2,5], AND AZIGULI WULAMU[1,2]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
[3]School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[4]Beijing Institute of Remote Sensing Equipment, Beijing 100854, China
[5]Surgery Simulation Research Lab, University of Alberta, Edmonton, AB T6G 2E1, Canada

Corresponding authors: Peng Chen (b20150280@ustb.cn) and Aziguli Wulamu (aziguli@ustb.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0823002, in part by the Key Research and Development Program of Ningxia Hui Autonomous Region under Grant 2019BFG02009, and in part by the National Natural Science Foundation of China under Grant 61801019.

ABSTRACT In recent years, deep learning methods have been widely applied in remote sensing image classification tasks, providing valuable information for natural monitoring and spatial planning. In an actual application like this, acquiring massive labeled data for deep convolutional networks is costly and difficult especially in the situation that the data sources are diverse and the requirements are changing. Transfer learning methods have already shown superior performance on exploiting domain invariance features in existing data for deep network-based categorization tasks. However, the data imbalance between source and target domains may bring negative transfer and weaken the classifier's ability. Moreover, it is still a difficult problem to extract object-level visual features among easy-mixed categories. In this context, **Multi-adversarial Object-level Attention Network (MOAN)** is proposed for partial transfer learning and selecting useful features. On the one hand, we present an improved object-level attention proposal network (OANet) for perceiving structural features of the main object in the picture, and weakening the unrelated regions. On the other hand, the extracted features are further enhanced by multi-adversarial framework in order to promote positive transfer, selecting and mapping valuable cross domain features from shared categories and suppressing others. This adversarial learning module can also generate pseudo tags for the samples in target domain so as to perceive integral visual signals, similar to the process in source domain. In addition, virtual adversarial training method is introduced in MOAN so as to regularize the model and maintain stability. Experimental analyses show that our MOAN can significantly promote positive transfer and restrain negative transfer in unsupervised classification problems. MOAN has good performances such as higher accuracies and lower loss values on several benchmark data sets.

INDEX TERMS Partial transfer learning, domain adaption, object-level attention, remote sensing scene classification, multi-adversarial learning, convolutional neural networks.

I. INTRODUCTION

It is always a tough work to acquire sufficient labeled data for training complex models in a special application, like remote sensing scene classification. The problem will worsen when the monitoring targets are changing in different periods and different regions. It is impracticable to label massive data accordingly for those effective but complicated deep networks. Transfer learning is regarded as a good solution for solving problems like this. As a typical transfer learning task, domain adaption has gained attention from many researchers in the past decade [1].

The associate editor coordinating the review of this manuscript and approving it for publication was L. Zhang.

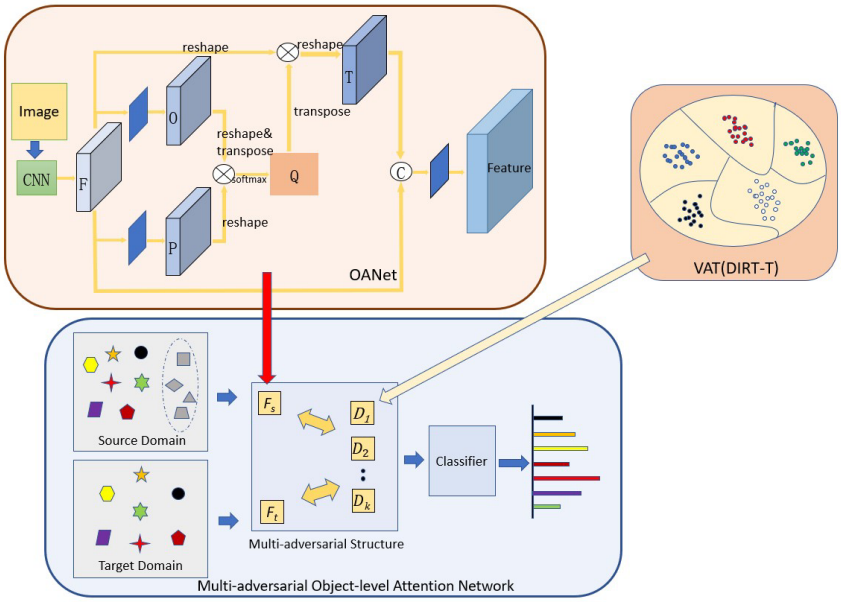


FIGURE 1. An overview of the proposed MOAN. MOAN consists of 2 main sub-modules: 1) an improved object-level attention mechanism (OANet) for extracting features and evaluating relationships among pixels; 2) a main framework composed of multi-adversarial units which will promote positive transfer and implement unsupervised learning between related categories in source and target domains. VAT denotes Virtual Adversarial Training. Particularly, we use the DIRT-T method in [8].

Some scholars have introduced Generative Adversarial Network (GAN) [2] to implement domain adaption, like [3] and [1]. For these models, there is a precondition that the label spaces in source domain and target domain are identical. However, in most real transfer learning scenarios, data imbalance is very common that the data and labels in source domain are more and more complicated than those in target domain, like from ImageNet [4] to NWPU-RESISC45 [5].

How to carry out a positive transfer and select useful domain-invariant features from large, complex source domain is the key point in actual remote sensing applications. Many existing methods adopt symmetric feature representation structures in which there is only one feature extractor in both source and target domains [6]. This is not suitable for the situation that the gaps between different domains are big, like the domain gap between remote sensing pictures and our daily life pictures. There are visual particularities in remote sensing pictures, different from daily life pictures that may be taken in everyday situations, perhaps by an digital camera. Instead, the remote sensing pictures can be obtained by the satellite, radar or other professional equipment and will be processed and analyzed for spatial perception or other applications. On the other hand, it is also a challenging topic to perceive accurate object-level features from remote sensing images in which the objects are distributed sparsely.

It has been shown that partial transfer learning is a good method to promote positive adaption by the utilization of class-wise multi-adversarial learning unites [7]. Due to the fact that there are always data imbalance in transfer learning tasks and the categories in source domain are much more than those in target domain, for remote sensing tasks, we propose the **Multi-adversarial Object-level Attention Network (MOAN)** in this article. There are 2 main submodules in the proposed MOAN: 1) attentional feature extractors for source and target domains respectively; 2) domain adoption module with multi-adversarial units. In order to obtain high quality visual representations and promote an positive transfer, thus utilize useful information and eliminate the irrelevant, we make different improvements. Effective training method is also applied to MOAN. The overview of MOAN is shown in Figure 1. Our contributions can be listed as follows:

1) An improved object-level attention proposal network (OANet) is introduced and embedded in MOAN. OANet adopts improved self-attention mechanism which can take a good use of contextual information and pixel-level spatial relationships, suitable for remote sensing images in which the object may be far away from another related one.
2) GAN-based partial transfer learning is utilized in MOAN. We use a group of generative-adversarial units in the framework for aligning cross-domain features class-wisely and partially.
3) We adopt different feature extractors for different domains and design an asymmetric framework for mappings which can extract features with the progressing "pseudo" tags in target domain and remove the the gradient reversal layer so as to prevent model instability.

4) We design an alternative training strategy for MOAN and select an ''warm booting'' method so as to obtain credible ''pseudo'' labels and propose attention regions effectively. Virtual adversarial training [8] is also used in MOAN.

The rest of this paper is divided into 4 parts. In section II, related works are reviewed briefly; the proposed MOAN is introduced and analyzed in section III; comparative experiments are conducted in section IV and the conclusion is given in section V.

II. RELATED WORK

A. DEEP TRANSFER LEARNING FOR SCENE CLASSIFICATION

With the development of satellite technologies, a large number of high resolution remote sensing images have been accumulated for different regions of interest. Although there are many mature technologies in the processing of hyper-spectral images [9], [10], it always a challenging topic to utilize advanced computer vision strategies in the remote sensing field and handle pictures in different bands including RGB images [11]. In recent years, deep learning technologies have become a widespread concern that Convolutional Neural Networks (CNN) are effectively applied in computer vision [4], language processing [12], [13], fault diagnosis [14], [15] and fault tolerant control [16] fields. More and more scholars adopt deep learning-based computer vision methods to handle scene classification for remote sensing images and these methods have shown better feature extraction and semantic comprehension abilities than those classic methods like Gist [4]. However, complex model structures and lots of network parameters pose new challenges for the utilization of Convolutional Neural Networks (CNN) [4] in remote sensing fields. Compared with acquiring massive labeled samples, those strategies with limited supervisory signals, like semi-supervised learning [17], [18], small-sample learning [19], un-supervised learning [1] and zero-short learning [20], are more economic and feasible.

This article mainly focus on transfer learning for remote sensing image classification. In this topic, Penatti Otavio *et al.* evaluated the performance of CNN networks trained on daily life pictures for the classification of remote sensing images [21], [22]. Nogueira *et al.* [21] and Scott *et al.* [23] fine-tuned CNN networks by transforming the final layers of CNN, like ResNet [24] and GoogLeNet [25], according to the categories of pictures. Their results show that only need a small amount of labeled data, the model is able to achieve excellent accuracies. Luus *et al.* even re-trained simplified CNN model by multi-view learning [26].

In order to improve the feature extracting ability of CNN models on this transfer learning task, Chaib *et al.* combined more than 1 CNN features by discriminant correlation analysis [27]; Cheng *et al.* coded the CNN features by semantic bags then dealt the categorization by one-versus-all support vector machine (SVM) classifiers [28]. These strategies take advantages of original CNN features but make little improvements on CNN architectures. In addition, most of them still need a few amount of labeled data which may be scarce in some real applications. Recently, attention mechanism is popular for improving CNN structures and have drawn attention from computer vision community. There are also researchers pay attention to unsupervised domain adaption to handle unlabeled data.

B. ATTENTION MECHANISM

In computer vision tasks, one important issue is to extract object-level visual features and recognize them. This issue is more important in remote sensing image processing due to the fact that the objects in large remote sensing pictures are always distributed unevenly and easily confused with each other. Classical methods adopt sparse or low-rank representations [29], [30] for image patches [31] and they are very effective in remote sensing area. Recently, deep learning-based attention mechanism [32], [33] is regarded as good solution for problems like this, perceiving object-level features which illustrate the main sematic information in remote sensing scenes.

These years, attention mechanism is widely applied in computer vision and natural language processing tasks. Spatial Transformer Network [32] is classical work which tried to generate rotation angles and grids of objects by the training of location network embedding in CNN architectures. SENET [34] is another representative work which aims to learn the weights for different channels. Besides, self-attention method [33], [35] is designed for evaluating the long-range relationships between different objects and regions by a pixel-level matrix multiplication.

In remote sensing field, Haut *et al.* proposed residual channel attention-based neural network model which integrated the attention module into the residual CNN layers [36]; Luo *et al.* also introduced channel attention mechanism into Fully Convolutional Networks to select appropriate features [37]; Wang *et al.* also improved Fully Convolutional Network by integrating class-specific attention model and ResNet [38]; Ba *et al.* incorporated spatial and channel-wise attention modules in CNN architectures to enhance features and detect fire smoke in satellite pictures [39].

C. GENERATIVE ADVERSARIAL NETWORK, DOMAIN ADAPTION AND UNSUPERVISED LEARNING

The monitoring requirements in remote sensing field is changing frequently resulted from the variations in target regions, time and even social development. Unsupervised domain adaption methods are urgently needed in which the objects in target domain can also be recognized by making effective use of existing labeled data and knowledge. Different other un-supervised learning methods like zero-shot learning which mainly focuses on newly-presented and never seen categories, in domain adaption, label space is shared between source domain and target domain [1], [3].

In recent years, GAN-based Domain Adaption methods are popular in many un-supervised classification and segmentation tasks [40]. The most common flow is to obtain the initial feature representations and classifier in source domain in which the data are sufficient, than align the features and map samples from target domain to source domain so as to classify them by the obtained classifier. Ganin *et al.* designed the symmetric adversarial learning-based domain adaption framework and the matched training method [1]. In the proposed Domain-Adversarial Neural Networks (DANN) in [1] and [6], samples in source domain and target domain are represented by a same feature extractor and the opposite optimization objectives between generator and discriminator are implemented by a gradient inversion layer [1]. In [3], Tzeng *et al.* presented the typical asymmetric domain adaption structure, Adversarial Discriminative Domain Adaptation (ADDA). In ADDA, 2 different feature extractors are optimized for source domain and target domain respectively. The one is obtained by the training with classification loss on source domain and the other can be acquired during the alternative training with adversarial loss between generator (feature extractor in target domain) and discriminator [3]. Although these models are very effective in the unsupervised classification of office supplies and hand-writing numbers [3], [41], there are still improvements for them in remote sensing fields due to the greater domain differences and data imbalance between daily life pictures and remote sensing images. Partial transfer learning is a valuable topic to promote positive transfer between related categories and avoid negative transfer between unrelated categories. Aiming at this issue, Cao *et al.* proposed Selective Adversarial Network (SAN) which matched feature distributions among shared categories by multi-adversarial learning framework [7]. SAN [7] inspired us a lot for the propose of MOAN in this article.

There are some other representative unsupervised methods in these years, e.g., multi-task learning. Although the ''task'' and ''domain'' are proceed in different styles in different models, these models have similar frameworks. *The first group* is represented by [42] and [43], including a feature mapping module for recognizing speech [42] or text [43], and a discriminator-like module to distinguish different domains, like different speakers, different languages or different review topics. References [42] and [43] regarded the information perception in different domains (or modalities in [44] and [45]) as different ''tasks'', while in [46], speech recognition and speaker recognition were regarded as 2 different tasks. The main structure of [42], [43] and [46] is similar to ADDA [3] and the proposed method in this article. There are always multiple sources, more than 2, in those applications of them [43]. Reference [47] did not adopt adversarial learning-based structure and the classifications in source and target domains were regarded as 2 tasks. These 2 classifiers were boosted by a joint learning and regularized by Maximum Mean Discrepancy (MMD). *The second group* of multi-task learning is based on self-supervised representation, like [48] and [49]. In methods like these, different newly-designed tasks are utilized to obtain better feature mappings. Common self-supervised tasks include rotation prediction [48], relative position evaluation [50], colorization [49], inpainting [51] and even the depth prediction and surface normal estimation [52], [53] (labeled in synthetic images; the task in [53] is also used in GAN-based framework, similar to group 1). In general, these multiple tasks can be proceeded and ensembled in an integrated framework [54], [55].

III. UNSUPERVISED PARTIAL TRANSFER LEARNING FOR REMOTE SENSING SCENE CLASSIFICATION

As analyzed in Section 1, data imbalance and domain differences are very common in the transfer from existing data to a new application. To solve the partial transfer task in remote sensing field, inspired by [3], [7] and [35], we propose a Multi-adversarial Object-level Attention Network(MOAN), as shown in Figure 1.

As shown in Figure 1, in MOAN, OANet is mainly used to perceive the basic features of images and provide visual information for multi-adversarial module. Multi-GAN structure is adopted in MOAN to eliminate the interference of irrelevant categories, and further improve the classifier's performance during the partial transfer learning. VAT denotes Virtual Adversarial Training [8] which instructs the training of GAN-based structure and makes the classifier and domain discriminators abide by locally-Lipschitz constraint so as to keep the classification boundary in target domain far away from the high-density region and obtain higher classification accuracies. In order to deal with the issue that the image attention in target domain cannot be effectively obtained without supervision, we introduce ''pseudo labels'' in target domain. These ''pseudo labels'' provide effective instruction for the optimizing of OANet and multi-adversarial units alternately. Relevant analyses and experimental results will be presented in the following Sections.

A. OBJECT-LEVEL ATTENTION NETWORK (OANet)

In general, satellite pictures are taken with huge spans in which the structures of ground objects are relatively complex. Therefore, globally perceiving objects is an important issue in scene classification. Self-attention mechanism can learn the correlations between any two pixels and obtain global features [33], [35]. Inspired by it, we propose an improved attention mechanism OANet to extract object-level and overall features in-depth in remote sensing images. OANet is shown in Figure 2.

Figure 2 illustrates the architecture of OANet. At first, original images are fed into the backbone (ResNet [24] is used in this article) to extract initial feature maps, as (1),

$$F \in \mathbb{R}^{M \times H \times W}, \tag{1}$$

where M is the number of channels; H and W indicate the size of single feature map; $\in$ denotes the tensor meets the multi-dimensional shape.

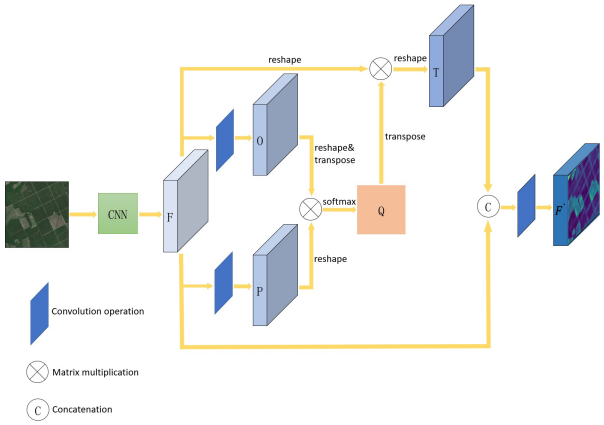


FIGURE 2. The overall structure of OANet. *F*, *O*, *P*, and *T* are different feature mappings in the module; *A* is attention mask proposed by the module.

Then F is further processed as the inputs of two convolutional layers to obtain feature representations $O \in \mathbb{R}^{M\times H\times W}$ and $P \in \mathbb{R}^{M\times H\times W}$ respectively. After that, we reshape and transpose O to $O' \in \mathbb{R}^{N\times M}$ and reshape P to $P' \in \mathbb{R}^{M\times N}$, where $N = W\times H$. In OANet, there is a matrix multiplication between O' and P', then the results are further processed in a softmax layer row by row. Through these steps, we can achieve the attention mask $A \in \mathbb{R}^{N\times N}$, which can be described as equation (2),

$$A_{ij} = \frac{\exp(O'_i \cdot P'_j)}{\sum_{i=1}^{N} \exp(O'_i \cdot P'_j)}, \tag{2}$$

where A_{ij} denotes the j_{th} point's weight associated with the i_{th} point in the feature maps, F.

In addition, we reshape F to $F_\Delta \in \mathbb{R}^{M\times N}$ and apply a matrix multiplication between F_Δ and the transpose of A, and the result is reshaped to $T \in \mathbb{R}^{M\times H\times W}$. Instead of performing a point-wise sum operation between T and F directly, we concatenate them and input the result into a convolution layer to get the final feature maps $F' \in \mathbb{R}^{M\times H\times W}$. The purpose of doing this is to adjust the combination mode between T and F through optimization training iteratively, rather than artificially defining the correlation weight between them, which can be expressed as equation (3),

$$F' = ReLU(f([F, T], \theta_f)), \tag{3}$$

where [,] denotes the concatenation operation; $f(,)$ donates the convolution operation with parameter θ_f; ReLU is the non-linear active function to prevent vanishing gradient.

In order to verify the effectiveness of the proposed OANet, we carry out comparative experiment with original self-attention mechanism on NWPU-RESIS45 [5] data set. In this experiment, Resnet-50 [24] is selected as the backbone and **initial classifier**. We utilize 3 indicators, Precision(P), Recall(R) and F1-score(F1) to evaluate the performance of different attention strategy as written in equations (4) to (6),

TABLE 1. Classification accuracies of ResNet-50 [24] with different attention proposal methods on NWPU-RESISC45.

Method	P(%)	R(%)	F1
ResNet-50(Basic)	81.32	73.23	77.06
ResNet-50 with Self-attention	86.50	75.36	80.55
ResNet-50 with OANet	88.13	76.53	81.92

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{6}$$

where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative respectively.

Experimental results are listed in Table 1. It can be seen that OANet has obvious advantages over basic Resnet-50 [24] and Self-attention mechanism in terms of precision, recall, and F1. The detailed and comprehensive comparison will be given in the experimental section (Section IV).

We also visualize the attention regions on the typical picture from NWPU-RESISC45 [5], as shown in Figure 3. In these attention sub-figures, brighter regions are with higher pixel values.

As illustrated in Figure 3, it can be found that self-attention and OANet significantly surpass basic Resnet-50 [24] in capturing the integrity of object-level attention regions, which indicates that attention proposal plays an important role in feature representation and classification. Compared with self-attention, OANet is more effective in highlighting the main object area and weakening the negative impacts from irrelevant objects. OANet has a better performance on acquiring basic features of remote sensing ground objects with complex structures in large span satellite pictures.

B. UNSUPERVISED PARTIAL TRANSFER LEARNING WITH MULTI-ADVERSARIAL NETWORKS

Partial transfer learning is proposed for the situation that the label space in source domain, L_s, is bigger than that in target domain and all the labels in target domain, L_t, are contained in source domain labels. As shown in Figure 1, $L_t \subset L_s$ (shown in the "source label space" and "target label space"). This is very common in real applications, such as transfer learning from ImageNet [56] to Caltech-256 [57] or from NWPU-RESISC45 [5] to UCM [58]. Most of Traditional transfer learning methods are designed for the situation that the label space is shared between source and target domains [3], [59]. That is $L_s = L_t$, the labels are identical in different domains. However, for most real applications, single transfer framework may be not suitable since the useful knowledge in source domain is always limited. In a huge data set in source domain, most feature mapping modules, like

(a) Original (b) ResNet-50 (c) Self-Attention (d) OANet

FIGURE 3. Attention visualization for a typical picture with basic backbone and 2 attention proposal methods.

single adversarial learning unit, can only roughly eliminate the domain differences. It is possible to bring negative transfer that the mapping will be influenced by other unrelated categories from source domain. The inconsistent between domains makes it difficult to obtain an appropriate adaption.

In this article, we propose the improved unsupervised domain adaption framework with partial transfer learning style, as shown in Figure 4. Multi-adversarial learning units are embedded in the framework. Unlike SAN [7], we set two feature extractors, F_s and F_t separately for the source and target domains. The aim of this is to improve the discriminating ability of domain predictor D through perceiving the inter-domain discrepancy deeply, and to model samples in different domains effectively. In addition, we introduce virtual adversarial training mechanism to further enhance the effectiveness of positive transfer.

In detail, we build the label classifier in source domain by the labeled data at first. The loss can be written as equation (7),

$$L_{advC} = \frac{1}{n_s} \sum_{x_i \in D_s} L_y(C(F_s(x_i)), y_i), \tag{7}$$

where C is the label classifier, and L_y is its loss function; x_i is the labeled data in source domain (in the dataset, D_s, in source domain), and y_i is the related label; n_s is the amount of instances in source dataset.

In addition, according to the number of categories in source domain, we build domain discriminators D_k, $k = 1, 2, \ldots, |L_s|$, where $|L_s|$ is the number of source labels. The domain discriminators can distinguish the domain for samples. The loss function can be written as equation (8),

$$L_{advD} = -\sum_{k=1}^{|L_s|}[\frac{1}{n_s} \sum_{x_i \in D_s} L_d^k \left(D_k \left(F_s \left(x_i\right)\right), d_i\right) + \frac{1}{n_t} \sum_{x_i \in D_t} L_d^k \left(D_k \left(F_t \left(x_i\right)\right), d_i\right)], \tag{8}$$

where n_s and n_t are the numbers of samples in source dataset and target dataset respectively; D_s and D_t denote the source dataset and target dataset; L_d is the loss function of the domain discriminator.

However, due to the presence of irrelevant categories, L_i, this adversarial learning-based domain adaptation may

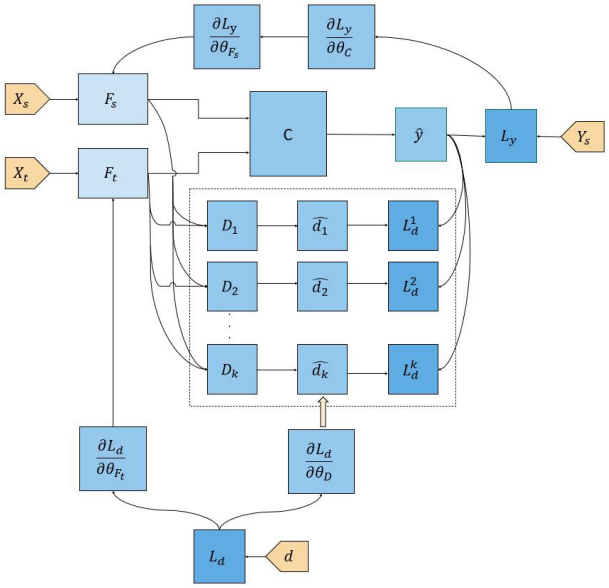


FIGURE 4. The multi-adversarial network for unsupervised partial transfer learning. F_s and F_t are the feature extractors in source domain and target domain respectively. C is the classifier which need to predict data label. X_s and X_t are the images from source and target domains respectively. The multi-GAN structure can align features from different domains and implement positive transfer by selecting out outlier classes. D is domain predictor. Y_s are labels in source domain; d are domain labels. L denotes the loss and θ means the parameters in each module; $\hat{y}$ are predicted data labels and $\hat{d}$ are predicted domain labels.

be hindered. How to eliminate the interference of irrelevant categories is the main factor for improving the performance of partial transfer learning. Since the samples in target dataset are unlabeled, we cannot distinguish the related labels beforehand from those redundant labels that only exist in source domain. Settle for second best, we can get "pseudo labels" through the pre-trained classifier in source domain. The prediction of "pseudo labels" is the result from the analyze of probability distribution of source labels. These probability distributions for each target sample on source labels are superimposed as weights which can be obtained from the softmax normalization, as written in equation (9). This equation is able to determine which labels are valid and which are not. Along with the continuous optimization for label classifier and domain discriminators, the prediction will

become more accurate.

$$W_k = \frac{1}{n_t} \sum_{x_i \in D_t} C_k \left(F_t \left(x_i \right) \right), \quad k = 1, 2, \ldots, |L_s|, \tag{9}$$

where C_k donates the probability that the target data x_i belongs to label k. Therefore, the greater the value of W_k, the greater the possibility that label k is a valid label, and conversely, the greater the possibility that it belongs to the collection of redundant labels.

In order to further reduce the impact from negative transfer, we take W_k as a constraint to optimize the domain discriminator, and the loss function of D can be improved from equation(8) to equation(10).

$$L'_{advD} = -\sum_{k=1}^{|L_s|} [W_k . (\frac{1}{n_s} \sum_{x_i \in D_s} L_d^k \left(D_k \left(F_s \left(x_i \right) \right), d_i \right) + \frac{1}{n_t} \sum_{x_i \in D_t} L_d^k \left(D_k \left(F_t \left(x_i \right) \right), d_i \right))] \tag{10}$$

It has been shown that adversarial learning is able to excavate domain-invariant features in many existing works. However, in our un-supervised partial transfer learning task, the target labels are missing and there are many irrelevant categories in source domain as mentioned above. For solving this problem, we apply multiple (L_s) domain discriminators to evaluate and adjust the features cross domains. The target mapping is able to obtain a weight for the ith sample associated with the kth category, as equation (9). It is reasonable to believe that the features that make the samples successfully puzzle the discriminators would be useful in target domain. And also, the target sample is more likely to be assigned to No. k category, the adversarial learning enhanced mapping is more effective, as equations (9) and (10).

During the alternative optimization, the feature extractors and domain discriminators will be improved accordingly. By extractors, F_s, and F_t, samples cross domains can be mapped into a space where they can be classified by the classifier. As a result, the per-class features in source domain will be selected in multi-adversarial learning process and those valuable features are enhanced partially, while the ''negative transfer'' are suppressed.

In this paper, we also introduce clustering hypothesis into partial transfer learning. In clustering hypothesis, the processed dataset consists of multiple subsets and each of them is associated with a typical category label. This clustering hypothesis has been successfully applied into many classification tasks [8]. Based on this assumption, the partition boundary between different collections should be far away from the high-density regions. Note that the optimization is highly rely on the predictions on unlabeled data. For the sake of improving the classification accuracy on the unlabeled target dataset, we minimize the conditional entropy of target distribution [60], which can be written as equation (11).

$$E = -\frac{1}{n_t} \sum_{x_i \in D_t} \sum_{k=1}^{|L_s|} C_k \left(F_t \left(x_i \right) \right) . ln C_k \left(F_t \left(x_i \right) \right). \tag{11}$$

It should be pointed out that since the target data is unlabeled, the classifier trained on the source domain can only give an approximate label in advance. As a result, the conditional probability here refers to a pre-determined label probability distribution archived according to the outputs of classifier.

Through minimizing conditional entropy, the prediction on unlabeled target data can be boosted effectively. We make the classifier and discriminators (domain classifiers) obey locally-Lipschitz constraint [60] in order to avoid the non-effective estimation and minimization of conditional entropy. This constraint can also help the classifiers (including domain classifiers) make decision boundaries which will bypass the sample dense areas rather than pass through them. This is the reason why we adopt virtual adversarial training [8] to make constrains on the optimization for classifier and adversarial learning units, as written in equation (12),

$$R = \sum_{k=1}^{|L_s|} [\frac{1}{n_t} \sum_{x \in D_t} \max_{\|r\| \leq \varepsilon} D_{kl}(C_k(F_t(x_i)) || C_k(F_t(x_i + r)))], \tag{12}$$

where ε is a hyper-parameter; $D_{kl}(.)$ represents the KL divergence calculation operation.

Considering all the factors above, the loss function of the whole framework can be written as equation (13).

$$L = L_{advC} + L'_{advD} + E + R \tag{13}$$

The optimization objectives can be written in equations (14) and (15),

$$\left(\widehat{\theta_{F_s}}, \widehat{\theta_{F_t}}, \widehat{\theta_C}\right) = arg \min_{(\theta_{F_s}, \theta_{F_t}, \theta_C)} L \left(\theta_{F_s}, \theta_{F_t}, \theta_C, \theta_{D_k} |_{k=1}^{|L_s|} \right), \tag{14}$$

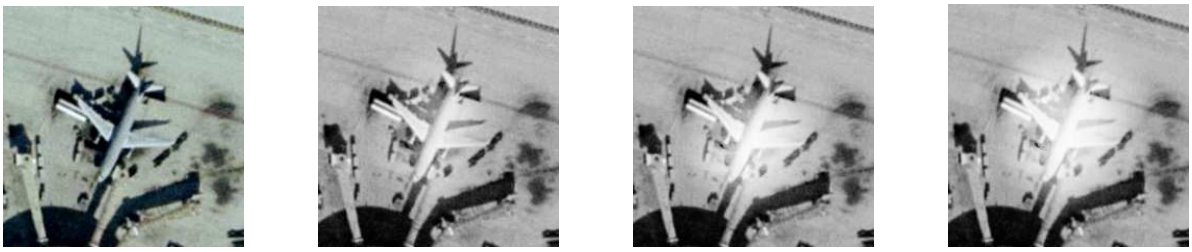
$$\left(\widehat{\theta_{D_1}}, \widehat{\theta_{D_2}}, \ldots, \widehat{\theta_{|L_s|}}\right) = arg \max_{(\theta_{D_1}, \theta_{D_2}, \ldots, \theta_{D_{|L_s|}})} L \left(\theta_{F_s}, \theta_{F_t}, \theta_C, \theta_{D_k} |_{k=1}^{|L_s|} \right), \tag{15}$$

where $\theta_{F_s}, \theta_{F_t}, \theta_C, \theta_{D_k} |_{k=1}^{|L_s|}$ are the parameters that need to be optimized.

It also should be mentioned that in the training of MOAN, we adopt a ''warm boosting'' strategy. That is, the attention proposal network, OANet, comes into play posterior to the adversarial training for multi-GAN units so as to generate credible labels which can provide more effective instruction for attention proposal (OANet mentioned in Section 3.1). These extracted attentions can also provide good instructions for the classifier to handle target data.

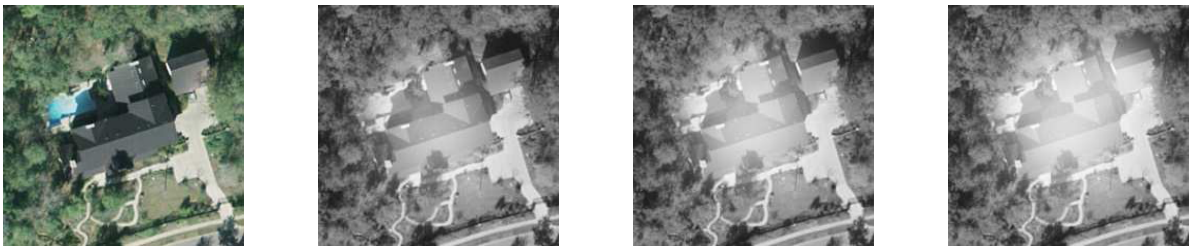
In Figures 5 and 6, we visualize the attention regions during the adversarial training of MOAN on 2 typical pictures. These pictures are randomly selected from another representative data set in remote sensing field, UCM [58]. Similar to Section 3.1, the higher the pixel values are, the brighter the regions will be.

In these 2 pictures, it can be seen that along with the increase of epoches, the extracted attention regions are concentrated near the the main objects, and become more obvious increasingly. This phenomenon shows that there is a



(a) Original (b) Epoch=1000 (c) Epoch=2000 (d) Epoch=3000

FIGURE 5. Attention visualization during the adversarial training (example 1).



(a) Original (b) Epoch=1000 (c) Epoch=2000 (d) Epoch=3000

FIGURE 6. Attention visualization during the adversarial training (example 2).

positive interaction between the multi-adversarial module and OANet.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

To evaluate the effectiveness of MOAN, we conduct comparative experiments on several public data sets with representative transfer learning algorithms in these years. All the experiments are carried out on a server with two Intel Xeon(R) E5-2640K CPUs @2.4GHz, a 64GB RAM and two 11GB GeForce RTX 2080 Ti GPUs. Development environment is pycharm on Ubuntu 16.04.1 system. We adapt TensorFlow 1.12 deep learning framework.

The parameters for training are set as this: batch-size, 50; initial learning-rate, 0.001; epoch, 6000; learning-rate-decay, 0.99; ε in equation (12) is set at 2.1. LeakyRelu is applied as the activation function in order to avoid the vanishing gradient. We also introduce the batch normalization strategy to prevent over-fitting. Resnet-50 [24] pretrained on ImageNet [56] is chosen as the backbone.

B. DATA SETS, BASELINES AND STATE OF THE ART

We carry out 3 groups of experiments in this article. The first group is on daily life pictures, Caltech [57]-Office [63], [64], similar to the experiments in [7] and other state of arts. We handle 4 transfer tasks in these experiments, C-256->W-10, C-256->A-10, C-256->D-10. C denotes Caltech [57] as the source domain. W, A and D are 3 distinct domains with shared label space, Amazon (A), Webcam (W) and DSLR (D) respectively [63]. The transfer is in 10 shared categories.

The second group is conducted on ImageNet (I-1000) [56] and Caltech(C-256) [57]. There are 84 shared categories in these 2 data sets and the transfer tasks can be setting as I-1000->C-84 and C-256->I-84. It should be noted that in order to eliminate the interference from the pre-trained model on ImageNet [56] and keep the rationality of experiment, I-1000 are derived from the training set of ImageNet [56], while I-84 comes from the validation set.

In remote sensing field, we also select 20 categories shared in NWPU-RESISC45 [5] and UCM [58] expect ''buildings'' that only in UCM data set. ''Agricultural'' in UCM is matched with ''Rectangular Farmland'' in NWPU-RESISC45 [5]. Since almost all the categories in UCM [58] can be matched to related ones in NWPU-RESISC45 [5], only N-45->U-20 is evaluated for the partial transfer learning.

Several mainstream deep transfer learning algorithms including DAN [41], DANN(RevGrad) [6], RTN [61], ADDA [3] and SAN [7] are selected as comparisons. We select Resnet-50 [24] pretrained on ImageNet [56] as the backbone for all the algorithms. ResNet [24] has achieved good effectiveness in many classification tasks, and its residual mechanism makes it possible to build very deep networks. The main idea of DAN [41] is to enhance the transferable ability of the specific layers which are embedded in the reproducing kernel hilbert space. There is an optimal process for multi-kernel selection and minimizing the Maximum Mean Discrepancies [65] designed in DAN so as to reduce domain discrepancy effectively. DANN (RevGrad) [6] is the first framework utilizing GAN-based structure for domain adaption. In DANN, a gradient reversal layer is used to connect the feature extractor and domain classifier during the

TABLE 2. Classification accuracies on Caltech-Office tasks with different models.

Method				C-265 → W-10	C-256→A-10	C-256 → D-10	Avg
ResNet-50[26]				65.32%	82.21%	71.53%	73.02%
DAN[43]				49.14%	78.42%	56.63%	61.40%
DANN(RevGrad)[8]				61.26%	80.35%	63.72%	68.44%
RTN[63]				78.52%	83.28%	75.41%	79.07%
ADDA[5]				80.35%	81.37%	80.75%	80.82%
mtUDA[49]				81.32%	81.26%	81.35%	81.31%
SAN[9]				**90.12%**	85.42%	87.63%	87.72%
WAN[64]				90.03%	85.86%	87.82%	87.90%
Ours	CE	VAT	OA				
	×	×	×	89.31%	85.13%	87.26%	87.23%
	✓	×	×	89.66%	85.52%	87.71%	87.63%
	✓	✓	×	89.80%	85.75%	87.92%	87.82%
	×	×	✓	89.75%	85.67%	87.85%	87.76%
	✓	✓	✓	89.83%	**86.13%**	**88.20%**	**88.05%**

TABLE 3. Classification accuracies on ImageNet-Caltech tasks with different models.

Method				I-1000→C-84	C-256→I-84	Avg
ResNet-50[26]				63.15%	58.78%	60.97%
DAN[43]				60.14%	53.21%	56.68%
DANN(RevGrad)[8]				62.73%	54.81%	58.77%
RTN[63]				71.83%	59.75%	65.79%
ADDA[5]				72.96%	61.32%	67.14%
mtUDA[49]				73.26%	61.41%	67.34%
SAN[9]				76.45%	64.74%	70.60%
WAN[64]				76.69%	64.83%	70.76%
Ours	CE	VAT	OA			
	×	×	×	76.13%	64.52%	70.33%
	✓	×	×	76.42%	64.71%	70.57%
	✓	✓	×	76.65%	64.92%	70.79%
	×	×	✓	76.63%	64.86%	70.75%
	✓	✓	✓	**77.20%**	**65.12%**	**71.16%**

backpropagation-based training for the opposite optimization objectives. RTN [61] is the upgrade of DAN [41]. In addition to kernel embedding, new residual layers and loss functions are also integrated in RTN [61]. ADDA [3] introduces the unsymmetrical adversarial mechanism into domain adaption and gives the overall framework which is the basis of some related research. SAN [7] is the first model for partial transfer learning, which can effectively improve the model's effectiveness by eliminating the outlier class in the source domain and minimizing domain differences in shared label space. WAN [62] is the advanced partial transfer learning method with 2 different domain discriminators, the one for discrimination and the other one for the weight generation for samples in feature space.

In addition to adversarial learning based methods above, we also adopt the novel multi-task method mtUDA [47] for comparison. Self-supervised learning method AMDIM [66] is also used in our experiments.

C. STATISTICAL RESULTS AND ANALYSES

Experimental results on different partial transfer learning tasks (as mentioned in section 4.2) are listed in Table 2, Table 3 and Table 4 respectively. In these tables, CE denotes "Conditional Entropy" as mentioned in Section 3.2;

TABLE 4. Classification accuracies on NWPU-UCM task with different models.

Method				N-45→U-20
ResNet-50[26]				78.82%
DAN[43]				71.65%
DANN(RevGrad)[8]				76.34%
RTN[63]				81.56%
ADDA[5]				85.92%
mtUDA[49]				86.32%
AMDIM[68]				88.13%
SAN[9]				88.35%
WAN[64]				88.63%
Ours	CE	VAT	OA	
	×	×	×	88.16%
	✓	×	×	88.42%
	✓	✓	×	88.71%
	×	×	✓	88.63%
	✓	✓	✓	**90.58%**

VAT denotes "Virtual Adversarial Training"; OA denotes "Object-level Attention" as mentioned in Section 3.1.

On Caltech-Office tasks, as listed in Table 2, the performances of DAN [41] and DANN(RevGrad) [6] are weaker than basic ResNet50 [24]. On C-265 -> W-10 and C-256 -> D-10 tasks, these 2 methods gain extremely low accuracies. This is probably because that DAN [41] and

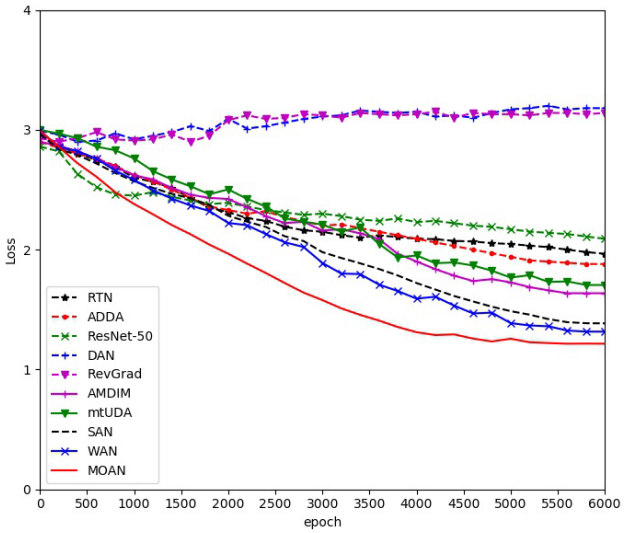


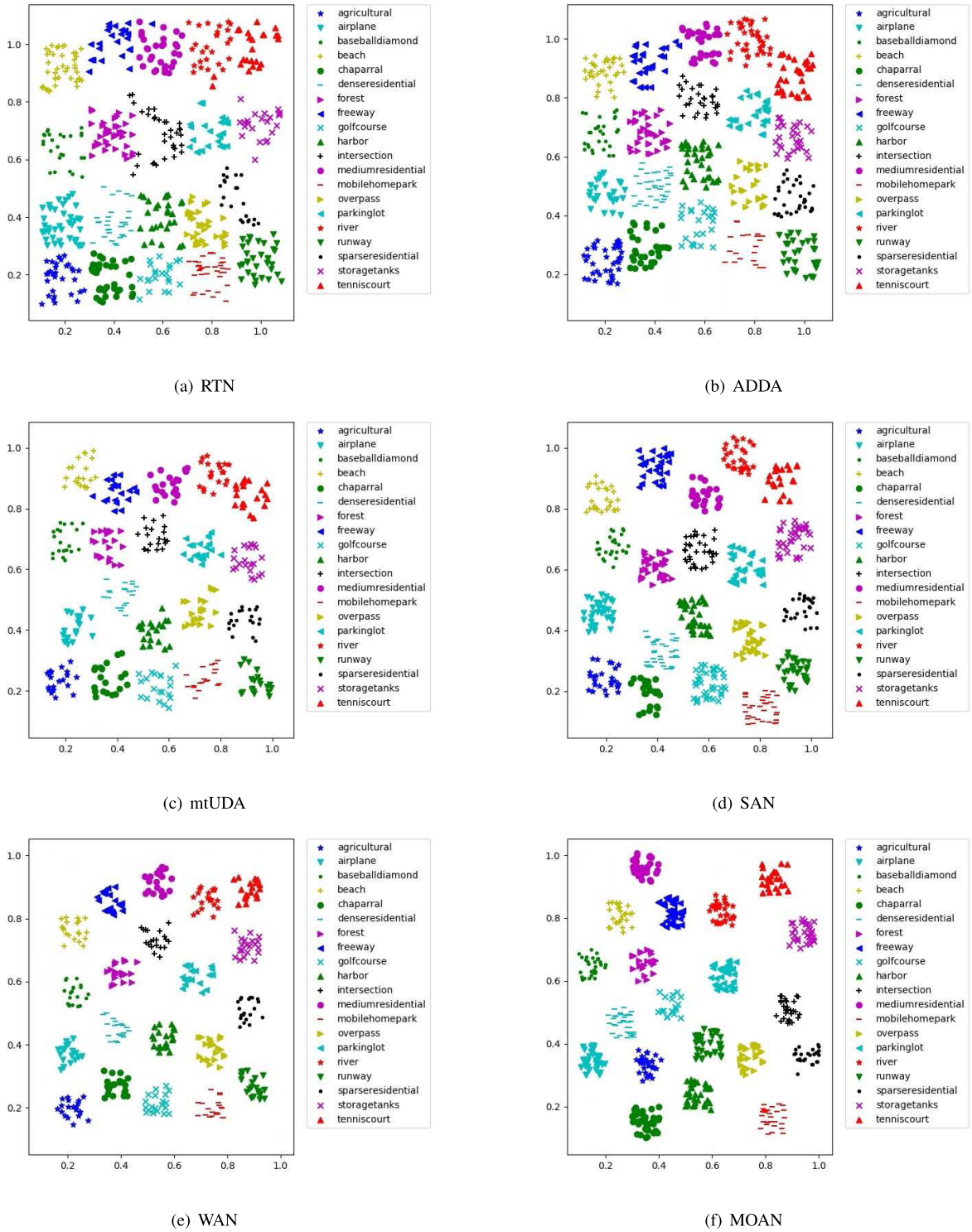
FIGURE 7. Convergence curves on NWPU-UCM task.

DANN(RevGrad) [6] are proposed for the symmetric domain adaption in which the label spaces in source and target domains are identical. For partial transfer learning with much noises, the specially designed structures in these methods are more susceptible to interferences. For example, In DAN [41], the kernel-based embedding and MMD-based (Maximum Mean Discrepancies-based) cross domain evaluation may amplify the negative effects from those irrelevant samples. Similarly, in DANN(RevGrad), there is only one feature extractor for all the samples in both domains. In partial transfer learning with huge source samples, the extractor cannot map the target sample effectively and the domain discriminator cannot distinguish the domain, too. Due to the fact that in Amazon pictures (A10 classification), the objects are obvious with fewer noises associated to source domain, these methods are not so weak and have similar accuracies to RTN [61] and ADDA [3]. The popular domain adaption algorithm ADDA [3] achieves an average accuracy at 80.82%. In ADDA [3], there is only one adversarial learning unit. Due to the introduction of multi-adversarial learning-based category selection mechanism, SAN [7] surpasses ADDA [3] with a huge advantage of 6.9%. The sample weighting strategy in WAN [62] makes it have a better average accuracy than SAN. Although MOAN is lower than SAN in C-256 -> W-10 task with 0.29%, it exceeds other methods in average accuracy.

On Image-Caltech tasks, there is a huge difference in the number of categories between source and target domains in which the shared categories are only 84. It is obvious that the outlier data in source domain are much more than the related data. This situation will lead to a serious negative transfer disturbed by unrelated signals. As a result, the accuracies of models on these tasks are a little lower than those on Caltech-Office tasks. It can be seen that in Table 3, the performances of all the methods on I-1000 -> C-84 task are better than C-256 -> I-84. This revels that the labeled source data are important for training feature extractors. Recent adversarial learning-based methods (like ADDA [3]) have higher accuracies than classical methods. The multi-task method, mtUDA [47], performs better than ADDA [3] on both 2 tasks. Those methods proposed for partial transfer learning, SAN and WAN [62], outperform other existing methods. On I-1000-> C-84 task, MOAN obtains an accuracy at 77.20%, and its average accuracy exceeds RTN [61], ADDA [3] and SAN [7] by 5.37%, 4.02% and 2.23% respectively. This indicates that MOAN has a better ability to eliminate interference from irrelevant categories effectively and prevent negative transfer.

In terms of remote sensing data set, the performances of DAN [41] and DANN(RevGrad) **on NWPU-UCM tasks** are inferior to that obtained by the baseline, ResNet50, and also weaker than RTN [61]. The trend is similar to the results listed in Tables 2 and 3. mtUDA [47] outperforms ADDA [3] with 0.4% higher. The partial transfer methods like SAN [7], WAN [62] and our MOAN are designed with special GAN-based structures to evaluate the samples and features, therefore showing obvious advantages in processing these partial transfer problems. As listed in Table 4, MOAN also shows better performance with higher accuracy than RTN [61], ADDA [3], mtUDA [47], SAN [7] and WAN [62] by 9.02%, 4.66%, 4.26%, 2.23% and 1.95% respectively. It is obvious MOAN is very suitable for the unsupervised remote sensing scene classification tasks. *It should also be noticed that the main idea of self-supervised methods is to learn a strong encoder with massive un-labeled samples, than train basic classifiers in a relative small labeled sample set without backpropagation to the encoder. Since the structure of basic classifier is relative simple, the classifier can be obtained by very limited samples. In transfer task, the encoder and classifier are trained by different data sets, e.g., training encoder on ImageNet [56] and MLP (Multi-Layer Perception) classifier on Places205 [66]. In the testing, the Places205 [67] samples are mapped by the encoder trained on ImageNet [56]. In this article, we also design experiments for the self-supervised method AMDIM [66]: training encoder on NWPU [5] and training the classifier on UCM [58]. We use unlabeled NWPU [5] and labeled UCM [58], according to the self-supervised idea.* AMDIM [66] obtains a similar accuracy to SAN [7], higher than mtUDA [47]. We believe that more samples can further improve the encoder in self-supervised style. MOAN outperforms AMDIM [66] in this task.

With regard to different components in MOAN, it can be seen that the introduction of Virtual Adversarial Training (VAT) has improved the average accuracies by 0.19%, 0.22% and 0.29% respectively on all 3 groups of experiments. OANet has further enhanced the average accuracies by 0.53%, 0.42% and 0.47%. **From the perspective of model structure**, we find that the adversarial learning-based models (e.g. ADDA [3], SAN [7] and WAN [62]) perform better than other models (like DAN [41] and Resnet-50 [24]). The models with multi-adversarial units (like SAN [7], WAN [62]



(a) RTN (b) ADDA

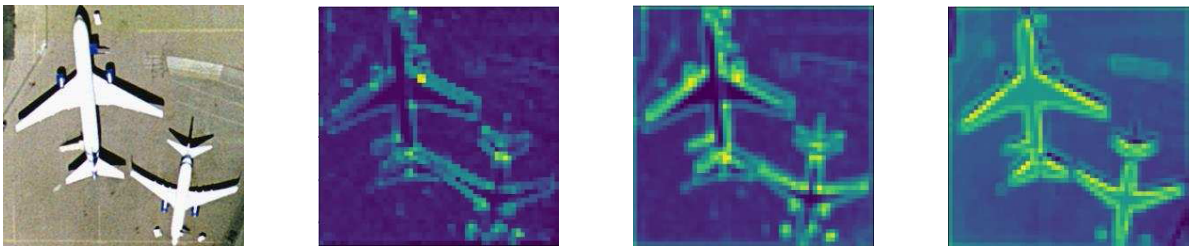
(c) mtUDA (d) SAN

(e) WAN (f) MOAN

FIGURE 8. t-SNE visualization of RTN, ADDA, SAN and MOAN with class information on UCM data set.

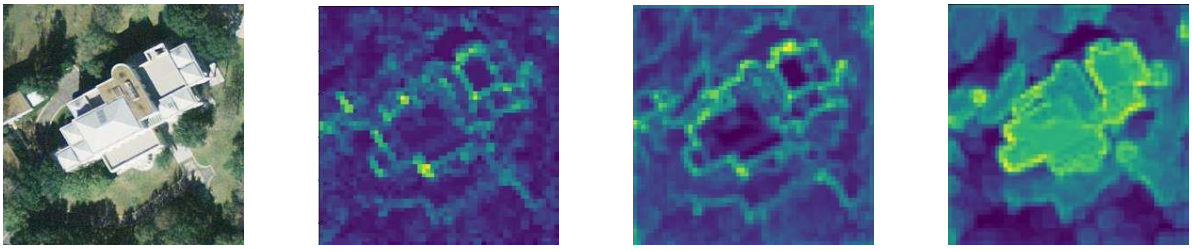
and MOAN) for feature evaluation are superior to those with single adversarial unit (e.g. DANN(RevGrad) [6] and ADDA [3]). The multi-task method mtUDA [47] and self-supervised method ADDIM also obtain acceptable accuracies. Although self-supervised methods have advantage in feature representation, they need an additional step to train classifiers with semantic labels.

It can be concluded that MOAN performs better than other methods on partial transfer learning tasks in an unsupervised style. In these tasks, the source data set is large and



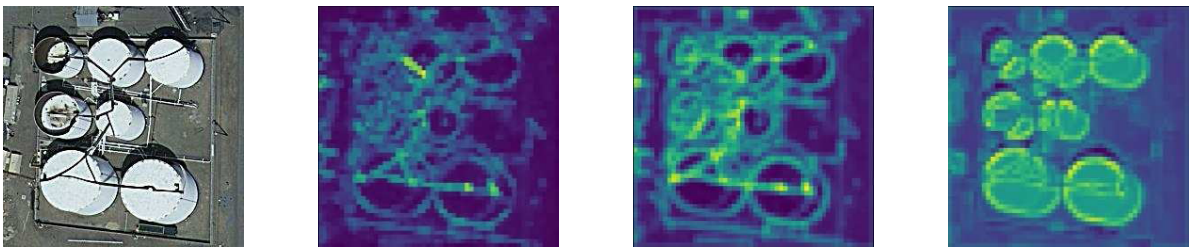
(a) Original (b) Before (c) After (Res-Attention) (d) After (OANet)

FIGURE 9. Feature visualization on "airplane40" picture.



(a) Original (b) Before (c) After (Res-Attention) (d) After (OANet)

FIGURE 10. Feature visualization on "sparseresidential70" picture.

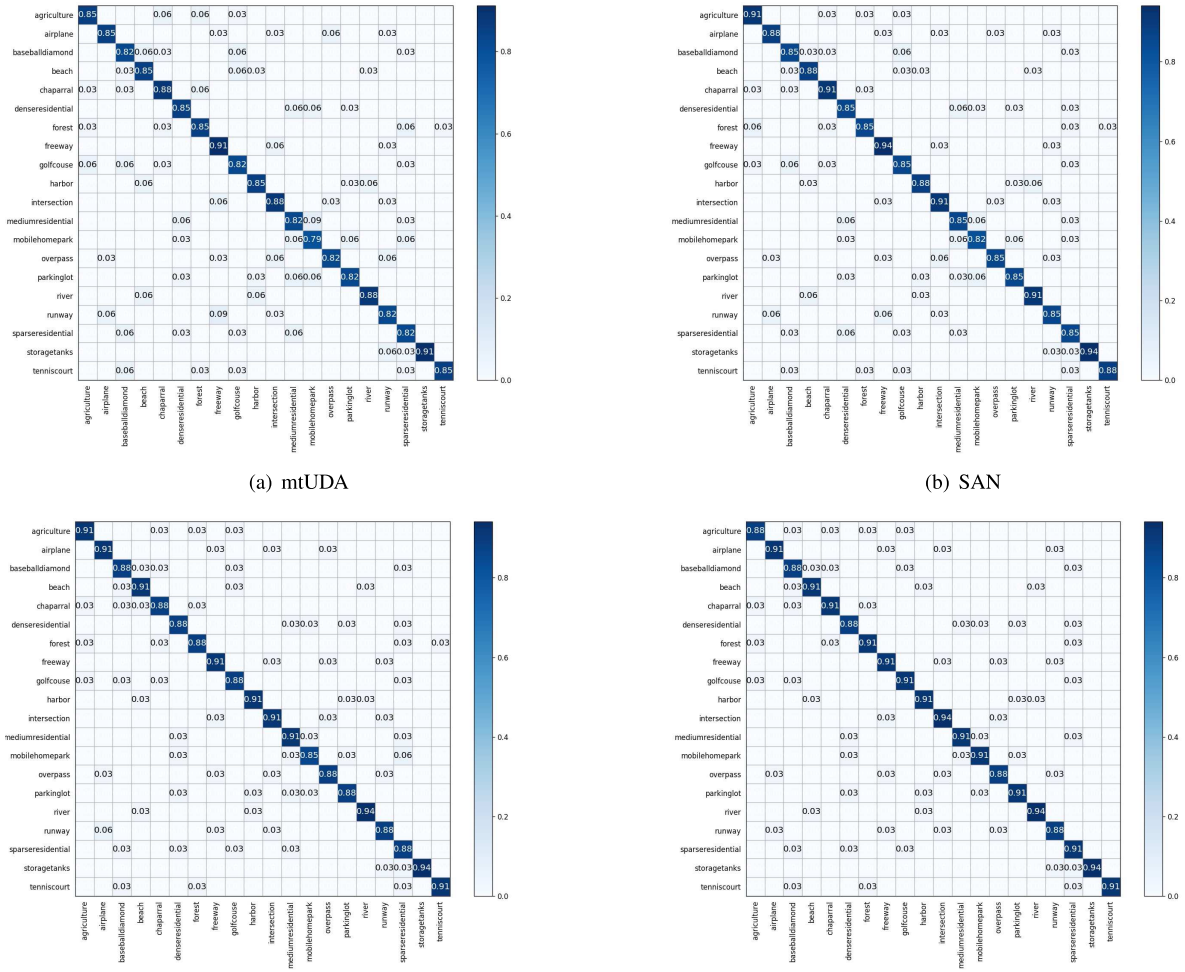


(a) Original (b) Before (c) After (Res-Attention) (d) After (OANet)

FIGURE 11. Feature visualization on "storagetanks27" picture.

complex and the target data set is relatively small. In remote sensing field, the advantage is more obvious as shown in Table 4. In these experiments, we can also easily find that the cross-domain noises impact negatively on the models' performances especially for those symmetrical domain adaption methods like DANN(RevGrad) [6] and ADDA [3]. It is a valuable research topic to modify existing methods so as to handle partial transfer tasks.

The loss curves of different methods on remote sensing task, NWPU-UCM, are shown in Figure 7. As the number of epoches increases, the losses of DAN [41] and DANN(RevGrad) [6] increase instead of falling. This is probably because that these 2 methods are mainly designed for the symmetric transfer tasks in which the label spaces are same in different domains. In DAN [41], the kernel-based embedding and MMD-based domain adaption cannot map the target samples effectively in the partial transfer tasks. In DANN [6], the gradient reversal layer may bring instability during the training of feature extractor and discriminator, and the instability is amplified by the negative transfer resulted from unrelated categories. Although DANN(RevGrad) [6] adopts adversarial structure, its performance is inferior to the backbone. This can also verify the importance of limiting negative transfer and promoting positive transfer. It can also be seen that the loss curve of RTN [61] tends to decrease generally, possibly because the method introduces entropy minimization constraint. ADDA [3] weakens the influence of negative transfer to some extent by mapping the target samples to the source space and adopting the alternative training method. This is the reason why ADDA [3] can converge, though the final loss value is still at a high level. SAN [7] exceeds the above methods both in the convergence rate and value, showing that the selective multi-adversarial structure can effectively avoid the interference of unrelated categories. mtUDA [47] and AMDIM [66] have similar downtrends with lower loss values than ADDA [3]. The adversarial learning-based partial transfer learning methods WAN [62] and SAN [7] have better convergence curves, falling faster with lower loss values. MOAN obtains the lowest loss value among all the models.



(a) mtUDA (b) SAN

(c) WAN (d) MOAN

FIGURE 12. Confusion matrixes of SAN and MOAN on UCM data set.

D. FEATURE VISUALIZATION AND CONFUSION MATRIXES

For NWPU-UCM task, **we visualize the t-SNE [68] embeddings** of samples in target domain in order to show the impacts of feature enhancement on remote sensing images. 5 models are used for comparison with MOAN, as shown in Figure 8.

It can be seen that RTN [61] cannot clearly distinguish the different categories. Although ADDA [3] adopts adversarial structure, it do not discriminate those target categories, too. Similar to the analyses in section IV.C, symmetrical domain adaption methods are more susceptible to noises from unrelated categories in partial transfer learning problems. Compared with ADDA [3], mtUDA [47] can generate more obvious regional boundaries. Since SAN [7], WAN [62] and MOAN are proposed for partial transfer problems and design different modules for accurate feature selection, their target samples are more concentrated with larger inter-class distances. MOAN performs better than other existing methods.

To illustrate the performance of OANet in feature extraction, **we also visualize the feature maps** of 3 images from UCM [58] data set, as shown in Figures 9, 10 and 11. Residual Attention network (Res-Attention) [69], the representative and effective attention mechanism, is selected as a comparison.

We can easily observe that after the processing by object-level attention network, in the maps, the features associated with the main objects (such as airplanes, buildings and oil tanks) are enhanced effectively. The attention areas are brighter after mapping, while the unrelated parts are weakened by the model. In addition, the overall structure of the object is more prominent. It is reasonable to believe that this advantage can help MOAN perform better in global scene recognition.

Figure 12 shows the confusion matrixes obtained by mtUDA [47], SAN [7], WAN [62] and MOAN. We use 2 significant digits. It can also be observed that MOAN has more excellent scene classification ability in remote sensing field, especially for those easily-confused categories like ‘‘mobile home park’’, ‘‘parking lot’’ and ‘‘residential’’. From (a) to (d), the accuracies increases gradually. On ‘‘mobile home park’’,

the accuracy obtained by MOAN is higher than mtUDA [47], SAN [7] and WAN [62] by 0.12%, 0.09% and 0.06% respectively. On ''parking lot'', the differences are 0.09%, 0.06% and 0.03%.

V. CONCLUSION

In this paper, we propose a novel unsupervised partial transfer learning framework, MOAN, for remote sensing scene classification. Different from conventional image recognition methods, object-level attention mechanism and multi-adversarial learning model are embedded in the framework.

With the ability to circumvent negative transfer by selecting out the irrelevant source data, MOAN can obtain higher accuracies in target domain than existing domain adaption methods (e.g. RTN [61]) in un-supervised partial transfer learning problems. Different from those state-of-the-art adversarial domain adaption methods like ADDA [3] and SAN [7], in MOAN, object-level attentions are improved and optimized alternatively in order to completely perceive the objects in pictures. This is the key point to boosting the performance on unsupervised remote sensing scene classification. Considering that most of existing attention proposal strategies are highly dependent on labels, it is a valuable attempt to propose attention regions by ''pseudo tags'' which are obtained during the multi-adversarial learning-based domain adaption. These attention regions can also promote predicted tags in target domain. Moreover, the virtual adversarial training mechanism is introduced in the training for adversarial learning units so as to keep the stability of the model and facilitate a positive, effective transfer.

Experimental results on 3 groups of transfer learning tasks show that our MOAN can significantly promote positive transfer and weaken negative transfer, thereby achieving higher accuracies. MOAN is valid in both daily life pictures and remote sensing images.

In the future, MOAN can be further improved in 4 aspects: 1) in addition to ''pseudo'' labels, ''proxy tasks'' like color prediction and relative position estimation can also be utilized to embed visual signals in a self-supervised multi-task style; 2) the proposed method is expected to obtain more guidance information from multiple feature embedding strategies like multi-view learning [70] and multi-modal learning [71]; 3) with high quality visual representations from 1) and 2), MOAN can be modified and applied in other complex remote sensing image archive, e.g., Bigearthnet [72]; 4) the proposed method can also be improved and used to deal with more complex spatial applications, like semantic segmentation.

REFERENCES

[1] Y. Ganin and V. S. Lempitsky, ''Unsupervised domain adaptation by backpropagation,'' in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 1180–1189. [Online]. Available: http://proceedings.mlr.press/v37/ganin15.html

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, ''Generative adversarial nets,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets

[3] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, ''Adversarial discriminative domain adaptation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2962–2971, doi: 10.1109/CVPR.2017.316.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[5] 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, doi: 10.1109/JPROC.2017.2675998.

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, ''Domain-adversarial training of neural networks,'' *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 59:1–59:35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[7] Z. Cao, M. Long, J. Wang, and M. I. Jordan, ''Partial transfer learning with selective adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2724–2732.

[8] R. Shu, H. H. Bui, H. Narui, and S. Ermon, ''A DIRT-T approach to unsupervised domain adaptation,'' in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=H1q-TM-AW

[9] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, ''Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image,'' *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019, doi: 10.1109/TCYB.2018.2810806.

[10] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, ''Simultaneous spectral-spatial feature selection and extraction for hyperspectral images,'' *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018, doi: 10.1109/TCYB.2016.2605044.

[11] Y. Yuan, J. Lin, and Q. Wang, ''Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization,'' *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016, doi: 10.1109/TCYB.2015.2484324.

[12] A. Pandey and D. Wang, ''A new framework for CNN-based speech enhancement in the time domain,'' *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019, doi: 10.1109/TASLP.2019.2913512.

[13] S. Wang, M. Huang, and Z. Deng, ''Densely connected CNN with multi-scale feature attention for text classification,'' in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 4468–4474, doi: 10.24963/ijcai.2018/621.

[14] Y. Wu, B. Jiang, and N. Lu, ''A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 10, pp. 2108–2118, Oct. 2019, doi: 10.1109/TSMC.2017.2757264.

[15] Y. Wu, B. Jiang, N. Lu, H. Yang, and Y. Zhou, ''Multiple incipient sensor faults diagnosis with application to high-speed railway traction devices,'' *ISA Trans.*, vol. 67, pp. 183–192, Mar. 2017.

[16] Y. Wu, B. Jiang, and Y. Wang, ''Incipient winding fault detection and diagnosis for squirrel-cage induction motors equipped on CRH trains,'' *ISA Trans.*, vol. 18, no. 9, pp. 1–8, 2019, doi: 10.1016/j.isatra.2019.09.020.

[17] P. Chen, P. Li, Q. Li, and D. Zhang, ''Semi-supervised fine-grained image categorization using transfer learning with hierarchical multi-scale adversarial networks,'' *IEEE Access*, vol. 7, pp. 118650–118668, 2019, doi: 10.1109/ACCESS.2019.2934476.

[18] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, ''Structured optimal graph based sparse feature extraction for semi-supervised learning,'' *Signal Process.*, vol. 170, May 2020, Art. no. 107456.

[19] Z. Jiang, Q. Wang, and Y. Yuan, ''Modeling with prejudice: Small-sample learning via adversary for semantic segmentation,'' *IEEE Access*, vol. 6, pp. 77965–77974, 2018, doi: 10.1109/ACCESS.2018.2884502.

[20] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, ''Zero-shot learning by convex combination of semantic embeddings,'' in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1312.5650

[21] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, ''Towards better exploiting convolutional neural networks for remote sensing scene classification,'' *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017, doi: 10.1016/j.patcog.2016.07.001.

[22] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 44–51, doi: 10.1109/CVPRW.2015.7301382.

[23] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017, doi: 10.1109/LGRS.2017.2657778.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[26] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015, doi: 10.1109/LGRS.2015.2483680.

[27] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017, doi: 10.1109/TGRS.2017.2700322.

[28] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017, doi: 10.1109/LGRS.2017.2731997.

[29] Z. Liu, J. Wang, G. Liu, and L. Zhang, "Discriminative low-rank preserving projection for dimensionality reduction," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105768.

[30] Z. Liu, X. Wang, J. Pu, L. Wang, and L. Zhang, "Nonnegative low-rank representation based manifold embedding for semi-supervised learning," *Knowl.-Based Syst.*, vol. 136, pp. 121–129, Nov. 2017, doi: 10.1016/j.knosys.2017.09.003.

[31] B. Du, M. Zhang, L. Zhang, R. Hu, and D. Tao, "PLTD: Patch-based low-rank tensor decomposition for hyperspectral images," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 67–79, Jan. 2017, doi: 10.1109/TMM.2016.2608780.

[32] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks

[33] Y. Yuan and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*. [Online]. Available: http://arxiv.org/abs/1809.00916

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[35] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[36] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019, doi: 10.1109/TGRS.2019.2924818.

[37] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3492–3507, Sep. 2019.

[38] J. Wang, L. Shen, W. Qiao, Y. Dai, and Z. Li, "Deep feature fusion with integration of residual connection and attention model for classification of VHR remote sensing images," *Remote Sens.*, vol. 11, no. 13, p. 1617, Jul. 2019, doi: 10.3390/rs11131617.

[39] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo, "SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention," *Remote Sens.*, vol. 11, no. 14, p. 1702, Jul. 2019, doi: 10.3390/rs11141702.

[40] B. Fang, R. Kou, L. Pan, and P. Chen, "Category-sensitive domain adaptation for land cover mapping in aerial scenes," *Remote Sens.*, vol. 11, no. 22, p. 2631, Nov. 2019, doi: 10.3390/rs11222631.

[41] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 97–105. [Online]. Available: http://proceedings.mlr.press/v37/long15.html

[42] Z. Peng, S. Feng, and T. Lee, "Adversarial multi-task deep features and unsupervised back-end adaptation for language recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 5961–5965, doi: 10.1109/ICASSP.2019.8682303.

[43] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, vol. 1, Jul./Aug. 2017, pp. 1–10, doi: 10.18653/v1/P17-1001.

[44] E. Meyerson and R. Miikkulainen, "Modular universal reparameterization: Deep multi-task learning across diverse domains," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 7901–7912. [Online]. Available: http://papers.nips.cc/paper/9004-modular-universal-reparameterization-deep-multi-task-learning-across-diverse-domains

[45] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016, doi: 10.1109/TIP.2015.2495116.

[46] L. Tóth and G. Gosztolya, "Reducing the inter-speaker variance of CNN acoustic models using unsupervised adversarial multi-task training," in *Proc. 21st Int. Conf. Speech Comput.*, Istanbul, Turkey, Aug. 2019, pp. 481–490, doi: 10.1007/978-3-030-26061-3_49.

[47] J. Zhang, W. Li, and P. Ogunbona, "Unsupervised domain adaptation: A multi-task learning-based method," *Knowl.-Based Syst.*, vol. 186, Dec. 2019, Art. no. 104975, doi: 10.1016/j.knosys.2019.104975.

[48] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-

[49] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 649–666, doi: 10.1007/978-3-319-46487-9_40.

[50] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 69–84, doi: 10.1007/978-3-319-46466-4_5.

[51] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2536–2544, doi: 10.1109/CVPR.2016.278.

[52] J. N. Kundu, N. Lakkakula, and V. B. Radhakrishnan, "UM-Adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1436–1445.

[53] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 762–771.

[54] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2070–2079, doi: 10.1109/ICCV.2017.226.

[55] B. C. Song, D. H. Kim, and S. H. Lee, "Metric-based regularization and temporal ensemble for multi-task learning using heterogeneous unsupervised tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2903–2912.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPRW.2009.5206848.

[57] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CaltechAUTHORS:CNS-TR-2007-001, 2007. [Online]. Available: https://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001

[58] 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. (GIS)*, San Jose, CA, USA, Nov. 2010, pp. 270–279, doi: 10.1145/1869790.1869829.

[59] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 3934–3941.

[60] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2004, pp. 529–536. [Online]. Available: http://papers.nips.cc/paper/2740-semi-supervised-learning-by-entropy-minimization

[61] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 136–144. [Online]. Available: http://papers.nips.cc/paper/6110-unsupervised-domain-adaptation-with-residual-transfer-networks
[62] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8156–8164.
[63] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 213–226, doi: 10.1007/978-3-642-15561-1_16.
[64] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073, doi: 10.1109/CVPR.2012.6247911.
[65] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," 2012, *arXiv:1207.6076*. [Online]. Available: http://arxiv.org/abs/1207.6076
[66] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 15509–15519. [Online]. Available: http://papers.nips.cc/paper/9686-learning-representations-by-maximizing-mutual-information-across-views
[67] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec 2014, pp. 487–495. [Online]. Available: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database
[68] G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[69] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458, doi: 10.1109/CVPR.2017.683.
[70] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognit.*, vol. 48, no. 10, pp. 3102–3112, Oct. 2015, doi: 10.1016/j.patcog.2014.12.016.
[71] P. Li, P. Chen, Y. Xie, and D. Zhang, "Bi-modal learning with channel-wise attention for multi-label image classification," *IEEE Access*, vol. 8, pp. 9965–9977, 2020, doi: 10.1109/ACCESS.2020.2964599.
[72] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Yokohama, Japan, Jul./Aug. 2019, pp. 5901–5904, doi: 10.1109/IGARSS.2019.8900532.

PENG LI received the B.E. degree from Zhengzhou University, Zhengzhou, China, in 2011, and the M.S. degree in computer science and technology from the University of Science and Technology Beijing, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree in computer science and technology. He was an Engineer with Shanghai Research and Development Department, Agricultural Bank of China, Shanghai, China, from 2016 to 2018. His research interests include computational intelligence technologies in remote sensing image processing and computer vision.

DEZHENG ZHANG received the Ph.D. degree in control theory and its applications from the University of Science and Technology Beijing, Beijing, China, in 2002. He is currently a Professor and a Ph.D. Supervisor at the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. He is also the Director of the Beijing Key Laboratory of Knowledge Engineering for Materials Science. His currently research interests include machine learning and knowledge engineering.

PENG CHEN received the Ph.D. degree in control science and engineering from the University of Science and Technology Beijing, Beijing, China, in 2020. He was an Engineer at the National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China, from 2012 to 2015. His current research interests include machine learning, intelligent optimization, and knowledge engineering.

XIN LIU received the B.S. degree in electronic information engineering and the Ph.D. degree in control science and engineering from the University of Science and Technology Beijing, China, in 2009 and 2015, respectively. From 2015 to 2017, she held a postdoctoral position in information and communication systems. Since 2017, she has been an Assistant Professor with the Computer Science and Technology Department, School of Computer and Communication Engineering, University of Science and Technology Beijing. From 2019 to 2020, she was a Visiting Scholar with the Department of Surgery, University of Alberta. Her research interests include high-level data analysis, intelligent information processing, data mining, and decision making.

AZIGULI WULAMU received the Ph.D. degree in control theory and its applications from the University of Science and Technology Beijing, Beijing, China, in 2004. She was a Visiting Scholar with Technische Universität Kaiserslautern, Germany, from August 2001 to October 2002. She is currently a Professor and a Ph.D. Supervisor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interests include knowledge engineering, knowledge graph, deep learning, and artificial intelligence technologies.