Stagewise Unsupervised Domain Adaptation With Adversarial Self-Training for Road Segmentation of Remote-Sensing Images

Lefei Zhang, *Senior Member, IEEE*, Meng Lan, Jing Zhang, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Road segmentation from remote-sensing images is a challenging task with wide ranges of application potentials. Deep neural networks have advanced this field by leveraging the power of large-scale labeled data, which, however, are extremely expensive and time-consuming to acquire. One solution is to use cheap available data to train a model and deploy it to directly process the data from a specific application domain. Nevertheless, the well-known domain shift (DS) issue prevents the trained model from generalizing well on the target domain. In this article, we propose a novel stagewise domain adaptation model called RoadDA to address the DS issue in this field. In the first stage, RoadDA adapts the target domain features to align with the source ones via generative adversarial networks (GANs)-based interdomain adaptation. Specifically, a feature pyramid fusion module is devised to avoid information loss of long and thin roads and learn discriminative and robust features. Besides, to address the intradomain discrepancy in the target domain, in the second stage, we propose an adversarial self-training method. We generate the pseudo labels of the target domain using the trained generator and divide it to labeled easy split and unlabeled hard split based on the road confidence scores. The features of hard split are adapted to align with the easy ones using adversarial learning and the intradomain adaptation process is repeated to progressively improve the segmentation performance. Experiment results on two benchmarks demonstrate that RoadDA can efficiently reduce the domain gap and outperforms state-of-the-art methods. The code is available at https://github.com/LANMNG/RoadDA.**

***Index Terms*— Remote sensing (RS), road segmentation, self-training, unsupervised domain adaptation (UDA).**

I. Introduction

ROAD segmentation from remote-sensing (RS) images is a crucial research topic in RS field. It aims to separate the road areas from the complex background and assign the right label to each pixel of the whole image. Road segmentation has many important applications, such as vehicle navigation [1], urban planning [2], [3], disaster assistance [4], and so on. Various methods have been proposed to effectively address the road segmentation task. However, the unsupervised learning-based methods usually meet low accuracy for depending on the predefined features [1], [5]. As the rapid development of deep learning technique, supervised learning-based convolutional neural networks (CNNs) have shown excellent ability of feature extraction and have been applied to many RS image-processing tasks [6]–[10]. Especially, with the aid of deep learning, road segmentation of RS images has made great progress [11], [12].

Many CNN-based road segmentation approaches achieve good performance on the public RS benchmark datasets. Cheng *et al.* [13] proposed a cascaded encoder–decoder network for road segmentation and evaluated it on the collected road dataset which contains 224 manually labeled Google Earth RGB images. Yang *et al.* [3] designed an recurrent convolutional neural network (RCNN) unit with shared weights of convolutional filters to build a deeper network. The model was test on the RoadTracer dataset (RTDS) which obtains the satellite images from Google Map labeled with OpenStreeMap (OSM) [14]. Zhou *et al.* [15] introduced the D-LinkNet and won the first prize of Road Extraction Challenge in DeepGlobe 2018, in which the released road dataset has 6226 DigitalGlobe RGB training images. It is exciting to witness various approaches to achieve better performance on different road scenarios, whereas these supervised learning-based road segmentation methods always require a large amount of labeled data to train the model and it could be extremely expensive and time-consuming to manually label the road areas from a lot of unlabeled high-resolution RS images in the practical applications.

Fortunately, we have access to some public road datasets released by previous researchers and abundant digital map resources, such as Google Maps and OpenStreetMap, which can be used to create labeled road dataset with minimum efforts. These data may help us to complete the road segmentation task on unlabeled target domain. However, as shown in Fig. 1(a1) and (b1), directly applying the road segmentation model, which is trained on the source domain, to infer the target road images may encounter significant performance drop. It is caused by the domain shift (DS) between training

Manuscript received February 27, 2021; revised May 20, 2021 and July 17, 2021; accepted August 7, 2021. Date of publication August 18, 2021; date of current version January 31, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62076188, in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, in part by the Fundamental Research Funds for the Central Universities under Grant 2042021kf0196, and in part by the supercomputing system in the Supercomputing Center of Wuhan University. The work of Jing Zhang was supported by Australian Research Council (ARC) under Project FL-170100117. *(Corresponding author: Lefei Zhang.)*

Lefei Zhang and Meng Lan are with the School of Computer Science, Institute of Artificial Intelligence, Wuhan University, Wuhan 430072, China (e-mail: zhanglefei@whu.edu.cn; menglan@whu.edu.cn).

Jing Zhang is with the School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: jing.zhang1@sydney.edu.au).

Dacheng Tao is with JD Explore Academy, 102600, China (e-mail: dacheng.tao@gmail.com).

Digital Object Identifier 10.1109/TGRS.2021.3104032

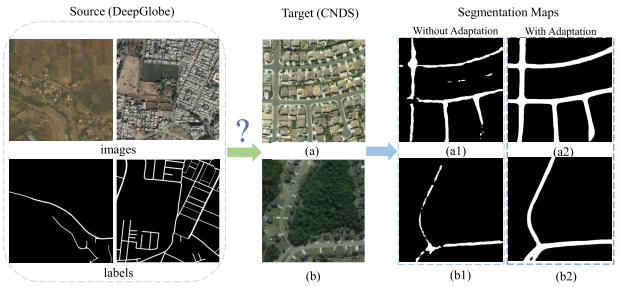


Fig. 1. How to exploit the available labeled road data (source domain) to improve the segmentation performance of unlabeled road data (target domain) are of practical importance. (a1) and (b1) Results obtained by directly applying the segmentation model trained on source domain to two target domain images. (a2) and (b2) Results of our model.

images and test images, since these images may have different types of road surfaces (unpaved, paved, dirt roads), rural and urban background areas, and so on. Hence, it is important and meaningful to find a suitable technique to address the DS issue and take advantage of the available labeled road data to conduct automated road segmentation on target RS images without manual labeling.

Recently, unsupervised domain adaptation (UDA) [16], [17] is introduced to solve this problem. UDA utilizes labeled data in one or more relevant source domains to execute new tasks in a target domain with unlabeled data and aims to mitigate the DS. Although UDA has been widely studied in RS image classification task [18], [19], few efforts have been made in the context of road segmentation. The recent work adversarial spatial pyramid network (ASPN) [20] introduces the random noise to the source domain as synthetic data and performs the adversarial domain adaptation on the output space of the source and target domains. However, the existing UDA methods for road segmentation neither fine-tune the segmentation model on the target data with pseudo labels for better performance nor tackle the possible intradomain discrepancy within the target domain caused during data collection. As a specific technique of semisupervised learning, self-training [21] could be a useful strategy for UDA. Specifically, it generates pseudo labels of unlabeled target data from the classifier and uses them to fine-tune the model, therefore adapting the model to the target domain. In addition, the adapted model can be used to update the pseudo labels and this procedure can be repeated to improve the performance.

Inspired by these techniques, we propose a novel stagewise UDA framework called RoadDA for road segmentation of RS images. RoadDA consists of two stages, namely the interdomain adaptation and the adversarial self-training, as illustrated in Fig. 2. In the first stage, given the labeled source data and unlabeled target data, a segmentation model serves as the generator to produce the predictions, while an interdomain discriminator predicts the domain labels of these predictions. The target domain is adapted to align with the source domain at the output level by optimizing the segmentation loss of source domain and the interdomain adversarial loss. Besides, a feature pyramid fusion module (FPFM) is devised to avoid information loss of long and thin roads and learn discriminative and robust features. In the second stage, to address the intradomain discrepancy in the target domain and further improve the segmentation performance, we design an adversarial self-training scheme. The segmentation model trained in the previous stage is utilized to generate pseudo labels of the target domain images. Since road information is the most important prior knowledge in road segmentation task, we pay more attention to the quality of road prediction of the pseudo labels. Therefore, these pseudo labels are checked by a quality estimator according to the confidence score of the road pixels. The retained pseudo labels along with their corresponding images are regarded as the easy split, while the left images in the target domain are regarded as the unlabeled hard split. Similarly, we use the same technique in the first stage to mitigate the intradomain gap by adapting the distribution of the hard split to that of the easy split. Moreover, we devise a progressive training scheme to iteratively update the pseudo labels from the trained segmentation model and then use them to retrain the segmentation model. During the inference phase, only the adapted segmentation model of the second stage is used for road segmentation without any extra computational requirements.

In summary, we conclude our contributions as follows.

1) We propose a novel stagewise UDA framework for road segmentation of RS images, which exploits available labeled road data to achieve promising road segmentation on unlabeled target images. This work could be useful in practical application scenarios.
2) We devise a new FPFM to avoid the information loss of long and thin roads in high-level feature maps and learn discriminative and robust features for better performance.
3) In addition to addressing the interdomain discrepancy between the source and target domains, we propose an adversarial self-training stage to mitigate the intradomain discrepancy in the target domain. It can progressively achieve intradomain adaptation and improve road segmentation performance on target domain.

The rest of the article is organized as follows. Section II briefly reviews the related work. In Section III, we describe the details of our proposed RoadDA model. Section IV provides extensive comparison experiments and ablation studies. Finally, we conclude the article in Section V.

II. Related Work

In this section, we review the related methods of road segmentation and UDA in the context of RS.

A. Supervised Road Segmentation

Various applications, such as vehicle navigation and urban planning, require a constantly updated road database, which could be created and updated using road segmentation model on high-resolution RS images. Before the era of deep learning, most of the road segmentation methods are derived from a pixel-level classification strategy. Yuan *et al.* [5] proposed an automatic road extraction approach, where a three-step method including segmentation, medial axis points selection, and road grouping was adopted for road segmentation. Similarly, Das *et al.* [22] introduced a multistage algorithm in which probabilistic support vector machine (SVM) and salient features were exploited to extract road regions from high-resolution multispectral satellite images.

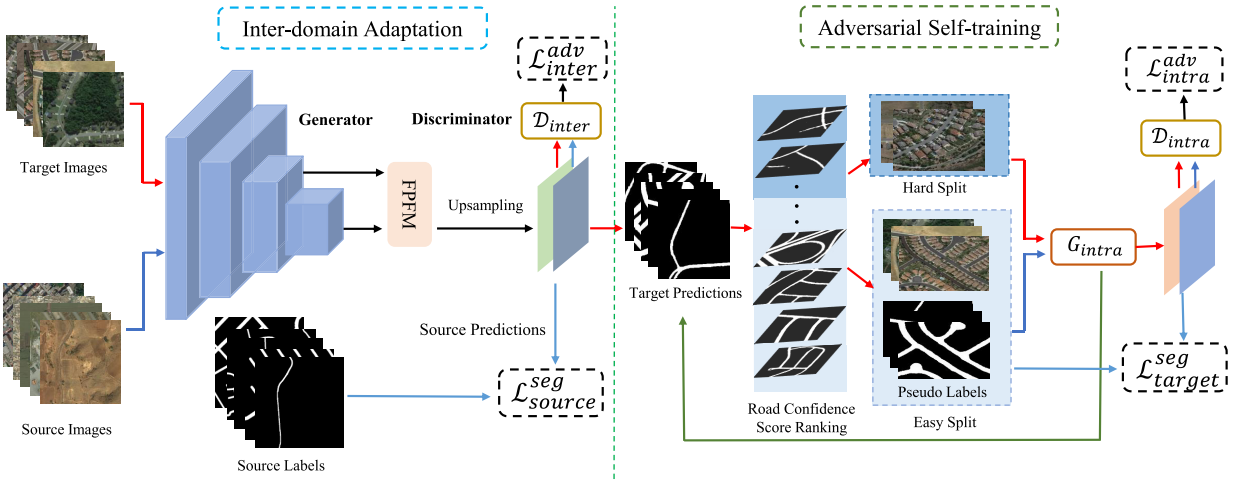


Fig. 2. Overview of our proposed RoadDA method, which consists of two stages. In the interdomain adaptation stage, given the source and unlabeled target data, a segmentation model equipped with a specifically devised FPFM acts as the generator to predict the segmentation results. The discriminator is trained to distinguish the domain label of the input, while the generator aims to generate similar distribution for both the source and target domains to fool the discriminator. In the adversarial self-training stage, we predict the segmentation maps of the target domain images using the trained model. The predictions are used to mine the easy target samples by calculating and ranking the road confidence scores and assigned to them as pseudo labels. These pseudo labels are used to help adapt the unlabeled hard split to the easy one and therefore reduce the intradomain discrepancy. The pseudo labels generation and intradomain adaptation are repeated to gradually improve the performance of the segmentation model.

As the rapid development of deep learning techniques, especially the deep CNNs, they have been widely applied in various kinds of computer vision tasks [23]–[27], owing to their strong capability of feature learning and end-to-end modeling. Some researchers attempt to introduce deep learning techniques to the field of road segmentation of RS images, which advance the development of this field [13], [28]–[30]. Cheng *et al.* [13] designed a cascaded end-to-end CNN for both road segmentation and centerline extraction tasks, where a symmetric encoder–decoder structure was employed for road segmentation. Ding and Bruzzone [12] proposed a direction-aware residual network for enhancing the structural completeness (COM) of the road networks, in which a direction supervision is introduced to strengthen the detection of linear features. Wei *et al.* [29] proposed a multistage framework to accurately extract the road surface and road centerline simultaneously. Li *et al.* [31] proposed to combine multiscale global context with adversarial networks. Specifically, the images with ground truth labels and images with prediction maps are fed into the discriminator network, which plays a min–max game with the segmentation model by back-propagating the classification error reversely. Although these approaches achieve remarkable performance, deep supervised learning-based methods require a large amount of labeled data, which are expensive and time-consuming to collect and annotate, especially for various unseen scenarios. Since we could get access to some public road datasets released by previous researchers, these data may be useful to help extract road information from unlabeled data by using domain adaptation techniques.

B. Unsupervised Domain Adaptation

UDA can be regarded as a particular case of domain adaptation that exploits labeled data in a relevant source domain to conduct a similar task in the target domain. The purpose of UDA methodologies is to reduce the DS, which typically deteriorates the performance of the models [32]–[35]. In recent years, various UDA methods have been proposed to address the DS problem in the semantic segmentation field. Tsai *et al.* [36] proposed a multilevel adversarial learning framework for UDA semantic segmentation, which employed adversarial learning in the structured output space at different feature levels. Li *et al.* [37] and Vu *et al.* [47] proposed the collaborative class conditional generative adversarial network (GAN) to bypass the dependence on the source data and improved the performance of the predicted model through generated target-style data. Pan *et al.* [38] proposed the intradomain adaptation of the target domain and designed the division criterion for target domain based on the mean of multiclass entropy map. Nevertheless, this criterion may be not suitable for binary road segmentation, since mean entropy-based criterion will misclassify all predicted background pixels as easy samples and mislead the intradomain adaptation. Therefore, we pay more attention to road pixels in the road segmentation task and design a new intradomain division strategy to better divide the target domain. Furthermore, we propose to combine the intradomain adaptation with self-training to further improve the segmentation performance.

UDA has also been introduced to the hyperspectral image classification. Tasar *et al.* [16] designed the color mapping GAN to generate fake training images with the same semantics as training images and used them to fine-tune the trained classifier. Ma *et al.* [39] used multiple classifiers and variational autoencoders to construct a GAN [40] to address the domain discrepancy. CaGAN [41] proposed a class-aware GAN for UDA multisource RS image classification. CaGAN selects the reliable per-category feature centers based on

the clustering and reduces the intraclass and the interclass discrepancies across domains by optimizing the class-level l_1-norm loss between the source and target category feature centers. Different from CaGAN, our RoadDA adopts the pseudo labels of easy split and address the intraclass discrepancy within target domain by adversarial self-training. Compared with the multiclass classification task of hyperspectral image, road segmentation can be regarded as a binary classification task of RGB RS images, where more attention should be paid to the road information than background pixels. Moreover, the road shape characteristics and the proportion of road and background pixels should also be considered to improve the model performance. In this sense, UDA methods of hyperspectral image classification that require considerations of spatial–spectral information and interclass relationships may not be suitable for road segmentation. Recently, ASPN [20] designed an ASPN for the domain adaptation in road segmentation of RS images, which focuses on extracting multilevel effective features and enhancing the feature representation. Few studies have been carried out on road segmentation of RS images and there are still many issues to be explored and settled.

III. Proposed Method

A. Overview

Our method aims to improve the performance of road segmentation on unlabeled target data using the UDA and self-training techniques. As depicted in Fig. 2, RoadDA consists of two stages: interdomain adaptation stage and adversarial self-training stage. In the first stage, a GAN equipped with a specifically devised FPFM is constructed to reduce the DS between the labeled source domain and unlabeled target domain. In the second stage, we use the trained generator to predict the pseudo labels of the target domain and then divide the target domain into easy and hard split based on the road confidence scores. An adversarial learning-based intradomain adaptation is performed to fine-tune the trained generator on the target domain and mitigate the intradomain discrepancy. We adopt the fine-tuned generator to get more accurate pseudo labels and use them for further intradomain adaptation. This procedure is iterated to progressively improve the segmentation performance until saturation. During the inference phase, only the adaptive generator is used for road segmentation.

B. Interdomain Adaptation

Interdomain adaptation is achieved by the GAN architecture. The input are the labeled source data $(\mathcal{X}_s, \mathcal{Y}_s)$ and unlabeled target data $(\mathcal{X}_t)$, where $\mathcal{X}_s, \mathcal{X}_t \in \mathbb{R}^{H\times W\times 3}$, and $\mathcal{Y}_s \in (0,1)^{H\times W}$ (0 for background pixel, 1 for road pixel). A segmentation model acts as the generator G_{inter} and predicts the segmentation probability map $P_s = G_{\text{inter}}(\mathcal{X}_s)$, $P_t = G_{\text{inter}}(\mathcal{X}_t)$ $(P_s, P_t \in \mathbb{R}^{H\times W\times 2})$ for source images and target images, respectively. As the intermediate representation learned in the segmentation process, the high-level semantic feature of high dimension actually has complex implicit semantics. Therefore, it may be less effective to perform domain alignment at the feature level and there is no guarantee that the joint image-label distributions are aligned between domains. The binary segmentation predictions of both source images and target images have strong similarity in context and layout. Therefore, we adopt an adversarial learning scheme to align the target domain to the source domain at the output level. We send the predictions P_s, P_t to the discriminator D_{inter} as input to correctly predict their domain labels, while the generator G_{inter} is trained to fool D_{inter}, that is, the generator G_{inter} is encouraged to generate similar prediction distributions on both the target and source domains.

Here, we describe the training process of G_{inter} and D_{inter}. For the discriminator, given the segmentation predictions on the source domain and target domain P_s, P_t, the fully convolutional discriminator is trained to classify the domain labels of the input samples. Here, we artificially set the labels of the source domain samples to 1 and the labels of the target domain samples to 0, thus the discriminator could be optimized with the binary cross-entropy (CE) domain classification loss. The original CE loss for discriminator could be formulated as follows:

$$\begin{aligned}\mathcal{L}^{D}_{\text{inter}}(P_s, P_t) = &-\sum_{h,w}\big[s\log\big(\mathbf{D}_{\text{inter}}(P_s)^{(h,w)}\big) \\ &+ (1-s)\log\big(1-\mathbf{D}_{\text{inter}}(P_s)^{(h,w)}\big)\big] \\ &-\sum_{h,w}\big[t\log\big(\mathbf{D}_{\text{inter}}(P_t)^{(h,w)}\big) \\ &+ (1-t)\log\big(1-\mathbf{D}_{\text{inter}}(P_t)^{(h,w)}\big)\big]\end{aligned} \tag{1}$$

when s = 1 and t = 0, it is rewritten as

$$\begin{aligned}\mathcal{L}^{D}_{\text{inter}}(P_s, P_t) = &-\sum_{h,w}\big[\log\big(\mathbf{D}_{\text{inter}}(P_s)^{(h,w)}\big) \\ &+ \log\big(1-\mathbf{D}_{\text{inter}}(P_t)^{(h,w)}\big)\big].\end{aligned} \tag{2}$$

For the training of generator G_{inter}, it contains two parts. First, the segmentation loss of images from the source domain, which is the CE loss

$$\mathcal{L}^{\text{seg}}_{\text{source}}(\mathcal{X}_s, \mathcal{Y}_s) = -\sum_{h,w,c}\mathcal{Y}_s^{(h,w,c)}\log\big(\mathbf{G}_{\text{inter}}(\mathcal{X}_s)^{(h,w,c)}\big) \tag{3}$$

where the source domain labels $\mathcal{Y}_s$ have been converted to the one-hot vector form and c denotes the number of classes. Second, the adversarial loss for the target images is defined as

$$\begin{aligned}\mathcal{L}^{\text{adv}}_{\text{inter}}(\mathcal{X}_t) = &-\sum_{h,w}\big[t\log\big(\mathbf{D}_{\text{inter}}(P_t)^{(h,w)}\big) \\ &+ (1-t)\log\big(1-\mathbf{D}_{\text{inter}}(P_t)^{(h,w)}\big)\big].\end{aligned} \tag{4}$$

According to the adversarial strategy, we set $t = 1$, which is opposite of the discriminator, then the adversarial loss could be rewritten as

$$\mathcal{L}^{\text{adv}}_{\text{inter}}(\mathcal{X}_t) = -\sum_{h,w}\log\big(\mathbf{D}_{\text{inter}}(\mathbf{G}_{\text{inter}}(\mathcal{X}_t))^{(h,w)}\big). \tag{5}$$

This loss forces the generator to produce a source-like distribution on the target domain to fool the discriminator and

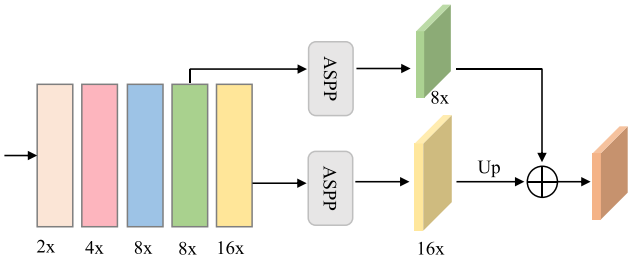


Fig. 3. Illustration of the proposed FPFM. 8× denotes a downsampling rate of 8.

finally the model reaches a balance state. In summary, the final training objective for the generator is

$$\mathcal{L}^{G}_{\text{inter}}(\mathcal{X}_s, \mathcal{X}_t) = \mathcal{L}^{\text{seg}}_{\text{source}}(\mathcal{X}_s) + \alpha_{\text{adv}}\mathcal{L}^{\text{adv}}_{\text{inter}}(\mathcal{X}_t) \tag{6}$$

where α_{adv} is the loss weight to balance the two losses.

During training, we alternatively optimize discriminator D_{inter} and generator G_{inter} using the loss function in (2) and (6), while in inference, only the segmentation model (generator) is used for road segmentation.

1) Feature Pyramid Fusion Module: In the GAN framework, a fully convolutional segmentation model acts as the generator and the backbone network is ResNet-101 [42]. Let the hierarchical features extracted by ResNet-101 be denoted as $[C_1, C_2, C_3, C_4, C_5]$, where the corresponding downsampling rates are [2, 4, 8, 16, 32]. High downsampling rate is originally designed to reduce the feature size and the computation cost and to increase the feature robustness. However, in the different scenarios of road segmentation from RS images, high downsampling rate may result in the spatial information loss of long and thin roads, which will affect the final road segmentation performance in dealing with different domain data. In contrast, if we keep high resolution in the deep layer, such as using the downsampling rate set of [2, 4, 8, 8, 8], it will reduce the effective receptive field of the model as well as increase the computation cost in the deep layer. To better handle the various scenarios in both the source and target domains, we design the FPFM to fuse multilevel and multiscale road features. As shown in Fig. 3, we employ the downsampling rate set of [2, 4, 8, 8, 16] by setting the stride of fourth stage in ResNet101 to 1, and select the features C_4 and C_5 as input to the atrous spatial pyramid pooling (ASPP) module [43] to generate the global context features C_4^* and C_5^* at different level. We then upsample C_5^* to the same size as C_4^* and add them together to explicitly enhance the feature representation ability and generate the robust fused feature at the downsampling rate of 8.

C. Adversarial Self-Training

After the training of the interdomain adaptation stage, we obtain the adapted segmentation model which can achieve considerable performance improvement compared with the model trained only on the source domain (source-only). However, according to our observation, there may be intradomain discrepancy in the target domain due to different illumination conditions and background context. Thus, we propose the adversarial learning based self-training stage to further improve the segmentation performance on the target domain. This stage is composed of three parts: intradomain division, intradomain adaptation, and self-training.

1) Intradomain Division: To address the intradomain discrepancy in the target domain, it should be decomposed into two parts and reduce their discrepancy. To this end, we propose a quality estimator to accomplish this task. Inspired by the observation that the performance of the segmentation model trained on the source domain will deteriorate on the target domain, we use the adapted segmentation model G_{inter} to predict the segmentation maps for the target domain and accordingly divide the target domain into an easy split and a hard split. Specifically, we design a quality estimator to estimate the road confidence scores of prediction results and rank them based on the scores. Finally, the target images are separated into an easy split $\mathcal{X}_{\text{te}}$ with pseudo labels and an unlabeled hard split $\mathcal{X}_{\text{th}}$.

Let $\mathcal{P}_t = G_{\text{inter}}(\mathcal{X}_t)$ denote the predicted segmentation map of the target images, and $\mathcal{P}^t_{i,j}$ denotes the probability of road at location (i, j). $\mathcal{M}_t$ represents the predicted binary mask (pseudo labels). $\mathcal{M}^t_{i,j}$ equals 1 when the probability of pixel at (i, j) is higher than a threshold of 0.5 and the pixel is predicted as road; otherwise, it equals 0. Since we focus on the road pixels rather than the large numbers of background pixels which may bias the estimation, we filter out the background and calculate the confidence score of the prediction as follows:

$$\mathcal{S}_{\text{cf}} = \frac{\sum_{i,j} \mathcal{P}^t_{i,j} \cdot \mathcal{M}^t_{i,j}}{\sum_{i,j} \mathcal{M}^t_{i,j}} \tag{7}$$

which is the mean value of the confidence of all predicted road pixels in each image. Then we rank the target images in a descending order according to their $\mathcal{S}_{\text{cf}}$ scores and employ a hyperparameter $\lambda \in (0, 1)$ as a ratio to divide the ranked target images into an easy split $\mathcal{X}_{\text{te}}$ with high-confidence pseudo labels $\mathcal{M}_{\text{te}}$ and an unlabeled hard split $\mathcal{X}_{\text{th}}$, where $|X_{\text{te}}| = \lambda|X_t|$ and $|X_{\text{th}}| = (1-\lambda)|X_t|$. In Section IV, we conducted a hyperparameter analysis experiment to investigate the influence of λ.

2) Intradomain Adaptation: In this section, the same GAN architecture as the first stage is constructed to align the hard split with the easy split at the output level and reduce the intradomain gap. The generator G_{intra} is initialized with the generator adapted in the first stage and takes $\mathcal{X}_{\text{te}}$ and $\mathcal{X}_{\text{th}}$ as the input, and the output are fed into the discriminator D_{intra}, which aims to predict the domain labels. D_{intra} is trained to minimize the binary CE classification loss, while G_{intra} is trained to minimize the segmentation loss and the adversarial loss. They are defined as follows:

$$\mathcal{L}^{D}_{\text{intra}}(\mathcal{X}_{\text{te}}, \mathcal{X}_{\text{th}}) = -\sum_{h,w}\Big[\log\big(\mathbf{D}_{\text{intra}}(\mathbf{G}_{\text{intra}}(\mathcal{X}_{\text{te}}))^{(h,w)}\big) + \log\big(1-\mathbf{D}_{\text{intra}}(\mathbf{G}_{\text{intra}}(\mathcal{X}_{\text{th}}))^{(h,w)}\big)\Big] \tag{8}$$

$$\mathcal{L}^{G}_{\text{inter}}(\mathcal{X}_{\text{te}}, \mathcal{X}_{\text{th}}) = \mathcal{L}^{\text{seg}}_{\text{target}}(\mathcal{X}_{\text{te}}) + \beta_{\text{adv}}\mathcal{L}^{\text{adv}}_{\text{intra}}(\mathcal{X}_{\text{th}}) \tag{9}$$

where β_{adv} is the loss weight to balance these two losses and

$$\mathcal{L}_{\text{target}}^{\text{seg}}(\mathcal{X}_{\text{te}}, \mathcal{M}_{\text{te}}) = -\sum_{h,w,c} \mathcal{M}_{\text{te}}^{(h,w,c)} \log\left(\mathbf{G}_{\text{intra}}(\mathcal{X}_{\text{te}})^{(h,w,c)}\right) \tag{10}$$

$$\mathcal{L}_{\text{intra}}^{\text{adv}}(\mathcal{X}_{\text{th}}) = -\sum_{h,w} \log\left(\mathbf{D}_{\text{intra}}(\mathbf{G}_{\text{intra}}(\mathcal{X}_{\text{th}}))^{(h,w)}\right). \tag{11}$$

As in the first stage, D_{intra} and G_{intra} are alternatively optimized using the loss functions in (8) and (9).

3) Self-Training: After the intradomain adaptation training on the target domain, the adapted segmentation model G_{intra} achieves better performance and predicts more accurate road segmentation maps. Therefore, we update the pseudo labels of the target images using this adapted generator and repeat the intradomain division and adaptation process to progressively improve the performance of the road segmentation model until saturation, as shown in Fig. 2. The training algorithm of RoadDA is summarized in Algorithm 1.

Algorithm 1 Training Algorithm of RoadDA

Stage1: Inter-domain adaptation
- **Input:** source domain data: $D_s = (\mathcal{X}_s, \mathcal{Y}_s)$; target domain data: $D_t = \mathcal{X}_t$.
- **Initialize:** G_{inter} and D_{inter}
- **while** $epoch < epoch_{max}$ **do**
  - $epoch = epoch + 1$
  - **for** $i < iteration_{max}$; i++ **do**
    - Derive B_s and B_t sampled from $\mathcal{X}_s$ and $\mathcal{X}_t$
    - **Train** G_{inter} on B_s and B_t by optimizing $\mathcal{L}_{inter}^{G}$
    - **Train** D_{inter} on B_s and B_t by optimizing $\mathcal{L}_{inter}^{D}$
  - **end for**
- **end while**
- **Output:** Adapted segmentation model $\boldsymbol{G_{inter}}$

Stage2: Adversarial self-training
- **Input:** target data: $\mathcal{X}_t$, adapted G_{inter}, hyperparameter λ.
- **Initialize:** Initialize G_{intra} using G_{inter}; D_{intra}
- **Step1:** generate pseudo labels $\mathcal{M}_t$ for $\mathcal{X}_t$ using G_{intra}
- **Step2:** divide $\mathcal{X}_t$ into $\mathcal{X}_{te}$ and $\mathcal{X}_{th}$ s.t. $|\mathcal{X}_{te}| = \lambda|X_t|$
- **Step3:** train G_{intra} on $\mathcal{X}_{te}$ and $\mathcal{X}_{th}$ by optimizing $\mathcal{L}_{intra}^{G}$ train D_{intra} on $\mathcal{X}_{te}$ and $\mathcal{X}_{th}$ by optimizing $\mathcal{L}_{intra}^{D}$
- **Step4:** evaluate the segmentation performance of G_{intra}
- **Repeat:** Step1-Step3
- **Until:** performance saturation of G_{intra}
- **Output:** $\boldsymbol{G_{intra}}$

D. Implementation Details

For the generator, ResNet-101 is pretrained on ImageNet. The discriminative fused feature from FPFM is directly upsampled to predict the final segmentation probability map. For the discriminator, we adopt the fully convolutional network similar to [44]. The network contains five convolution layers with 4×4 kernel and stride of 2. Leaky rectified linear unit (ReLU) with a slope of 0.2 follows each convolution layer as activation function except the last layer.

We implemented our model with PyTorch on a single NVIDIA V100 GPU. The generators (G_{inter}, G_{intra}) were optimized using the stochastic gradient descent (SGD) optimizer with the momentum of 0.9 and a weight decay of 10^{-4}. We used the Adam optimizer to optimize the discriminator with the momentum of 0.9 and 0.99. The initial learning rates are 4×10^{-4} and 1×10^{-4} for generators and discriminators, respectively, and both of which are decreased using the polynomial decay policy with a power of 0.9. The loss weights α_{adv} and β_{adv} are set to 0.1 and 0.01, respectively. We set the batch size to 4 and λ to 0.7.

IV. Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of RoadDA in terms of both objective evaluation metrics and subjective visual comparisons.

A. Datasets

In the UDA experiments of road segmentation, we consider the adaptation setting that includes different spatial resolution, background region, and different types of road surfaces. Here, we employ the RTDS [45] and DeepGlobe dataset [46] as the source domains, and the CasNet dataset (CNDS) [13] as the target domain. The segmentation model is trained on the labeled source data and unlabeled target data and is evaluated on the test set of CNDS.

1) *DeepGlobe Dataset (DeepGlobe)* This dataset is proposed in the road extraction challenge of DeepGlobe 2018 [46]. The images of this dataset are captured over Thailand, Indonesia, and India with the spatial resolution of 0.5 m/pixel. Images are sampled uniformly between rural and urban areas and cropped to extract useful region by geographic information system (GIS) experts. The dataset consists of 8570 RGB images with a size of 1024×1024, where 6226 images are randomly sampled for training, and 1243 images and 1101 images are chosen as the validation and test sets, respectively. Especially, the official available dataset only provides labels for the training set. The dataset is available at https://competitions.codalab.org/competitions/18467.
2) *RTDS*: RTDS is collected and first used in [45]. RTDS is a large corpus of high-resolution satellite images and ground-truth road network graphs, in which images are obtained from Google Map with 0.6 m/pixel resolution and ground-truth road network are from the OSM. RTDS covers the urban core of 40 cities in six countries. In each city, the center area of approximately 24 km^2 is selected as the sample of the dataset for a total of 300 RGB images having a size of 4096×4096 pixels. Following [45], images from 25 cities are randomly selected as the training set and images of the 15 remaining cities serve as the test set. The dataset is available at https://github.com/mitroadmaps/roadtracer.
3) *CNDS*: the CNDS, built by Cheng *et al.* [13], consists of 224 RGB RS images collected from Google Earth with manual annotations. The images in this dataset

Fig. 4. Some visual examples from different datasets. (a) DeepGlobe. (b) RTDS. (c) CNDS.

TABLE I
DETAILED INFORMATION OF THE DATASETS

Datasets	Resolution	Size	Training Data	Test Data
DeepGlobe	0.5 m/pixel	512×512	44,165	-
RTDS	0.6 m/pixel	512×512	51,106	-
CNDS	1.2 m/pixel	512×512	28,420	589

are at least 600 $\times$ 600 pixels with the spatial resolution of 1.2 m/pixel. Moreover, these images contain complex background and diversified road shapes. Following [13], we randomly select 180 images as the training set, 14 images for validation, and the rest 30 images as the test set. The dataset is available at https://shibiaoxu.github.io/TGRS2017.html.

Some image examples are depicted in Fig. 4.

B. Data Augmentation

Because of the limitation of the GPU resource, we cannot directly train the model using the high-resolution RS images. Besides, the number of samples in the original dataset is insufficient to train a model with good feature representation and generalization. Therefore, data preprocessing and augmentation are carried out before training.

For each image in these datasets, we randomly crop 20 patches having a size of 512 $\times$ 512 and then filter out the patches where the number of road pixels is less than 4000. Since the total number of training samples in CNDS is still insufficient, we flip each cropped training sample in the horizontal direction and rotate the original and flipped samples at the step of 90° for three times. The detailed information of the final processed and augmented datasets is reported in Table I.

C. Evaluation Metrics

Four common metrics are used to evaluate the quantitative performance of different road segmentation models, including intersection-over-union (IoU), COM, correctness (COR), and F1 score. COM measures the proportion of matched road pixels in the ground truth map, COR reports the percentage of matched road areas in the predicted segmentation map. The F1 score is a harmonic average between COM and COR which can measure the robustness of methods. The four metrics are defined as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \quad \text{COM} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$
$$\text{COR} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad \text{F1} = \frac{2 \times \text{COM} \times \text{COR}}{\text{COM}+\text{COR}} \tag{12}$$

where TP, FP, and FN denote the number of true positive pixels, false positive pixels, and false negative pixels, respectively. A larger metric value indicates better performance.

D. Comparative Methods

To demonstrate the superiority of the proposed method, we compare it with several state-of-the-art methods, including AdaptSegNet [36], Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation (ADVENT) [47], bidirectional learning for domain adaptation of semantic segmentation (BDL) [48], and IntraDA [38]. Here, we briefly describe these approaches as follows.

1) *Source-only*: This is a baseline model that is only trained on the labeled source domain data while directly tested on the target domain.
2) *Target-only*: Target-only represents the typical supervised road segmentation method which is trained on the target domain with labels. It is the oracle model to provide the upper bound of performance and indicate the effectiveness and reliability of those UDA methods.
3) *AdaptSegNet*: AdaptSegNet first adopts adversarial learning in the output space for UDA semantic segmentation. Compared with the adaptation in the feature level, structured output spaces of the source and target domains contain the global context and spatial similarities, which is beneficial to the adaptation. Moreover, a multi-level adversarial network is designed to enhance the model.
4) *ADVENT*: Based on AdaptSegNet, ADVENT calculates the entropy of output space to avoid low-confident predictions on the target domain. The entropy map is fed into the discriminator for adversarial training.
5) *BDL*: BDL adopts a bidirectional learning framework which first carries out image-to-image translation and then uses the translated images to complete the adversarial domain adaptation. The image translation and adversarial adaptation promote each other to progressively improve performance.
6) *IntraDA*: In addition to the adaptation between the source and target domains, IntraDA tries to address the intra-DS, although the entropy-based criterion designed to separate the multiclass target domain is not suitable for binary road segmentation tasks.

TABLE II
QUANTITATIVE RESULTS ON THE SETTING OF DEEPGLOBE ⟶ CNDS. ADV. LEARNING: ADVERSARIAL LEARNING

DeepGlobe ⟶ CNDS							
Methods	backbone	Input size	Strategy	IoU	COM	COR	F1
Source-only	ResNet-101	512 × 512	Segmentation	52.84	57.47	89.57	70.01
AdaptSegNet [36]	ResNet-101	512 × 512	Adv. learning	54.13	59.04	**91.86**	71.88
ADVENT [47]	ResNet-101	512 × 512	Adv. learning	59.40	72.76	90.66	80.72
BDL [48]	ResNet-101	512 × 512	Style transfer & Adv. learning	72.35	79.91	89.97	84.64
IntraDA [38]	ResNet-101	512 × 512	Adv. learning	72.73	80.06	90.21	84.83
RoadDA	ResNet-101	512 × 512	Adv. learning & self-training	**74.92**	**82.11**	89.88	**85.81**
Target-only	ResNet-101	512 × 512	Segmentation	85.35	91.23	93.51	92.35

TABLE III
QUANTITATIVE RESULTS OF ADAPTING RTDS TO CNDS. ADV. LEARNING: ADVERSARIAL LEARNING

RTDS ⟶ CNDS							
Methods	backbone	Input size	Strategy	IoU	COM	COR	F1
Source-only	ResNet-101	512 × 512	Segmentation	45.96	48.39	**91.36**	63.26
AdaptSegNet [36]	ResNet-101	512 × 512	Adv. learning	53.34	58.03	89.48	70.40
ADVENT [47]	ResNet-101	512 × 512	Adv. learning	54.44	58.03	90.71	71.26
BDL [48]	ResNet-101	512 × 512	Style transfer & Adv. learning	55.47	59.65	90.75	71.98
IntraDA [38]	ResNet-101	512 × 512	Adv. learning	56.06	60.26	91.16	72.55
RoadDA	ResNet-101	512 × 512	Adv. learning & self-training	**61.76**	**67.56**	90.81	**77.48**
Target-only	ResNet-101	512 × 512	Segmentation	85.35	91.23	93.51	92.35

E. UDA Results on Different Adaptation Setting

1) DeepGlobe ⟶ CNDS: In Table II, we report the road segmentation performance of our method and other comparison methods on the CNDS test set. For a fair comparison, the baseline model and target-only model adopt the same segmentation model as RoadDA and all the methods also use the same ResNet-101 backbone. Overall, RoadDA achieves the best performance of 74.92% IoU and 85.81% F1 score compared with the state-of-the-art UDA methods, which shows the reliability and robustness of our model. Source-only and target-only models produce the worst and best results, respectively, as expected. Compared with the methods with only interdomain adaptation, such as AdaptSegNet and ADVENT, the intradomain adaptation and self-training adopted in our model bring considerable performance improvement. For instance, RoadDA is 15.52% higher than ADVENT in terms of IoU. Moreover, our method also outperforms IntraDA, which also uses intradomain adaptation. We argue that the performance gain owes to the employment of more effective division criterion for the target domain and self-training.

Fig. 5 presents the visual road segmentation results of different methods on four representative test images from the CNDS test set. It can be observed that RoadDA obtains the best segmentation results close to the ground-truth and could handle the RS images with different complex backgrounds. Moreover, most of the road contours in the images can be captured by RoadDA, demonstrating the great potential of using domain adaptation techniques to solve the unsupervised road segmentation problem.

2) RTDS ⟶ CNDS: Here, we use the RTDS as the source domain and report the evaluation results on the CNDS test set in Table III. All the methods employ the same configuration as the previous experiment. As shown in Table III, our proposed method achieves 61.76% IoU and 77.48% F1 score, which is superior to all the UDA comparison methods. Specifically, RoadDA is 15.8% higher than source-only and 7.32% higher than ADVENT in terms of IoU. Compared with IntraDA equipped with intradomain adaptation, our model also brings 5.7% improvement in IoU metric.

Comparing the experimental results of the two adaptation settings, we can find that the adaptation performance of DeepGlobe is better than RTDS. This may be caused by the inherent properties of the domain itself. The images in DeepGlobe dataset have more similar spatial information to the target dataset, such as the background context and road shape, while the images in RTDS are captured in the big city centers where roads are too wide or indistinguishable from surrounding buildings. This may give us some hints for choosing the suitable source domain in practice.

F. Model Analysis

1) Ablation Study: In this section, we evaluate the effects of the proposed modules in RoadDA on the setting of DeepGlobe ⟶ CNDS. The input image size is 512 × 512.

In Table IV, we summarize the performance of RoadDA using different variants of FPFM on the setting of DeepGlobe ⟶ CNDS. RoadDA with 32× means that we only use the deep feature of a downsampling rate of 32 as input for

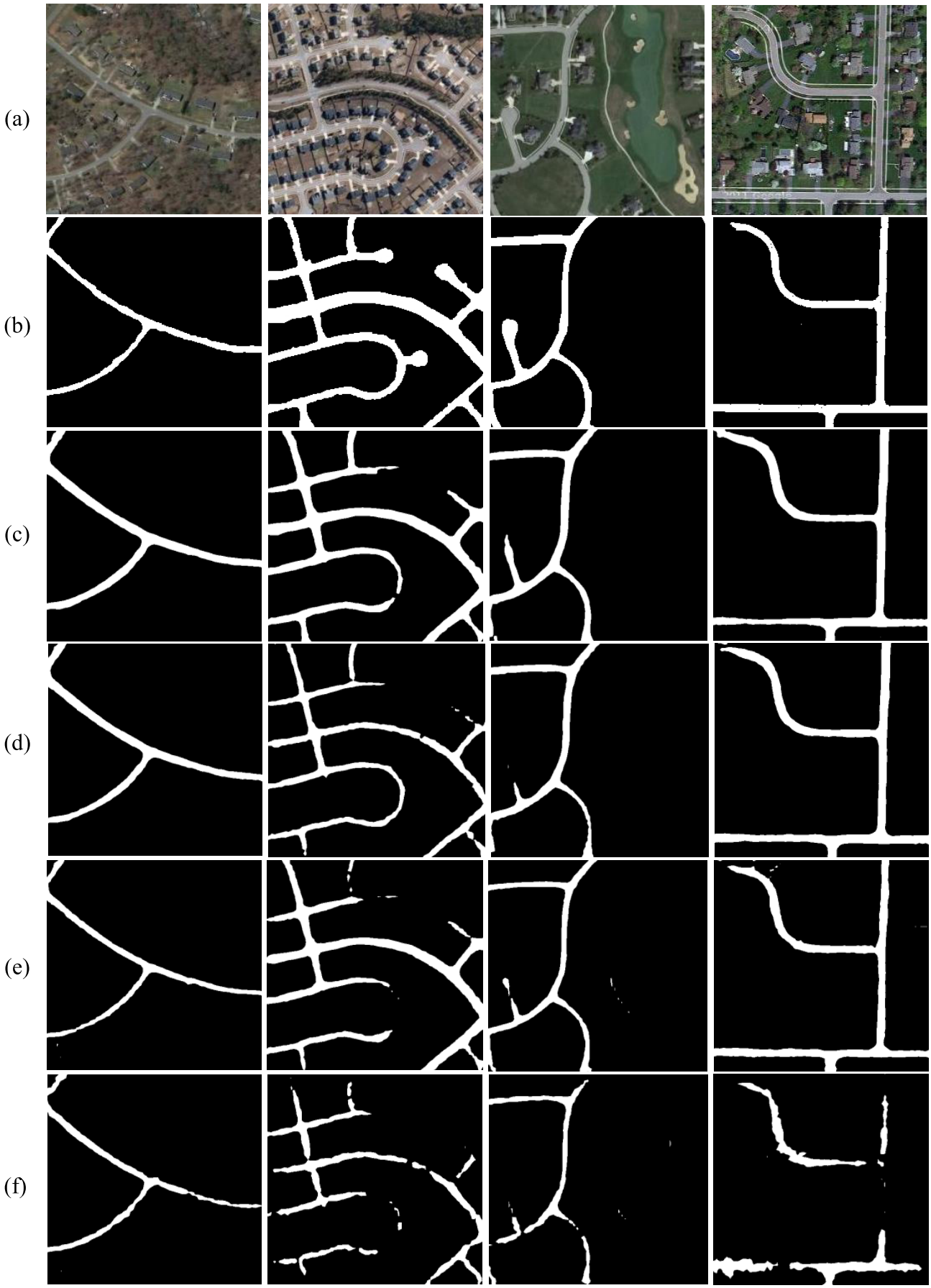


Fig. 5. Visual results of typical methods on CNDS. (a) Target images. (b) Ground-truth. (c) Our RoadDA. (d) IntraDA. (e) ADVENT. (f) Source-only.

TABLE IV

ABLATION STUDY RESULTS OF THE PROPOSED FPFM ON THE SETTING OF DEEPGLOBE ⟶ CNDS

Model	32×	16×	8×	Fusion	IoU
RoadDA	✓	×	×	×	73.24
RoadDA	×	✓	×	×	73.65
RoadDA	×	×	✓	×	74.17
RoadDA	✓	✓	×	✓	73.98
RoadDA	×	✓	✓	✓	**74.92**

TABLE V

ABLATION STUDY RESULTS OF THE PROPOSED TECHNIQUES IN THE ADVERSARIAL SELF-TRAINING STAGE ON THE SETTING OF DEEPGLOBE ⟶ CNDS. IA: INTRADOMAIN ADAPTATION. ST: SELF-TRAINING

Model	IA	ST	IoU	COM	COR	F1
RoadDA	×	×	61.74	73.48	90.57	81.13
RoadDA	✓	×	73.64	80.72	90.03	85.12
RoadDA	×	✓	74.08	81.15	89.97	85.33
RoadDA	✓	✓	**74.92**	**82.11**	89.88	**85.81**

ASPP module and then directly generate the segmentation map without the fusion operation. Among different variants of RoadDA only using the single-level feature, the setting of 8× performs best by obtaining a 74.17% IoU which shows that the high-resolution features preserve more road spatial information and beneficial to the performance. RoadDA with 32×, 16×, and Fusion denotes using the FPFM which takes the two levels of features of the downsampling rate of 32 and 16 as input. Overall, RoadDA using the proposed setting of FPFM achieves the best performance of 74.92% IoU compared with all other variants.

We also conducted an ablation study on the second adversarial self-training stage. The results are summarized in Table V. As can be seen, RoadDA without intradomain adaptation and self-training, which only employs the interdomain adaptation, achieves 61.74% IoU and 81.13% F1 score. After using the intradomain adaptation and self-training strategies, the performance is improved. The intradomain adaptation provides 11.9% IoU improvement and the self-training brings 12.34% IoU gains. RoadDA using both intradomain adaptation and self-training achieves a gain of 13.18% IoU over the vanilla baseline, which confirms the effectiveness of the proposed techniques in the adversarial self-training stage.

2) Division Ratio λ: For intradomain division, λ controls the number of samples in the easy and hard splits and also has an impact on the distribution of simple samples and difficult samples. To investigate its influence, we conducted the experiments using λ from 0.6 to 0.9 on both adaptation settings. As reported in Table VI, RoadDA achieves the best performance for both settings when $\lambda = 0.7$.

3) Iteration for Self-Training: In the adversarial self-training stage, we iterate the intradomain adaptation process to progressively improve model performance, while it is unknown how many iterations it will take to achieve performance saturation. Therefore, we analyze the relationship between the iteration and the IoU on two adaptation setting.

TABLE VI

IOU RESULTS OF DIFFERENT HYPERPARAMETER λ ON BOTH ADAPTATION SETTINGS. DPG: DEEPGLOBE DATASET

λ	0.6	0.7	0.8	0.9
DPG ⟶ CNDS	73.54	**74.92**	74.01	72.87
RTDS ⟶ CNDS	60.25	**61.76**	61.33	60.62

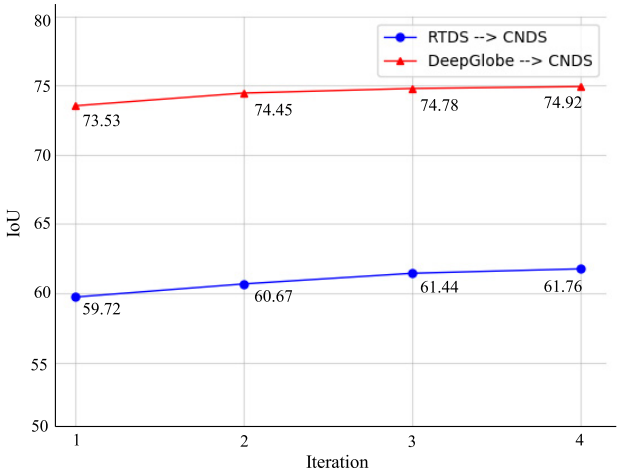


Fig. 6. IoU results versus the iterations of self-training.

TABLE VII

RESULTS OF TWO CHOICES OF DOMAIN ADAPTATION SPACE ON THE SETTING OF DEEPGLOBE ⟶ CNDS

Methods	IoU	COM	COR	F1
Feature-level adaptation	51.42	56.36	87.65	68.61
RoadDA (output space)	74.92	82.11	89.88	85.81

As depicted in Fig. 6, the performance gradually increases and tends to be saturated after three or four iterations. Thus, three or four iterations may be a proper choice to save the training time while achieve a decent performance.

4) Domain Adaptation Space: Here, we investigate the influence of domain adaptation space on the segmentation results. As shown in Table VII, when RoadDA performs the feature-level domain adaptation, the model suffers significant performance deterioration and only achieves 51.42% IoU which is 23.5% lower than RoadDA using output space domain adaptation. It indicates that the structured prediction space is more appropriate for domain adaptation in the binary road segmentation task since RS images always contain complex structures and consequently complex feature distributions may obtain for different domains.

5) Model Training: Fig. 7 presents the curves of the segmentation loss and adversarial loss against the number of iterations on the DeepGlobe ⟶ CNDS setting. The loss curves in the interdomain adaptation stage and the intradomain adaptation are listed in the left side and right side, respectively. As can be seen, with the increasing iterations, the segmentation loss and discriminator loss become smaller until convergence, while the adversarial loss increases gradually until convergence. Finally, the adversarial learning reaches a balanced state where the model generates similar distributions on both

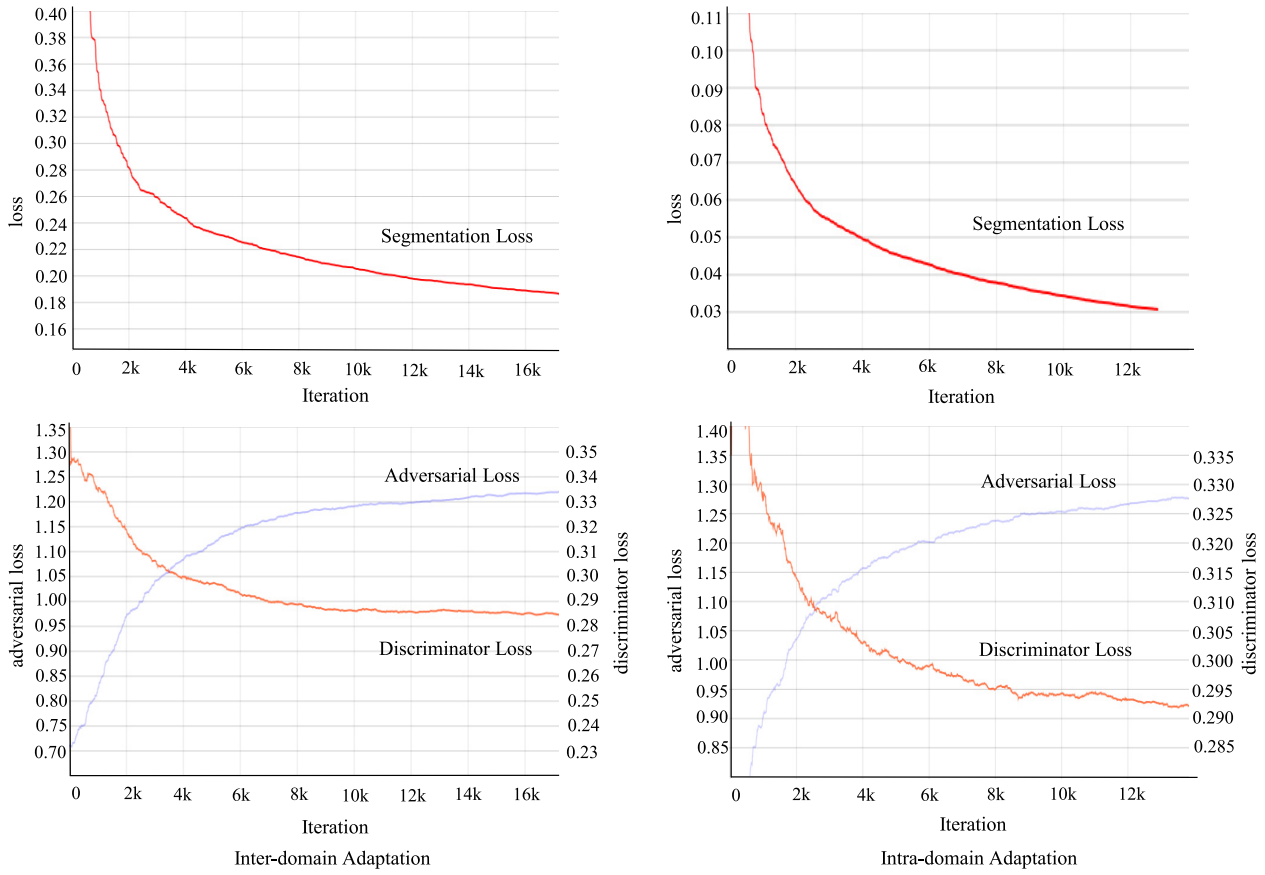


Fig. 7. Segmentation loss and adversarial loss versus the iterations while training RoadDA on DeepGlobe and CNDS.

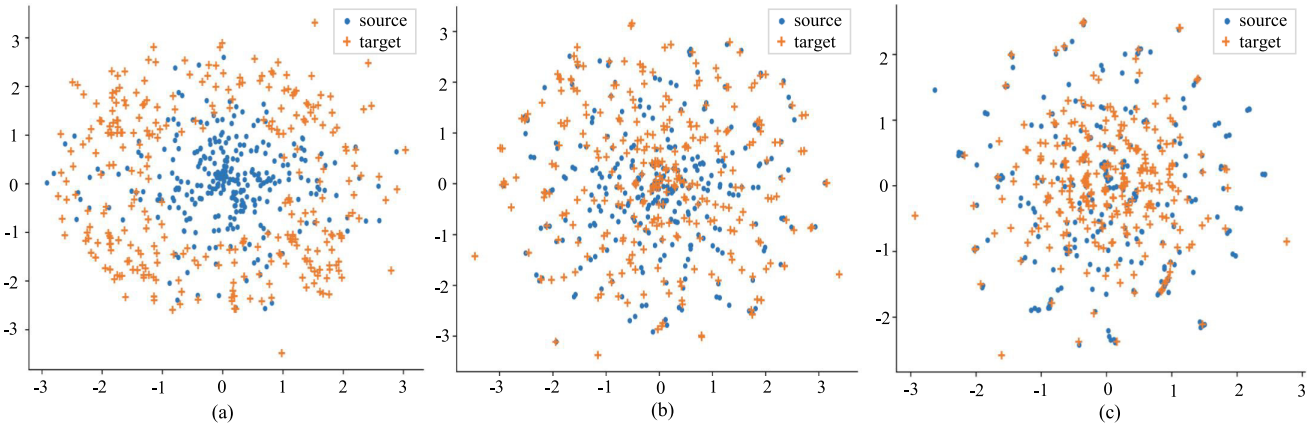


Fig. 8. T-SNE visualization of the features obtained by different models on the setting of DeepGlobe ⟶ CNDS. (a) Source-only. (b) RoadDA using interdomain adaptation only. (c) RoadDA.

the target and source domains. In addition, we show the t-distributed stochastic neighbor embedding (t-SNE) visualization [49] of the features before the prediction layer of two variants of our RoadDA and the source-only model in Fig. 8. The results demonstrate that using the two-stage adaptation in the output-level space, the features in the target domain is gradually aligned with those in the source domain.

V. Conclusion

In this article, we propose a novel two-stage UDA framework for road segmentation of the RS images. In the first interdomain adaptation stage, our model equipped with a specifically designed FPFM efficiently mitigates the domain discrepancy between the labeled source domain and the unlabeled target domain, resulting in an adapted segmentation model that can generalize well to the target domain. In the second adversarial self-training stage, our model effectively mine easy target samples and assign pseudo labels to them, which are used to guide the intradomain adaptation to mitigate the discrepancy within the target domain. Our model outperforms state-of-the-art domain adaptation methods on two benchmark settings for RS road segmentation, showing a promising

application potential in real-world scenarios. In the future works, we will attempt to use the RS image dataset of river, which has a similar shape with the road, as the source domain to guide the segmentation of the road in the target domain.

ACKNOWLEDGMENT

This work was done during Meng Lan's internship at JD Explore Academy.

REFERENCES

[1] C. Unsalan and B. Sirmacek, "Road network detection using probabilistic and graph theoretical methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4441–4453, Nov. 2012.

[2] R. Cao *et al.*, "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 82–97, May 2020.

[3] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[4] M. O. Sghaier and R. Lepage, "Road damage detection from VHR remote sensing images based on multiscale texture analysis and dempster shafer theory," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4224–4227.

[5] J. Yuan, D. Wang, B. Wu, L. Yan, and R. Li, "LEGION-based automatic road extraction from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4528–4538, Nov. 2011.

[6] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[7] M. E. Paoletti, J. M. Haut, N. S. Pereira, J. Plaza, and A. Plaza, "Ghostnet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 22, 2021, doi: 10.1109/TGRS.2021.3050257.

[8] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[9] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[10] L. Zhang, J. Zhang, W. Wei, and Y. Zhang, "Learning discriminative compact representation for hyperspectral imagery classification," *IEEE Trans. Geoence Remote Sens.*, vol. 57, no. 10, pp. 8276–8289, Oct. 2019.

[11] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[12] L. Ding and L. Bruzzone, "DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 16, 2020, doi: 10.1109/TGRS.2020.3034011.

[13] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[14] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 184–199, Mar. 2021.

[15] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[16] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[17] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 433–443.

[18] A. Elshamli, G. W. Taylor, A. Berg, and S. Areibi, "Domain adaptation using representation learning for the classification of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4198–4209, Sep. 2017.

[19] 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.

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

[21] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10687–10698.

[22] S. Das, T. T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3906–3931, Oct. 2011.

[23] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[24] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[25] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7789–7817, May 2021.

[26] W. Wang, W. Zhai, and Y. Cao, "Deep inhomogeneous regularization for transfer learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 221–225.

[27] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[28] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Inf. Sci.*, vol. 535, pp. 156–171, Oct. 2020.

[29] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

[30] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[31] Y. Li, B. Peng, L. He, K. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2279–2287, Jul. 2019.

[32] M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, "Enhanced transport distance for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13936–13944.

[33] J. Dong, Y. Cong, G. Sun, Y. Liu, and X. Xu, "CSCL: Critical semantic-consistent learning for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 745–762.

[34] L. Song *et al.*, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.

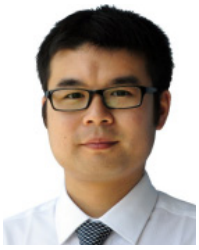
[35] Z. Chen, J. Zhang, and D. Tao, "Progressive LiDAR adaptation for road detection," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 693–702, May 2019.

[36] 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.*, Jun. 2018, pp. 7472–7481.

[37] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9641–9650.

[38] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3764–3773.

[39] X. Ma, X. Mou, J. Wang, X. Liu, J. Geng, and H. Wang, "Cross-dataset hyperspectral image classification based on adversarial domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4179–4190, May 2021.
[40] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*. [Online]. Available: http://arxiv.org/abs/1406.2661
[41] Q. Xu, X. Yuan, and C. Ouyang, "Class-aware domain adaptation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 17, 2020, doi: 10.1109/TGRS.2020.3031926.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[44] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[45] F. Bastani *et al.*, "RoadTracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4720–4728.
[46] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
[47] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.
[48] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.
[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.



Lefei Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively.

He was a Big Data Institute Visitor with the Department of Statistical Science, University College London, London, U.K., and a Hong Kong Scholar with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. He is a Professor with the School of Computer Science, Wuhan University. His research interests include pattern recognition, image processing, and remote sensing.

Dr. Zhang serves as an Associate Editor for *Pattern Recognition* and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Meng Lan received the B.S. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2018, where he is pursuing the Ph.D. degree.

His research interests include deep learning and computer vision.

Jing Zhang (Member, IEEE) is a Research Fellow with the School of Computer Science, The University of Sydney, Sydney, NSW, Australia. He has authored or coauthored more than 30 articles in prestigious journals and proceedings at leading conferences. His research interests include computer vision and deep learning.

Dr. Zhang is a Senior Program Committee Member of the AAAI Conference on Artificial Intelligence and the International Joint Conference on Artificial Intelligence and a Member of ACM. He serves as a reviewer for many journals and conferences.

Dacheng Tao (Fellow, IEEE) is the President of the JD Explore Academy, China, and a Senior Vice President of JD.com. He is an Advisor and the Chief Scientist of the Digital Science Institute, The University of Sydney, Sydney, NSW, Australia. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences.

Dr. Tao is a fellow of the Australian Academy of Science, AAAS, and ACM. He was a recipient of the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award.