frontiers | Frontiers in Earth Science

TYPE Methods
PUBLISHED 25 August 2022
DOI 10.3389/feart.2022.974325

OPEN ACCESS

EDITED BY
Xinghua Li,
Wuhan University, China

REVIEWED BY
Xu Xia,
Nankai University, China
Linwei Yue,
China University of Geosciences
Wuhan, China

*CORRESPONDENCE
Congan Xu,
xcatougao@163.com

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Earth Science

RECEIVED 21 June 2022
ACCEPTED 18 July 2022
PUBLISHED 25 August 2022

CITATION
Wu J, Tang Z, Xu C, Liu E, Gao L and Yan W (2022), Super-resolution domain adaptation networks for semantic segmentation via pixel and output level aligning.
Front. Earth Sci. 10:974325.
doi: 10.3389/feart.2022.974325

Super-resolution domain adaptation networks for semantic segmentation *via* pixel and output level aligning

Junfeng Wu[1], Zhenjie Tang[2,3], Congan Xu[1]*, Enhai Liu[2,3], Long Gao[1] and Wenjun Yan[1]

[1]Remote Sensing Image Interpretation Research Group, Naval Aviation University, Yantai, Shandong, China, [2]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China, [3]Hebei Province Key Laboratory of Big Data Calculation, Tianjin, China

Recently, unsupervised domain adaptation (UDA) has attracted increasing attention to address the domain shift problem in the semantic segmentation task. Although previous UDA methods have achieved promising performance, they still suffer from the distribution gaps between source and target domains, especially the resolution discrepancy in the remote sensing images. To address this problem, this study designs a novel end-to-end semantic segmentation network, namely, Super-Resolution Domain Adaptation Network (SRDA-Net). SRDA-Net can simultaneously achieve the super-resolution task and the domain adaptation task, thus satisfying the requirement of semantic segmentation for remote sensing images, which usually involve various resolution images. The proposed SRDA-Net includes three parts: a super-resolution and segmentation (SRS) model, which focuses on recovering high-resolution image and predicting segmentation map, a pixel-level domain classifier (PDC) for determining which domain the pixel belongs to, and an output-space domain classifier (ODC) for distinguishing which domain the pixel contribution is from. By jointly optimizing SRS with two classifiers, the proposed method can not only eliminate the resolution difference between source and target domains but also improve the performance of the semantic segmentation task. Experimental results on two remote sensing datasets with different resolutions demonstrate that SRDA-Net performs favorably against some state-of-the-art methods in terms of accuracy and visual quality. Code and models are available at https://github.com/tangzhenjie/SRDA-Net.

Abbreviations: AAN, appearance adaptation network; ASPP, Atrous Spatial Pyramid Pooling; CNNs, convolutional neural networks; CycleGan, cycle generative adversarial network; FCAN, Fully Convolutional Adaptation Networks; FCN, fully convolutional network; GANs, generative adversarial networks; IoU, intersection-over-union; MSE, mean squared error; MSI, multispectral image; PatchGAN, patch generative adversarial network; PDC, pixel-level domain classifier; ODC, output-space domain classifier; RAN, representation adaptation networks; SRDA-Net, Super-Resolution Domain Adaptation Network; SRS, super-resolution and segmentation; and UDA, unsupervised domain adaptation.

KEYWORDS

remote sensing, semantic segmentation, domain adaptation, super resolution, deep learning

1 Introduction

Remote sensing imagery semantic segmentation, aiming at assigning a semantic label for each pixel, has enabled various high-level applications, such as land-use survey, urban planning, and environmental protection (Zheng et al., 2017; Pan B. et al., 2019; Mou et al., 2020). Deep convolutional neural networks (CNNs) have already shown amazing performance in the semantic segmentation task (Long et al., 2015; Chen et al., 2018; Wang Q. et al., 2019; Pan B. et al., 2020). To guarantee the superior representation ability, CNNs usually require a large number of manually labeled training data. However, the manually annotating process for each pixel is time-consuming and labor-intensive.

UDA tries to learn a well-performed model for the target domain only under the supervision of the source data and has become a powerful technology to handle the problem of insufficient labeling. Most UDA-related works focus on aligning features of source and target domains in a deep network by extracting domain-invariant features (Zhang et al., 2018; Wu et al., 2019). In recent years, some works begin seeking to minimize the domain shift at the pixel level, by means of turning source domain images into target-like images by adversarial training (Zhang et al., 2018; Li et al., 2019). In addition, some studies are proposed to address this problem by reducing the spatial structure domain discrepancies in the output space (Tsai et al., 2018; Vu et al., 2019).

However, these typical algorithms mainly address the semantic segmentation problem on natural scene image, and the performance would be influenced when applied on remote sensing images because of the spatial resolution difference. Spatial resolution (Pan Z. et al., 2019; Liu et al., 2019) is one of the important characteristics of remote sensing images. Unlike natural scene images, the sensors used to acquire remote sensing images usually have significant differences, which results in different spatial resolutions. For the same object, there are often large differences in resolution in remote sensing images obtained by different sensors. For example, a car in a 4 m-resolution remote sensing image can never be the same size as a car in a 1 m-resolution image, which has a great impact on domain adaptation semantic segmentation. On the other hand, if we only considered UDA for remote sensing images with the same resolution (Liu and Su, 2020; Tasar et al., 2020), the available data should be severely compressed. Therefore, we may conclude that UDA for remote sensing images should not only narrow the gaps between source and target domains but also address the issue of different resolutions.

To the best of our knowledge, there are few UDA algorithms for remote sensing images that explicitly consider the resolution problem. The existing algorithms usually neglect the resolution problem when the resolution differences between the source and target domains are not obvious (Yan et al., 2020; Jun et al., 2020) or deal with the problem by simple interpolation (Zhaoxiang et al., 2021) or adjust the parameters of kernel function (Liu and Qin, 2020). For instance, Yan et al. (2020) proposed a triplet adversarial domain adaptation method to learn a domain-invariant classifier in output space by a novel domain discriminator, without considering the resolution problem between the source and target domains. Instead of matching the distributions in output space, Zhaoxiang et al. (2021) proposed to eliminate the domain shift by aligning the distributions of the source and target data in the feature space, where the resolution problem was dealt with interpolation. Liu and Qin (2020) minimized the feature distributions distance between the source and target domains through metric under different kernel functions, which reduced the effect of resolution problem by adjusting the parameters of kernel function. However, the existing UDA methods for remote sensing images have not explicitly studied the resolution problem.

In this article, explicitly considering the resolution problem, a novel end-to-end network is designed, which can simultaneously conduct Super-Resolution and Domain Adaptation, to improve the segmentation performance from low-resolution remote sensing data to high-resolution remote sensing data. Figure 1 briefly depicts the problem setting: source domain (low-resolution remote sensing images) with labels and target domain (high-resolution remote sensing images) without labels. SRDA-Net is motivated by two recent research works: 1) super-resolution and semantic segmentation can promote each other, and 2) adversarial training-based UDA methods for semantic segmentation. Recently, some studies have shown that super-resolution and semantic segmentation can boost each other. For instance, researchers indicate that super-resolution results can be improved by semantic priors, such as semantic segmentation probability maps (Wang et al., 2018) or segmentation labels (Rad et al., 2019). In the field of remote sensing, high-resolution images contain more detailed information, and this is very important for image segmentation (Lei et al., 2019). Lei et al. (2019) proposed to embed image super-resolution into the segmentation network to improve the performance on both super-resolution and segmentation tasks. Furthermore, most of the UDA methods successfully reduce the domain discrepancies drawing support from the adversarial training. For instance, Zhang et al. (2018) applied the adversarial loss to the lower layers of the segmentation network because the lower layers mainly capture the appearance information of the images. Tsai et al. (2018)

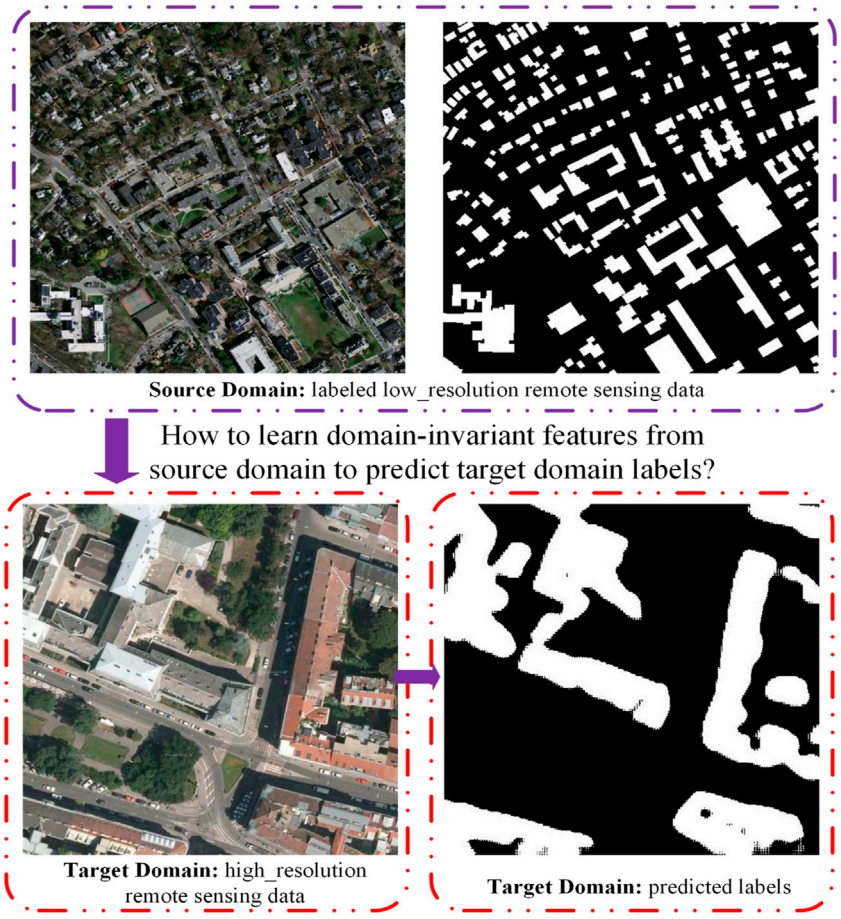

FIGURE 1
Description of the problem setting: given a source domain composed of labeled low-resolution remote sensing data, and a target domain made up of unlabeled high-resolution remote sensing data, this task intends on predicting the label map for image from the target domain using a semantic segmentation model trained by source domain images.

employed the adversarial feature learning in the output space over the base segmentation model. Vu et al. (2019) also reduced the discrepancies of feature distributions in output space using adversarial entropy minimization. To be specific, the SRDA-Net consists of three networks: a multi-task model for super-resolution and semantic segmentation (SRS), a pixel-level domain classifier (PDC), and an output-space domain classifier (ODC). By integrating a super-resolution network and a segmentation network into one architecture, SRS can eliminate the resolution gap between the source and target domains, and further enhances the semantic segmentation capability. PDC is fed with high-resolution images generated by the super-resolution network, and outputs their domain (source or target domain) for each pixel. ODC is fed with the predicted label distributions from the segmentation network, and then outputs the domain class for each pixel label distribution. Similar to generative adversarial networks (GANs) (Goodfellow et al., 2014), the SRS model can be regarded as a generator, while PDC/ODC models can be treated as two discriminators. Through the adversarial training, the SRS model can learn domain-invariant features at both the pixel and output-space levels.

To summarize, the major contributions of SRDA-Net can be stated as follows:

- A new UDA method named SRDA-Net is proposed for semantic segmentation, to adapt the changes from low-resolution remote sensing images to high-resolution remote sensing images.
- A multi-task model composed of super-resolution and segmentation is built, which not only eliminates the resolution difference between the source and target domains but also obtains improvements on the semantic segmentation task.
- Two domain classifiers are designed at the pixel level and output space, to pursue domain alignment. With the help

of adversarial training, the domain gap can be effectively reduced.

2 Related works

This section briefly reviews some important works about semantic segmentation, single image super resolution, and unsupervised domain adaptation.

2.1 Semantic segmentation

Semantic segmentation aims to assign a semantic label to each pixel in an image. It plays an important role in many fields, such as autonomous driving and urban planning. In 2014, fully convolutional network (FCN) (Long et al., 2015) presents amazing performance in some pixel-wise tasks (such as semantic segmentation). After that, the models based on FCN have made significant improvements on several segmentation benchmarks (Maggiori et al., 2017, (Badrinarayanan et al., 2016, Ronneberger et al., 2015). Some model variants are then proposed to exploit the contextual information by adopting multi-scale inputs (Chen et al., 2014, 2018, Ding et al., 2020b) or employing probabilistic graphical models (Zheng et al., 2017). For instance, Chen et al. (2014, 2018) proposed a dilated convolution operation to aggregate multi-scale contextual information. Ding et al. (2020b) introduced a two-stage multi-scale training strategy to incorporate enough context information. In order to describe objects consistently, Zheng et al. (2020) proposed a standalone end-to-end edge-aware neural network (EaNet) for urban scene semantic segmentation. Moreover, the attention mechanism is also utilized for semantic segmentation (Fu et al., 2019, Ding et al., 2020a).

2.2 Single image super resolution

Single image super resolution (Freeman and Pasztor, 1999) attempts to recover high-resolution images from the corresponding low-resolution ones, which has been applied broadly in many occasions, such as product quality inspection, medical diagnosis, and remote sensing image reconstruction. Given the HR image I_y and the degradation function D, the LR image I_x can be obtained by the following degradation process:

$$I_x = D\left(I_y; \delta\right), \tag{1}$$

where δ is the parameter of the degradation function. The Single image super resolution process is as follows:

$$\hat{I_y} = F\left(I_x; \theta\right), \tag{2}$$

where F is the super resolution model, and θ is the parameter.

The conventional non-CNNs method mainly focuses on the domain and feature priors. For example, interpolation methods such as bicubic and Lanczos generate the high-resolution pixels by the weighted average of neighboring low-resolution pixels. However, CNN-based methods (Haut et al., 2018; Arun et al., 2020; Mei et al., 2020; Liu et al., 2021; Jiang et al., 2020) consider the super resolution as a mapping from the low-resolution space to high-resolution space in an end-to-end manner, showing great breakthrough. For example, Arun et al. (2020) designed a 3-D super resolution neural network for hyperspectral images. Han et al. (2019) proposed a multi-level and multi-scale to solve the super-resolution problem of multispectral image (MSI). Wei et al. (2020) utilized the deep unfolding technique to construct the network. Lei and Shi (2022) proposed a new hybrid-scale self-similarity exploitation network for remote sensing image SR. Moreover, some researchers proposed the perceptual loss (Johnson et al., 2016) and adversarial training (Lei et al., 2020; Li et al., 2020) to improve perceptual quality of super resolution result.

2.3 Unsupervised domain adaptation

Since the distributions of source domain and target domain data are different, we find a measure criterion defined on feature space to make the source domain and target domain data as close as possible. Then the predictive function based on the source domain data can be utilized to the target domain data.

We denote that $\mathcal{X}$ is the instance set, $\mathcal{Z}$ is the feature set, $\mathcal{D}_S$ and $\widetilde{\mathcal{D}_S}$ are the distributions of source domain data defined on $\mathcal{X}$ and $\mathcal{Z}$, and $\mathcal{D}_T$ and $\widetilde{\mathcal{D}_T}$ are the distributions of target domain data defined on $\mathcal{X}$ and $\mathcal{Z}$. The $\mathcal{H}$ distance, expressed as Eq. 3, is generally utilized in most methods to measure the distance between two domains.

$$d_{\mathcal{H}}\left(\widetilde{D_S}, \widetilde{D_T}\right) = 2 \sup_{h \in \mathcal{H}} |P_{\widetilde{D_S}}[I(h)] - P_{\widetilde{D_T}}[I(h)]|, \tag{3}$$

where h is the predictive function, and $\mathcal{H}$ is the set of h.

The work mainly focuses on visual semantic segmentation, the review of UDA is limited in this task as well. Many UDA-based segmentation approaches (Zhang et al., 2018, Wu et al., 2019, Lee et al., 2019, Tsai et al., 2019) use adversarial training to minimize cross-domain discrepancy in the feature space. Some works (Tsai et al., 2018, Vu et al., 2019) propose to align the predicted label distributions in the output space. Tsai et al. (2018) carried out the alignment on the prediction of the segmentation network, and Vu et al. (2019) proposed to do it on entropy minimization of the prediction probability. In contrast, pixel-level domain adaptation (Zou et al., 2020, Tasar et al., 2020, Li et al., 2019) makes use of generative networks to turn source domain images into target-like images. Li et al. (2019) presented a bidirectional learning system for semantic segmentation, which is a closed loop to learn the segmentation adaptation model and the image translation model

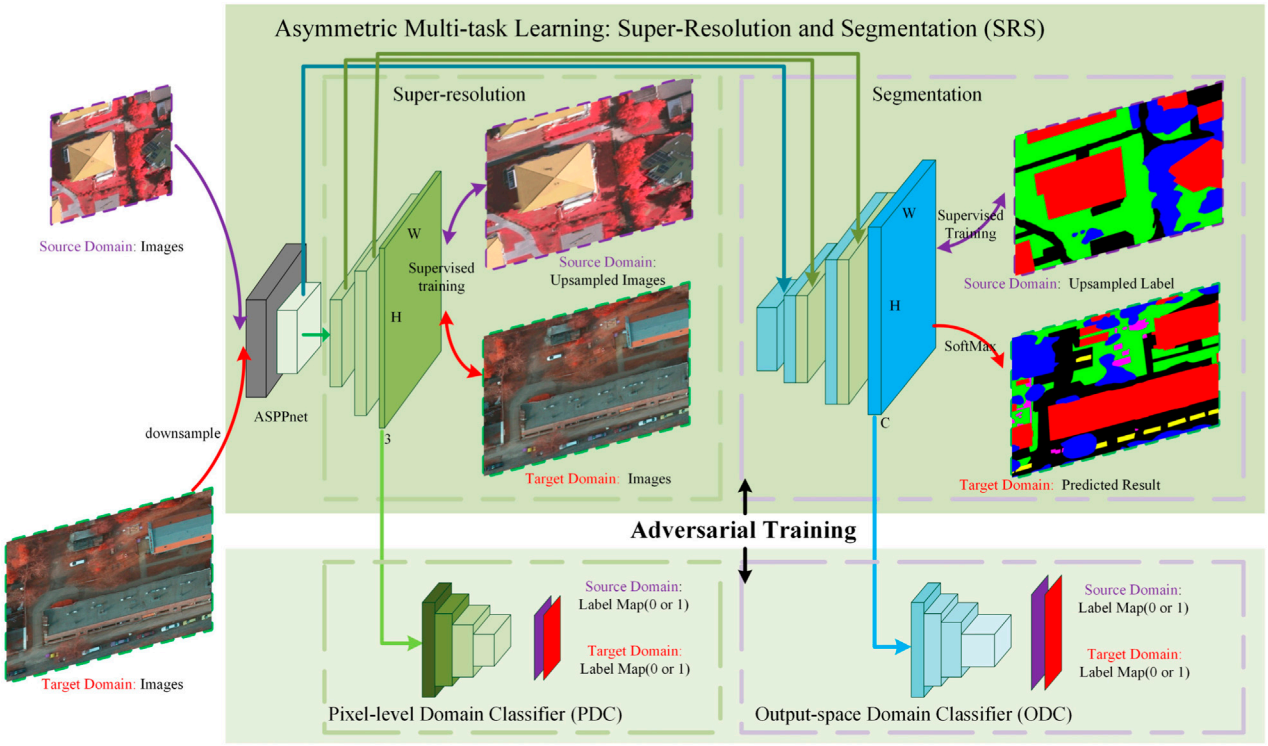

FIGURE 2
Overview of the proposed Super-Resolution Domain Adaptation Network (SRDA-Net). The two models, super-resolution model and the segmentation model, are integrated together based on asymmetric multi-task learning, as presented in the upper part. At the stage of training, two images randomly selected from source domain and target domain, respectively, are input into the SRS model. The purple and red curves, respectively, indicate the input/output of the source and target domains, and, the two-way arrow indicates the data flow involved in the training process. Note that source images are used for training the super-resolution model and segmentation model in the supervised mode, while the target images only participate the super-resolution training. The two domain classifiers, PDC and ODC, are displayed in left and right bottom parts, respectively. The super-resolution images from SRS are fed into PDC, and the predicted label distributions from SRS are fed into ODC. The SRS can be optimized by adversarially training SRS and the two classifiers. In the testing stage, the downsampled test images are input into the SRS for predicting the segmentation maps.

alternatively, causing the domain gap to be reduced gradually at the pixel level. In addition, a curriculum learning strategy is proposed in the study of Zhang et al. (2019) by leveraging information from global label distributions and local super-pixel distributions of the target domain. Moreover, self-supervised learning approach (Pan F. et al., 2020) is usually used in UDA.

3 The proposed approach

In this section, we discuss the methodology of the proposed SRDA-Net, and the overall framework is shown in Figure 2. In order to reduce the resolution domain gap, we integrate the super-resolution into the segmentation model to eliminate the impact of different resolutions. By optimizing the SRS as well as two domain classifiers (PDC and ODC) with adversarial optimizing, the domain gap in pixel-level and output-space can be gradually reduced, thus improving the performance.

3.1 Problem description

It is worthy of defining cross-domain semantic segmentation with mathematical notations, before illustrating the method in detail. Formally, let us suppose $\mathcal{S}$ as a source domain from a low-resolution remote sensing dataset, where low-resolution images $I_{\mathcal{S}}$ and pixel-level annotations $A_{\mathcal{S}}$ are provided; $\mathcal{T}$ as a target domain from high-resolution remote sensing dataset, which only provides high-resolution images $I_{\mathcal{T}}$. Note that the label space of $\mathcal{S}$ and $\mathcal{T}$ is the same, denoted as $\mathbb{R}^C$, where C denotes the number of categories. In a word, given $I_{\mathcal{S}}$, $A_{\mathcal{S}}$, and $I_{\mathcal{T}}$, our goal is to reduce the domain gap (including resolution difference) between $\mathcal{S}$ and $\mathcal{T}$, and learn a segmentation model to predict pixel-wise category of $\mathcal{T}$. In the following, we first describe the asymmetric multi-task (super-resolution and segmentation) model. Then, the adversarial domain adaptation (pixel level and output space) is presented in details.

3.2 Multi-task model: Super-resolution and segmentation

Over the past few years, CNN-based methods have been widely applied to solve the semantic segmentation problem. However, those methods may perform worse when generalizing to the unseen images, especially the domain gap between the training (source domain) and test (target domain) images are obvious. This problem is critical for remote sensing images because the resolution of them usually changes dramatically, which seriously affects the generalization ability of the segmentation models. Therefore, it is important to eliminate the resolution difference for cross-domain semantic segmentation in remote sensing.

Recently, some studies (Wang et al., 2018; Lei et al., 2019) show that super-resolution and semantic segmentation can boost each other. Although super-resolution and segmentation are two different and challenging tasks, they may have certain relationship. Super-resolution can provide images with more details, which is helpful for improving the segmentation accuracy. Label maps from segmentation dataset or semantic segmentation probability maps may contribute to recover textures faithful to semantic classes during the super-resolution process.

Based on the earlier discussions, we propose a novel model based on the aysmmetric multi-task learning, which consists of super-resolution and segmentation models. In order to make super-resolution and segmentation boost each other, two strategies are adopted: 1) introducing a pyramid feature fusion structure between the two tasks; and 2) imposing the cross-entropy segmentation loss to train the segmentation network, for the generated high-resolution images of the source domain. During training, a source domain image, pairing with a downsampled target domain image, is taken as the input to the SRS network. The source domain images are used to train both the super-resolution network and the segmentation network, while the target images only participate in the super-resolution training process, shown in Figure 2. At the testing stage, the downsampled test images are input into the SRS network to obtain the pixel-wise scope maps.

To be specific, due to GPU memory limitation, we use the residual Atrous Spatial Pyramid Pooling (ASPP) Module (Wang L. et al., 2019) as the shared feature extractor. For the super-resolution model, we only use a few deconvolutions to recover the high-resolution images, without using PixelShuffle (Shi et al., 2016). To transfer the low-level features effectively from super-resolution stream to segmentation stream, we introduce the pyramid feature fusion structure (Lin et al., 2017) between the two streams. Moreover, the super-resolution results of the source domain are also fed to the segmentation stream. Meanwhile, the segmentation stream also ensures to recover textures faithful to semantic classes during super-resolution stream.

The proposed SRS model is optimized through the following loss:

$$\begin{aligned}
\mathcal{L}_{SRS} &= \alpha\mathcal{L}_{seg} + \beta(\mathcal{L}_{idT} + \mathcal{L}_{idS}) \\
\mathcal{L}_{seg} &= \mathcal{L}_{cel}(\mathbf{S}(I_S), \uparrow A_S) + \mathcal{L}_{cel}(\mathbf{S}(\downarrow \mathbf{R}(I_S)), \uparrow A_S) \\
\mathcal{L}_{idT} &= \mathcal{L}_{mse}(\mathbf{R}(\downarrow I_T), I_T) \\
\mathcal{L}_{idS} &= \mathcal{L}_{per}(\mathbf{R}(I_S), \uparrow I_S) + 0.5 \times \mathcal{L}_{fp} \\
\mathcal{L}_{fp} &= \mathcal{L}_{L1}(\mathbf{E}(\downarrow \mathbf{R}(I_S)), \mathbf{E}(I_S)),
\end{aligned} \tag{4}$$

where $\mathcal{L}_{cel}$ represents the 2D cross-entropy loss, the standard supervised pixel-wise classification objective function (Wang Q. et al., 2019); $\mathcal{L}_{mse}$ is the pixel-wise mean squared error (MSE) loss, which is widely applied to optimize the objective function for image super resolution; $\mathcal{L}_{per}$ is the perceptual loss (Johnson et al., 2016); $\mathcal{L}_{L1}$ represents the L1 norm loss; and $\mathcal{L}_{fp}$ is the fixpoint loss (Kotovenko et al., 2019). The $\uparrow$ and $\downarrow$ denote upsampling and downsampling operations, respectively. $\mathbf{S}$, $\mathbf{R}$, and $\mathbf{E}$ denote segmentation model, super-resolution model, and the shared feature extractor, respectively. Note that, in order to easily superimpose the style of the target domain and stabilize the adversarial training process, we use $\mathcal{L}_{per}$ and $\mathcal{L}_{fp}$ to train the super-resolution model of source domain images. α and β denote the weighting factors for semantic segmentation and super-resolution, respectively.

3.3 Adversarial domain adaptation

Although the proposed asymmetric multi-task model can eliminate the resolution difference between the source and target domains, some other domain gaps (e.g., color, texture *etc.*) still exist. Affected by various human and natural factors, such as sensors, weather conditions, and imaging locations, these differences are inherent in remote sensing imagery. Therefore, how to learn the domain-invariant features for remote sensing imagery is a critical problem.

An effective framework to deal with the aforementioned problem is adversarial learning. It consists of two main parts: a generator network and a discriminator network. Its main idea is to train the discriminator to predict the domain label of the data, while the generator network attempts to fool it, as well as implements the segmentation task on source domain data. Through training the two networks alternately, the feature domain gap can be gradually reduced, thus obtaining the domain-invariant representations.

The proposed method also takes the adversarial learning to alleviate the domain gap. Specially, the PDC and ODC are designed as discriminators, and the SRS is adopted as the generator. By the adversarial training, SRS will learn the domain-invariant features that fool the PDC and ODC.

3.4 Pixel-level adaptation

The proposed SRS model eliminates the resolution difference between the source and target domains, while it does not reduce the gap in other aspects. To address this problem ulteriorly, PDC is designed to receive the high-resolution images from the source or target domain and classify the domain for each pixel. Concretely, the PatchGAN (Li and Wand, 2016) is applied as PDC, and the network architecture is shown at the bottom left of Figure 2.

The loss objective of PDC can be formulated as following:

$$\begin{aligned}\mathcal{L}_{PDC} &= \mathbb{E}_{I_{fake}\sim p_{\text{data}}(I_{fake})}\left[\left(I_{fake}-1\right)^2\right]+\mathbb{E}_{I_{true}\sim p_{\text{data}}(I_{true})}\left[\left(I_{true}\right)^2\right]\\ I_{true} &= \mathbf{D}_{pdc}(I_{\mathcal{T}})\in\mathbb{R}^{H\times W\times 1}\\ I_{fake} &= \mathbf{D}_{pdc}(I_{\mathcal{S}}^{R})\in\mathbb{R}^{H\times W\times 1}\\ I_{\mathcal{S}}^{R} &= \mathbf{R}(I_{\mathcal{S}})\in\mathbb{R}^{H\times W\times 3},\end{aligned} \tag{5}$$

where $\mathbf{D}_{pdc}$ is the PDC model, H and W denote the height and width of the high-resolution target domain image, respectively.

Accordingly, the inverse of PDC loss is calculated by:

$$\mathcal{L}_{PDC_{inv}} = \mathbb{E}_{I_{true}\sim p_{\text{data}}(I_{true})}\left[\left(I_{true}-1\right)^2\right]+\mathbb{E}_{I_{fake}\sim p_{\text{data}}(I_{fake})}\left[\left(I_{fake}\right)^2\right]. \tag{6}$$

Finally, the adversarial objective functions is given as:

$$\min_{\theta_{SRS}}\mathcal{L}_{SRS}+\mathcal{L}_{PDC}, \tag{7}$$

$$\min_{\theta_{PDC}}\mathcal{L}_{PDC_{inv}}, \tag{8}$$

where θ_{SRS} and θ_{PDC} denote the network parameters of SRS and PDC, respectively. During the training phase, the parameters of the two models are updated in turns using Eq. 7 and Eq. 8.

3.5 Output-space adaptation

Different from the image classification task that is based on global features, the generated high-dimensional features for the semantic segmentation encode complex detailed representations, which will result in contextual relationships among neighboring pixels. Therefore, adaptation only in the pixel space may not be enough for semantic segmentation. On the other hand, although segmentation outputs are in the low-dimensional space, they contain rich information, for example, scene layout and context. Moreover, in the segmentation task of remote sensing, images from the source or target domain should share strong similarities both in spatial and local representations. For example, the rectangular road region may cover the part of cars, pedestrians, and green plants that often grow around the buildings. Thus, we adapt the low-dimensional softmax outputs of segmentation predictions *via* an adversarial learning scheme.

To be specific, we design ODC to distinguish domain source for the distribution of pixels, which receives the segmentation softmax output: $P=\mathbf{S}(I)\in\mathbb{R}^{H\times W\times C}$, where C is the number of categories. We forward P to ODC using a cross-entropy loss $\mathcal{L}_{ODC}$ for the two classes (i.e., source and target). The ODC loss can be written as:

$$\begin{aligned}\mathcal{L}_{ODC} &= -\sum_{h,w}(1-z)\log\left(P_{fake}\right)+z\log\left(P_{true}\right)\\ P_{fake} &= \mathbf{D}_{odc}(P_{\mathcal{T}})\in\mathbb{R}^{H\times W\times 1}\\ P_{true} &= \mathbf{D}_{odc}(P_{\mathcal{S}})\in\mathbb{R}^{H\times W\times 1}\\ P_{\mathcal{T}} &= \mathbf{S}(\downarrow I_{\mathcal{T}})\in\mathbb{R}^{H\times W\times C}\\ P_{\mathcal{S}} &= \mathbf{S}(I_{\mathcal{S}})\in\mathbb{R}^{H\times W\times C},\end{aligned} \tag{9}$$

where $\mathbf{D}_{odc}$ denotes the ODC model.

Accordingly, the inverse of ODC loss is defined as:

$$\mathcal{L}_{ODC_{inv}} = -\sum_{h,w}(1-z)\log\left(P_{true}\right)+z\log\left(P_{fake}\right), \tag{10}$$

In the end, the adversarial objective functions are expressed as follows:

$$\min_{\theta_{SRS}}\mathcal{L}_{SRS}+\mathcal{L}_{ODC}, \tag{11}$$

$$\min_{\theta_{ODC}}\mathcal{L}_{ODC_{inv}}, \tag{12}$$

where θ_{SRS} and θ_{ODC} represent the parameters of SRS and ODC networks, respectively. They can be optimized in turns by minimizing Eq. 11 and Eq. 12 during the training stage.

3.6 Final objective function

To initialize parameters of the network better, we first use the following loss function to pre-train the model:

$$\min_{\theta_R}\beta\left(\mathcal{L}_{idT}+\mathcal{L}_{idS}\right)+\mathcal{L}_{PDC} \tag{13}$$

$$\min_{\theta_{PDC}}\mathcal{L}_{PDC_{inv}} \tag{14}$$

where β denotes a weighting factor for super-resolution, θ_R is the parameters of **R** network. During training stage, the **R** and PDC networks are optimized in turns using Eq. 13 and Eq. 14.

For the whole models training, including SRS, PDC and ODC, the objective functions can be formulated as:

$$\min_{\theta_{SRS}}\mathcal{L}_{SRS}+\mathcal{L}_{PDC}+\mathcal{L}_{ODC}, \tag{15}$$

$$\min_{\theta_D}\mathcal{L}_{PDC_{inv}}+\mathcal{L}_{ODC_{inv}}, \tag{16}$$

where θ_D denotes the network parameters of PDC and ODC. During the training phase, the parameters of SRS, PDC, and ODC are optimized in turns by minimizing Eq. 15 and Eq. 16. Eq. 15 and Eq. 16 together constitute the adversarial training as the generator and discriminator loss. The training procedure of our proposed SRDA-Net is illustrated in Algorithm 1.

Require:
Source Domain low-resolution image I_S, Target Domain high-resolution image I_T, Source Domain low-resolution label A_S, The weighting factors for semantic segmentation and super-resolution: α, $\beta = 10$.
Ensure:
High-resolution source domain image with style of target domain: I_S^R
Predict label of target domain: A_T
repeat
% Super-Resolution images by the **R** model
$I_S^R, I_T^R = \mathbf{R}(I_S, \downarrow I_T) \in \mathbb{R}^{H\times W\times 3}$
% Segmentation softmax outputs from the **S** model
$P_S, P_T = \mathbf{S}(I_S, \downarrow I_T) \in \mathbb{R}^{H\times W\times C}$
% Predict label of target domain image
$A_T = max(P_T) \in \mathbb{R}^{H\times W\times 1}$
% Distinguish the pixels of super-resolution images
$I_{fake}, I_{true} = \mathbf{D}_{pdc}(I_S^R, I_T) \in \mathbb{R}^{H\times W\times 1}$
% Distinguish the pixel distributions of softmax outputs
$P_{true}, P_{fake} = \mathbf{D}_{odc}(P_S, P_T) \in \mathbb{R}^{H\times W\times 1}$
% Adversarial training
R and **S** can be optimized according to equation (23).
$\mathbf{D}_{pdc}$ and $\mathbf{D}_{odc}$ are updated by minimizing the inverse loss (24).
until convergence

Algorithm 1. the proposed SRDA-Net.

4 Experimental results

In this section, we validate the performance of the proposed SRDA-Net. First, the experimental dataset and implementation details are described, and then, the experimental results are reported and analyzed to demonstrate the effectiveness of SRDA-Net. Finally, the two strategies to achieve mutual promotion of super-resolution and segmentation in SRS are discussed.

4.1 Datasets description

1) *Mass-Inria:* the following two UDA datasets are used for single-category semantic segmentation.

- Massachusetts Buildings Dataset (Volodymyr, 2013) contains 151 aerial images of the Boston area at 1 m spatial resolution. The ground truth provides two semantic classes: building and non-building. The whole dataset is divided into three parts: a training set with 137 images, a testing set with 10 images, and a validation set with four images. Among these sets, the training set is considered as the source domain.
- Inria Aerial Image Labeling Dataset (Maggiori et al., 2017) is composed of 360 tiles with a resolution of 0.3 m on 10 cities across the globe. Ground truth provides two semantic classes, *building* and *non-building* classes. We split the training set (image 1 to 5 of each location for validation, and 6 to 36 of each location for training). We consider the training set of this dataset as the target domain. We finally validate the results of the algorithm on the validation set of this dataset.

2) *Vaih-Pots:* we use the following two UDA datasets for multi-category semantic segmentation.

- ISPRS Vaihingen 2D Semantic Labeling Challenge contains 33 images of different sizes at 9 cm spatial resolution, taken over the city of Vaihingen (Germany). Each image consists of a true orthophoto extracted from a larger orthophoto mosaic. There are six labeled categories: impervious surface, building, low vegetation, tree, car, and clutter/background. This dataset is considered to be a source domain.
- ISPRS Potsdam 2D Semantic Labeling Challenge dataset is composed of 38 ortho-rectified aerial IRRGB images with a size of 6,000, ×, 6,000 at 5 cm spatial resolution, taken over the city of Potsdam in Germany. The ground truth is provided for 24 tiles alike Vaihingen dataset. We randomly choose 12 images as the training set, and other 12 images as the testing set.

Note that the resolution gap of Mass-Inria (around 3.333 times) is greater than Vaih-Pots (around 2.0 times).

4.2 Evaluation metric and implementation details

1) Evaluation metric

The intersection-over-union (IoU) is adopted as the main evaluation metric, and it is defined as:

$$\mathrm{IoU}\left(P_m, P_{gt}\right) = \left|\frac{P_m \cap P_{gt}}{P_m \cup P_{gt}}\right|, \tag{17}$$

where P_m is the prediction and P_{gt} is the ground truth. Mean IoU (mIoU) is used to evaluate model performance on all classes.

2) Implementation details

Network architectures: in SRS, due to GPU memory limitation, we choose the residual ASPP module (Wang L. et al., 2019) to capture contextual information, as the shared feature extractor. For the super-resolution stream, we only use a few deconvolutions to recover the high-resolution images. As for two discriminators, we apply the patch generative adversarial network (PatchGAN) Li and Wand (2016) classifier as the PDC Network, and for ODC network we choose is similar to Tsai et al. (2018), which consists of five convolution layers with kernel of 4 × 4 and stride of 2, where the channel number is 64, 128, 256, 512, and 1, respectively.

Training and testing details: in the training stage, Adam optimization is applied with a momentum of 0.9. For the Mass to Inria experiments, α is set to 2.5, and β set to 10. Due to different resolutions, the Mass images and labels are cropped to 114 × 114 pixels, and then the labels are interpolated to 380 × 380 pixels. The Inria images are cropped to 380 × 380 pixels

TABLE 1 Comparison results of domain adaptation from Mass to Inria val datasets.

Method %	BaseNet	Source domain	Target domain	IoU
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	Mass	Inria	32.9
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	Mass	↓ Inria	35.0
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	↑ Mass	Inria	48.5
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	Mass	Inria	32.9
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	Mass	↓ Inria	41.8
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	↑ Mass	Inria	49.7
NoAdapt	ResidualASPP Wang et al. (2019a)	Mass	Inria	31.9
SRS	ResidualASPP Wang et al. (2019a)	Mass	Inria	36.7
SRS + PDC	ResidualASPP Wang et al. (2019a)	Mass	Inria	46.0
SRS + ODC	ResidualASPP Wang et al. (2019a)	Mass	Inria	39.4
Full (SRDA-Net)	ResidualASPP Wang et al. (2019a)	Mass	Inria	52.8

and resized to 114 × 114 pixels. During the stage of testing, images from Inria are cropped to 625 × 625 patches without overlap and resized to 188 × 188 pixels. In the Vaih to Pots experiments, α and β are set to 5 and 10, respectively. The low-resolution image is cropped to 160 × 160 pixels and the high-resolution is cropped to 320 × 320 pixels during training stage. During the testing stage, images of Pots are cropped to 500 × 500 pixels without overlap and resized to 250 × 250 pixels. In the actual training process, we first pre-train the model with learning rate 2×10^{-4}. Then the framework is trained with a learning rate of 1.5×10^{-4}. For image and label, bicubic interpolation and nearest neighbor interpolation are used, respectively. Since the resolution difference between the source and target domains causes a great influence on domain adaptation, the interpolation of labels generates greater gain than error.

Our stepwise experiments:

- NoAdapt: as a contrast model, NoAdapt is directly trained without domain adaptation from the source domain to the target domain.
- SRS: based on NoAdapt model, SRS is trained to eliminate resolution gap between the source and target domains.
- SRS + PDC: based on the SRS model, PDC is added further to the training process by the adversarial learning.
- SRS + ODC: on the basis of the SRS model, ODC is introduced into the training process by the adversarial learning.
- SRDA-Net (SRS + PDC + ODC): the proposed SRDA-Net model.

Other comparison experiments:

- AdaptSegNet: in this work, Tsai et al. (2018) employ the adversarial feature learning in output space of the base segmentation model. Instead of having only one discriminator over the feature layer, Tsai et al. (2018) propose to install another discriminator on one of the intermediate layers as well.
- CycleGan-FCAN: Fully Convolutional Adaptation Networks (FCAN) (Zhang et al., 2018) is a two-stage method, where appearance adaptation networks (AANs) first adapts source domain images to appear as if drawn from the "style" in the target domain, then representation adaptation networks (RANs) attempt to learn domain-invariant representations. To better adapt the source images to appear as if drawn from the target domain, we replace AAN in FCAN with the cycle generative adversarial network (CycleGan) Zhu et al. (2017).

In the experiments, we reported their no adaptation and final results, for comparison with our stepwise experiments.

4.3 Mass → Inria

The experimental results of the methods mentioned earlier for the shift from Mass to Inria are summarized in Table 1, including AdaptSegNet (Tsai et al., 2018), CycleGan-FCAN (Zhang et al., 2018), and our stepwise experiments: NoAdapt, SRS, SRS + PDC, SRS + ODC, and SRDA-Net. The bold values denote the best scores in the corresponding column.

From Table 1, it can be seen that our proposed method (SRDA-Net) achieves the best result: IoU of 52.8%. Under the same training condition, result (52.8%) of SRDA-Net outperforms that of AdaptSegNet (best result of 48.5%) and CycleGan-FCAN (best result of 49.7%), increased by 8.87 and 6.24%, respectively. Moreover, in order to explore the effect of resolution problem on the domain adaptation results, we construct experiments on two training data settings (source

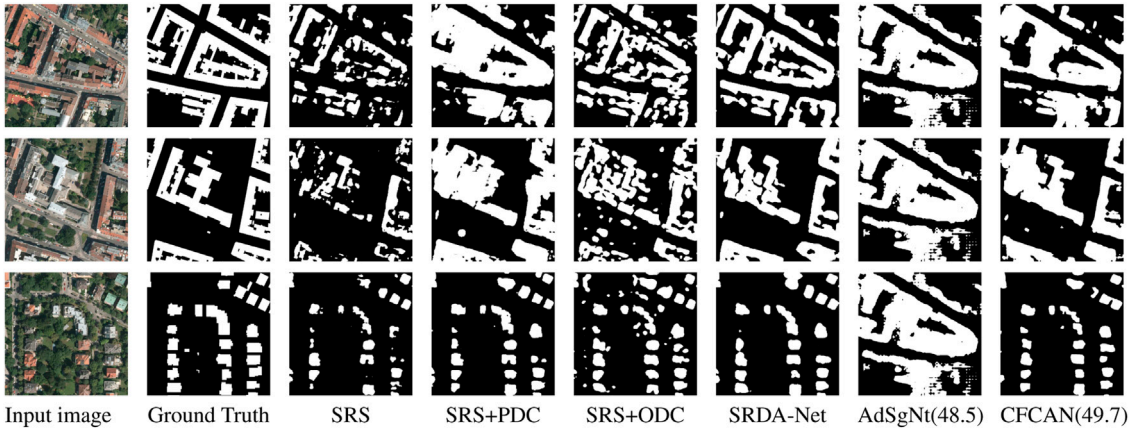

FIGURE 3
Qualitative results of the Inria val dataset (Source domain: Massachusetts Buildings).

TABLE 2 Comparison results of domain adaptation from Vaih to Pots val dataset.

Methods %	BaseNet	Source	Target	Impervious	Building	Vegetation	Tree	Car	Clutter	mIoU
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	Vaih	Pots	51.8	45.5	46.2	11.8	35.3	18.5	34.9
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	Vaih	↓ Pots	59.4	54.2	47.0	26.3	52.2	32.2	45.2
AdaptSegNet Tsai et al. (2018)	ResNet-101 He et al. (2016)	↑ Vaih	Pots	55.1	55.6	43.0	31.5	60.6	1.6	41.2
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	Vaih	Pots	51.8	45.5	46.2	11.8	35.3	18.5	34.9
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	Vaih	↓ Pots	50.1	42.5	33.1	31.6	44.1	22.6	37.3
CycleGan-FCAN Zhang et al. (2018)	ResNet-101 He et al. (2016)	↑ Vaih	Pots	47.9	51.2	43.0	41.7	61.1	23.8	44.8
NoAdapt	ResidualASPP Wang et al. (2019a)	Vaih	Pots	29.1	36.3	37.6	19.3	2.8	23.4	24.7
SRS	ResidualASPP Wang et al. (2019a)	Vaih	Pots	26.5	32.0	35.2	17.3	32.0	17.5	26.7
SRS + PDC	ResidualASPP Wang et al. (2019a)	Vaih	Pots	58.3	51.1	51.8	27.9	62.5	20.5	45.4
SRS + ODC	ResidualASPP Wang et al. (2019a)	Vaih	Pots	51.2	21.7	17.9	12.3	54.2	13.0	28.4
Full (SRDA-Net)	ResidualASPP Wang et al. (2019a)	Vaih	Pots	60.2	61.0	51.8	36.8	63.4	18.3	48.6

domain: Mass, target domain: Downsampled Inria; source domain: Upsampled Mass, and target domain: Inria) for each comparison method. From the results, we observe that the setting (source domain: Upsampled Mass and target domain: Inria) achieves best result, and the setting (source domain: Mass and target domain: Downsampled Inria) performs better than general training data setting, which confirms that when the resolution difference between the source and target domains is smaller, domain adaptation semantic segmentation networks will achieve better results. Moreover, when the resolution difference between the source and target domains is larger, the gain obtained by eliminating the resolution difference is greater than the error introduced by interpolation. In comparison, our method does not need to consider this and obtained better results owing to eliminating resolution gap by integrating super resolution into the segmentation model.

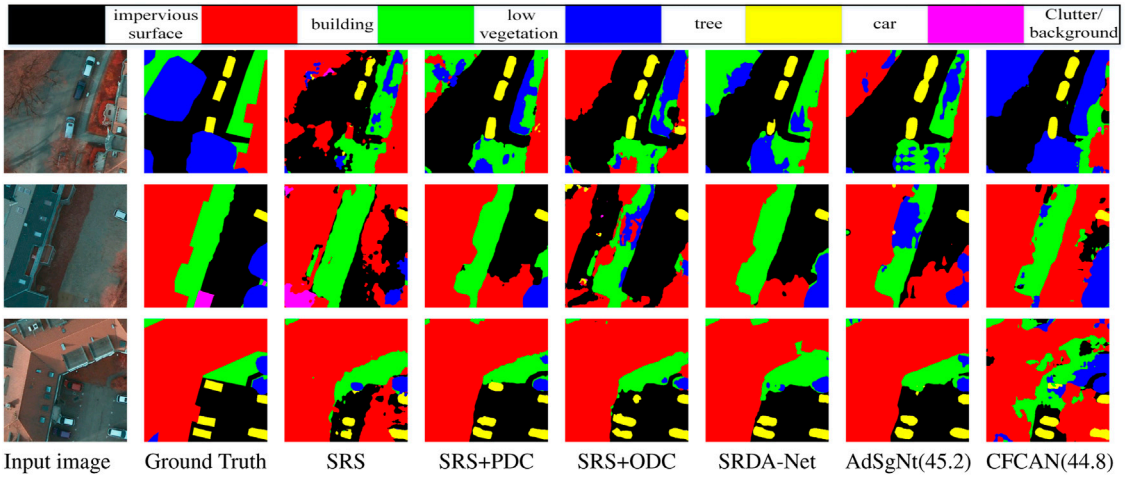

FIGURE 4
Qualitative results of the Pots val dataset (Source domain: ISPRS Vaihingen 2D semantic dataset).

As for the results compared with the baseline method, the adaptation results of the three methods outperform their corresponding results of NoAdapt. In addition, SRS improves the IoU significantly, from 31.9 to 36.7%, increasing by 4.8%, which shows the effectiveness of combining image super-resolution in the segmentation network to eliminate the resolution gap between the source and target domains.

In order to explore the semantic segmentation performance further, Figure 3 shows the visualization results of our step-by-step and the AdaptSegNet/CycleGan-FCAN methods. The images in the first column are selected from the Inria val dataset. The second column shows the ground truth, and the remaining columns illustrate the prediction results of SRS, SRS + PDC, SRS + ODC, SRDA-Net, AdaptSegNet (48.5%), and CycleGan-FCAN (49.7%). On the whole, after adding the PDC or ODC, some segmentation mistakes are removed effectively. According to the results of SRS + PDC and SRS + ODC, PDC plays a more important role in learning domain-invariant features than ODC (improvement of 9.3 vs. 2.7%). When the domain gap is reduced by integrating PDC and ODC to SRS, a better segmentation result can be obtained. Moreover, we can observe that visualization segmentation results of SRDA-Net outperform the results of best AdaptSegNet/CycleGan-FCAN.

TABLE 3 Ablation experiments of SRS.

	NoAdapt	+strategy1	+strategy2	SRS
IoU	31.9	33.5	36.2	36.7

4.4 Vaih → Pots

The results of AdaptSegNet (Tsai et al., 2018), CycleGan-FCAN (Zhang et al., 2018), and our stepwise experiments are listed in Table 2, which are adapted from Vaih to Pots. The bold fonts represent the best scores of the corresponding columns. It can be observed that the proposed method obtains the best performance with mIoU of 48.6%. Compared with the best comparative result (45.2%, obtained by AdaptSegNet), the SRDA-Net contributes 3.4% relative mIoU improvement. Under the condition of no adaptation, among the three baseline methods, mIoU of our method is slightly lower. This is because that parameters of residual ASPP module are less than half of ResNet-101, which limits the learning power of the network. According to the mIoUs of SRS+PDC and SRS + ODC, PDC (18.7%) is more effective at learning domain-invariant features than ODC (1.7%).

From the results of comparison experiments, we find that the setting (source domain: Mass and target domain: Downsampled Inria) achieves best result, and the setting (source domain: Upsampled Mass and target domain: Inria) performs better than general training data setting. However, compared with Mass → Inria, the gap of resolution between Vaih and Pots is relatively small, thus it is difficult to determine whether to upsample the source domain or downsample the target domain to obtain better results. However, there is no need to our proposed method (SRDA-Net).

For reporting the effect of our algorithm, Figure 4 gives the three typical example labeling results. From the visual results, we

e. Qualitative results of super-resolution with style transfer from CycleGAN and SRS+PDC

FIGURE 5
Qualitative results of super-resolution with style transfer from CycleGAN and SRS+PDC.

can find that the segmentation results are getting more and more refined in our step-by-step experiments, and our SRDA-Net obtains the finer segmentation results.

4.5 Study of two strategies in SRS

To make super-resolution and segmentation promote each other, we propose two strategies in the SRS model: 1) a pyramid feature fusion structure between the two tasks; 2) a cross-entropy segmentation loss is applied to the generated high-resolution source domain images to train the segmentation network. In this section, we construct ablation experiments of SRS and SRS+PDC vs. CycleGan experiments to illustrate the effectiveness of two strategies. From Table 3, we can observe that both strategies improve the segmentation performance compared to the baseline model (NoAdapt). The SRS (strategy1 + strategy2) achieves the best segmentation accuracy, which shows that both strategies can transfer detailed information from super-resolution to improve segmentation performance.

4.6 SRS+PDC vs. CycleGan

The SRS+PDC module is essentially a super-resolution style transfer model, which improves the effect of semantic segmentation. As shown in the results in Table 1 and Table 2, SRDA-Net performs better than CycleGan-FCAN because the SRS+PDC module reconstructs more texture information. For further clarification, the qualitative

super-resolution results of source domain images with style transfer from CycleGan and SRS+PDC are shown in Figure 5. It can be observed that CycleGan generates monotonous and unnatural textures, like buildings in Figure 5. Moreover, we find that some objects in the results of CycleGan get distorted, like cars in Figure 5. The reason is that the upsampled source domain images are blurry, which drops some information and confirms that SRS+PDC captures the characteristics of images and produce more natural and realistic textures to help improve semantic segmentation results.

5 Conclusion

In this article, we propose a novel end-to-end framework named SRDA-Net to explicitly address the resolution adaptation problem in the field of semantic segmentation. SRDA-Net can simultaneously deal with the super-resolution task and the domain adaptation task, thus meeting the requirement of semantic segmentation for remote sensing images, which usually involve various resolution images. To be specific, a multi-task model is built to simultaneously accomplish the semantic segmentation, as well as eliminate the difference in resolution between the source and target domains. By means of the adversarial learning, the pixel level and output space domain classifiers are designed to guide the SRS model to learn domain-invariant features, which can eliminate the domain gap effectively. In order to verify the effectiveness of the proposed method, two datasets are constructed, which have different resolutions in their source and target domains: Mass-Inria and Vaih-Pots. Extensive experiments demonstrate the effectiveness of SRDA-Net when domain adaptation involving the resolution difference.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, JW and ZT; methodology, JW and EL; software, JW and ZT; validation, JW, EL, and CX; formal analysis, CX and JW; investigation, JW and CX; resources, JW and EL; data curation, ZT; writing—original draft preparation, ZT and JW; writing—review and editing, JW and L.G.; visualization, JW and EL; supervision, JW and CX; project administration, WY; funding acquisition, CX and WY. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Beijing-Tianjin-Hebei Basic Research Cooperation Project under the grant number F2021203109, National Natural Science Foundation of China under the grant number 62001251, 62001252, 61790550, 61790554, and 61971432, and the grant number of young Elite Scientists Sponsorship Program by CAST is 2020-JCJQ-QT-011.

Acknowledgments

The authors thank all reviewers for their careful reading and valuable suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Arun, P. V., Buddhiraju, K. M., Porwal, A., and Chanussot, J. (2020). Cnn-based super-resolution of hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* 58, 6106–6121. doi:10.1109/TGRS.2020.2973370

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2016). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi:10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *Comput. Sci.* 4, 357–361.

Ding, L., Tang, H., and Bruzzone, L. (2020a). Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 59, 426–435. doi:10.1109/TGRS.2020.2994150

Ding, L., Zhang, J., and Bruzzone, L. (2020b). Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture. *IEEE Trans. Geosci. Remote Sens.* 58, 5367–5376. doi:10.1109/TGRS.2020.2964675

Freeman, W. T., and Pasztor, E. C. (1999). "Learning low-level vision," in Proceedings of the Seventh IEEE International Conference on Computer Vision (Kerkyra, Greece: IEEE), 25–47. doi:10.1109/ICCV.1999.790414

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA: IEEE), 3141–3149. doi:10.1109/CVPR.2019.00326

Goodfellow, I. J., Pouget-Abadiey, J., Mirza, M., Xu, B., Warde-Farley, D., Ozairz, S., et al. (2014). "Generative adversarial nets," in Proceedings of the 27th International Conference on Neural Information Processing Systems (NeraIPS), 2672–2680. doi:10.1145/3422622

Han, X.-H., Zheng, Y., and Chen, Y.-W. (2019). "Multi-level and multi-scale spatial and spectral fusion cnn for hyperspectral image super-resolution," in IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (Seoul, South Korea: IEEE), 4330–4339. doi:10.1109/ICCVW.2019.00533

Haut, J. M., Fernandez-Beltran, R., Paoletti, M. E., Plaza, J., Plaza, A., and Pla, F. (2018). A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Trans. Geosci. Remote Sens.* 56, 6792–6810. doi:10.1109/TGRS.2018.2843525

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA: IEEE), 770–778. doi:10.1109/CVPR.2016.90

Jiang, J., Sun, H., Liu, X., and Ma, J. (2020). Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Trans. Comput. Imaging* 6, 1082–1096. doi:10.1109/TCI.2020.2996075

Johnson, J., Alahi, A., and Li, F.-F. (2016). "Perceptual losses for real-time style transfer and super-resolution," in IEEE European Conference on Computer Vision (IEEE), 694–711.

Jun, Z., Jiao, L., Bin, P., and Zhenwei, S. (2020). Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* 58, 7920–7930. doi:10.1109/TGRS.2020.2985072

Kotovenko, D., Sanakoyeu, A., Ma, P., Lang, S., and Ommer, B. (2019). "A content transformation block for image style transfer," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA: IEEE), 10024–10033. doi:10.1109/CVPR.2019.01027

Lee, S., Kim, D., Kim, N., and Jeong, S.-G. (2019). "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in IEEE/CVF International Conference on Computer Vision (ICCV) (Seoul, South Korea: IEEE), 91–100. doi:10.1109/ICCV.2019.00018

Lei, S., and Shi, Z. (2022). Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* 60, 1–10. doi:10.1109/TGRS.2021.3069889

Lei, S., Shi, Z., Wu, X., Pan, B., Xu, X., and Hao, H. (2019). "Simultaneous super-resolution and segmentation for remote sensing images," in IEEE International Geoscience and Remote Sensing Symposium(IGARSS) (Yokohama, Japan: IEEE), 3121–3124. doi:10.1109/IGARSS.2019.8900402

Lei, S., Shi, Z., and Zou, Z. (2020). Coupled adversarial training for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* 58, 3633–3643. doi:10.1109/TGRS.2019.2959020

Li, C., and Wand, M. (2016). "Precomputed real-time texture synthesis with markovian generative adversarial networks," in European Conference on Computer Vision (ECCV), 702–716.

Li, J., Cui, R., Li, B., Song, R., Li, Y., Dai, Y., et al. (2020). Hyperspectral image super-resolution by band attention through adversarial learning. *IEEE Trans. Geosci. Remote Sens.* 58, 4304–4318. doi:10.1109/TGRS.2019.2962713

Li, Y., Yuan, L., and Vasconcelos, N. (2019). "Bidirectional learning for domain adaptation of semantic segmentation," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA: IEEE), 6929–6938. doi:10.1109/CVPR.2019.00710

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI, USA: IEEE), 936–944. doi:10.1109/CVPR.2017.106

Liu, D., Li, J., and Yuan, Q. (2021). A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* 59, 7711–7725. doi:10.1109/TGRS.2021.3049875

Liu, G., Gousseau, Y., and Tupin, F. (2019). A contrario comparison of local descriptors for change detection in very high spatial resolution satellite images of urban areas. *IEEE Trans. Geosci. Remote Sens.* 57, 3904–3918. doi:10.1109/TGRS.2018.2888985

Liu, W., and Qin, R. (2020). A multikernel domain adaptation method for unsupervised transfer learning on cross-source and cross-region remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.* 58, 4279–4289. doi:10.1109/TGRS.2019.2962039

Liu, W., and Su, F. (2020). Unsupervised adversarial domain adaptation network for semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* 17, 1978–1982. doi:10.1109/LGRS.2019.2956490

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, MA, USA: IEEE), 3431–3440. doi:10.1109/CVPR.2015.7298965

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. (2017). "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in IEEE International Geoscience and Remote Sensing Symposium(IGARSS) (Fort Worth, TX, USA: IEEE), 3121–3124. doi:10.1109/IGARSS.2017.8127684

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., and Shi, H. (2020). "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Seattle, WA, USA: IEEE), 5689–5698. doi:10.1109/CVPR42600.2020.00573

Mou, L., Hua, Y., and Zhu, X. (2020). Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* 58, 7557–7569. doi:10.1109/TGRS.2020.2979552

Pan, B., Shi, Z., Xu, X., Shi, T., Zhang, N., and Zhu, X. (2019a). Coinnet: Copy initialization network for multispectral imagery semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* 16, 816–820. doi:10.1109/LGRS.2018.2880756

Pan, B., Xu, X., Shi, Z., Zhang, N., Luo, H., and Lan, X. (2020a). Dssnet: A simple dilated semantic segmentation network for hyperspectral imagery classification. *IEEE Geosci. Remote Sens. Lett.* 17, 1968–1972. doi:10.1109/LGRS.2019.2960528

Pan, F., Shin, I., Rameau, F., Lee, S., and Kweon, I. S. (2020b). "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Seattle, WA, USA: IEEE), 3763–3772. doi:10.1109/CVPR42600.2020.00382

Pan, Z., Ma, W., Guo, J., and Lei, B. (2019b). Super-resolution of single remote sensing image based on residual dense backprojection networks. *IEEE Trans. Geosci. Remote Sens.* 57, 7918–7933. doi:10.1109/TGRS.2019.2917427

Rad, M. S., Bozorgtabar, B., Marti, U.-V., Basler, M., Ekenel, H. K., and Thiran, J.-P. (2019). "Srobb: Targeted perceptual loss for single image super-resolution," in IEEE International Conference on Computer Vision (ICCV) (Seoul, South Korea: IEEE), 2710–2719. doi:10.1109/ICCV.2019.00280

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI) (Fort Worth, TX, USA: IEEE), 234–241. doi:10.1109/IGARSS.2017.8127684

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA: IEEE), 1874–1883. doi:10.1109/CVPR.2016.207

Tasar, O., Happy, S. L., Tarabalka, Y., and Alliez, P. (2020). Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.* 58, 7178–7193. doi:10.1109/TGRS.2020.2980417

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. (2018). "Learning to adapt structured output space for semantic segmentation," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, UT, USA: IEEE), 7472–7481. doi:10.1109/CVPR.2018.00780

Tsai, Y.-H., Sohn, K., Schulter, S., and Chandraker, M. (2019). "Domain adaptation for structured output via discriminative patch representations," in IEEE/CVF International Conference on Computer Vision (ICCV) (Seoul, South Korea: IEEE), 1456–1465. doi:10.1109/ICCV.2019.00154

Volodymyr, M. (2013). *Machine learning for aerial image labeling*. Toronto, Canada: University of Toronto.

Vu, T.-H., Jain, H., Bucher, M., Cord, M., and Pérez, P. (2019). "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA: IEEE), 2512–2521. doi:10.1109/CVPR.2019.00262

Wang, L., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., et al. (2019a). "Learning parallax attention for stereo image super-resolution," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA: IEEE), 12242–12251. doi:10.1109/CVPR.2019.01253

Wang, Q., Gao, J., and Li, X. (2019b). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Image Process.* 28, 4376–4386. doi:10.1109/TIP.2019.2910667

Wang, X., Yu, K., Dong, C., and Loy, C. C. (2018). "Recovering realistic texture in image super-resolution by deep spatial feature transform," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, UT, USA: IEEE), 606–615. doi:10.1109/CVPR.2018.00070

Wei, W., Nie, J., Li, Y., Zhang, L., and Zhang, Y. (2020). Deep recursive network for hyperspectral image super-resolution. *IEEE Trans. Comput. Imaging* 6, 1233–1244. doi:10.1109/TCI.2020.3014451

Wu, Z., Wang, X., Gonzalez, J., Goldstein, T., and Davis, L. (2019). "Ace: Adapting to changing environments for semantic segmentation," in IEEE International Conference on Computer Vision (ICCV) (Seoul, South Korea: IEEE), 2121–2130. doi:10.1109/ICCV.2019.00221

Yan, L., Fan, B., Liu, H., Huo, C., Xiang, S., and Pan, C. (2020). Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.* 58, 3558–3573. doi:10.1109/TGRS.2019.2958123

Zhang, Y., David, P., Foroosh, H., and Gong, B. (2019). A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 1823–1841. doi:10.1109/TPAMI.2019.2903401

Zhang, Y., Qiu, Z., Yao, T., Liu, D., and Mei, T. (2018). "Fully convolutional adaptation networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, UT, USA: IEEE), 6810–6818. doi:10.1109/CVPR.2018.00712

Zhaoxiang, Z., Kento, D., Akira, I., and Guodong, X. (2021). Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training. *IEEE Geosci. Remote Sens. Lett.* 18, 746–750. doi:10.1109/LGRS.2020.2982783

Zheng, C., Zhang, Y., and Wang, L. (2017). Semantic segmentation of remote sensing imagery using an object-based markov random field model with auxiliary label fields. *IEEE Trans. Geosci. Remote Sens.* 55, 3015–3028. doi:10.1109/TGRS.2017.2658731

Zheng, X., Huan, L., Xia, G.-S., and Gong, J. (2020). Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss. *ISPRS J. Photogrammetry Remote Sens.* 170, 15–28. doi:10.1016/j.isprsjprs.2020.09.019

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in IEEE International Conference on Computer Vision (ICCV) (Venice, Italy: IEEE), 2242–2251. doi:10.1109/ICCV.2017.244

Zou, Z., Shi, T., Li, W., and Zhang, Z. (2020). Do game data generalize well for remote sensing image segmentation? *Remote Sens.* 12, 275. doi:10.3390/rs12020275