A Stepwise Domain Adaptive Segmentation Network With Covariate Shift Alleviation for Remote Sensing Imagery

Jiaojiao Li, Member, IEEE, Shunyao Zi, Rui Song, Member, IEEE, Yunsong Li, Member, IEEE, Yinlin Hu, Member, IEEE, and Qian Du, Fellow, IEEE

Abstract—Semantic segmentation for remote sensing images (RSIs) is critical for the Earth monitoring system. However, the covariate shift between RSI datasets under different capture conditions cannot be alleviated by directly using the unsupervised domain adaptation (UDA) method, which negatively affects the segmentation accuracy in RSI. We propose a stepwise domain adaptive segmentation network with covariate shift alleviation (Cov-DA) for RSI parsing to solve this issue. Specifically, to alleviate domain shift generated by different sensors, both the source and target domains are projected into a colorspace with normalized distribution through an elaborate colormapspace mapping unified module (CMUM). The color distributions of these two domains tend to be more uniform. Furthermore, in the target domain, the multistatistics joint evaluation module (MJEM) is proposed to capture different statistical characteristics of subscenarios for selecting plain scenarios regarded as high-confidence segmentation results to assist the further improvement of segmentation performance. In addition, a pyramid perceptual attention module (PPAM) containing omnidirectional features without computational burdens is added to our network for effectively enhancing the multiscale feature capture ability. In the cross-city DA experiments based on the International Society for Photogrammetry and Remote Sensing (ISPRS) and aerial benchmarks, the superiority of our algorithm is significantly demonstrated. Furthermore, we release a large-scale Martian terrain dataset noted as “Mars-Seg” containing 5 K images with pixel-level accurate annotations regarding issues, such as the lack of semantic segmentation datasets for unknown scenes.

Index Terms—Covariate shift alleviation, semantic segmentation, stepwise, unsupervised domain adaptation (UDA).

I. INTRODUCTION

WITH the rapid development of imaging technology, remote sensing images (RSIs) with finer texture features are utilized in scene parsing frequently. As a key technology, semantic segmentation for RSI processing is exploited in many practical applications, such as military reconnaissance, agricultural planning, and deep space exploration [1]–[5]. Specifically, semantic segmentation requires assigning each pixel in RSI to a unique semantic label. To achieve high segmentation performance, a mass of labeled training data is necessary and needed for superior results while adopting supervised learning. Nevertheless, in reality, label annotation for segmentation datasets is an extremely expensive labor, which leads to the lack of labeled data in several typical scenes. In addition, the prior shift [6] inflicted by significant differences in content distribution between source and target domain images prevents a well-trained model with limited generalization capabilities from obtaining ideal segmentation results. Therefore, given a dataset with limited labeling, achieving high-precision segmentation is still a major challenge in the current research field.

Unsupervised domain adaptation (UDA) as an efficient technology can assist the model in learning the domain-invariant feature representation between the source and target domains by introducing the idea of the generative adversarial network (GAN) [7], thereby effectively reducing the prior shift in domain shift. At present, the UDA-based semantic segmentation methods have achieved excellent results in the raw image, but which cannot be directly applied to RSI [6]. The reason is that the RSI capture conditions (e.g., shooting angles, camera parameters, illumination conditions, and atmospheric radiation) vary widely, which changes the confounding factor describing the distribution difference between the source and target domain datasets, resulting in covariate shift [8]. In addition, the range of captured RSI is relative to the large-scale scene, so that some subscenarios contain a large number of irregular objects. The spatial diversity and background complexity created by these irregular objects lead
to poor predictive results through using UDA-based methods, which requires the UDA approach to focus on these informative subscenarios and improve their accuracy during the DA process [9]. Besides, the existing models improve the feature extraction capability via building multiple modules or branches while ignoring the additional computational burden. What is more, it is critical to comprehend some typical unknown scenarios for researchers like Mars terrain exploration missions. However, such datasets containing finely labeled information are still very scarce but urgently needed.

In this article, we propose a stepwise domain adaptive segmentation network with covariate shift alleviation (Cov-DA) for RSI to address the aforementioned drawbacks. The principle comparison between the conventional UDA method and Cov-DA is shown in Fig. 1. An image preprocessing structure noted as the colorspace mapping unified module (CMUM) is presented to the Cov-DA front end, which projects raw image data from the source and target domains into a unified color distribution mapping space to alleviate covariate shift created by different capture conditions. Furthermore, given the dissatisfactory segmentation effect of complex subscenarios in RSI, we first evaluate the complexity of the subscenarios and then redivide the target domain into two novel domains consisting of a plain domain and a complex domain according to the content complexity, to be specific, the plain domain is considered to be an additional useful region with high confidence results to assist extracting abundant features of the complex domain in the target domain. Consequently, we perform the stepwise UDA training to help the network focus on informative subscenarios. Considering that a single-statistical feature cannot be regarded as a criterion for an image’s information richness and content complexity, a multistatistics joint evaluation module (MJEM) is proposed to select a plain domain with high confidence segmentation results. We also designed a pyramid perceptual attention module (PPAM) to facilitate feature extractor extracting multiscale features without computational burdens. Most importantly, we released a Martian terrain dataset named “Mars-Seg” that includes 1024 grayscale images with $1024 \times 1024$ spatial resolution and 4155 RGB images with $560 \times 500$ spatial resolution, covering Mars terrain scene images taken by multiple rovers with detailed annotation. In general, the main contributions of our work can be summarized as follows.

1) We first propose a novel algorithm named Cov-DA to realize UDA semantic segmentation based on the data characteristics for RSI parsing.

2) An effective unsupervised CMUM is presented to unify color space mapping between source and target domains, which significantly alleviates the covariate shift caused by different capture conditions.

3) To further enhance the segmentation performance, the MJEM is proposed to assist the network focus on more complex subscenarios in the target domain through evaluating the complexity of RSI. Besides, the PPAM is proposed to reduce the computational cost of the proposed network efficiently.

4) More importantly, we publish a Mars terrain dataset with scientific annotations, including nine classes for semantic segmentation research, and verified the superiority of Cov-DA on this dataset. Besides, the experimental results on the other two public RSI benchmarks named International Society for Photogrammetry and Remote Sensing (ISPRS) and Aerial also demonstrate the innovation of our proposed algorithm in the RSI research field.

II. RELATED WORK

A. Domain Adaptation

Domain adaptation is a special branch of transfer learning, which can effectively alleviate the problem of domain shift. To be specific, when there is a significant difference in the feature distribution between the source domain and the target domain, a well-trained model cannot achieve ideal prediction results in the target domain due to its limited generalization ability [10]. The DA methods can help the network continuously adapt and learn the feature distribution of the target domain during the training process and align the feature distribution of the two domains as much as possible. At present, DA has been widely used in the field of computer vision. Some typical DA-based methods are as follows: deep adaptation network (DAN) [11] is one of the earliest methods of depth DA based on maximum mean discrepancy (MMD). In this method, the MMD between the source domain features and the target domain features is added to the objective function and optimized to realize the domain adaptation from the source domain to the target domain. Then, Gannin and Lempitsky [12] proposed a domain adaptive adversarial neural network impacting the research field of DA greatly, which is the first model that applies the idea of single adversarial learning to DA. Soon after, Pei et al. [13] proposed multi-adversarial DA (MADA), which utilizes multiple category domain discriminators to align the data distribution of two domains to improve domain adaptability. Each category domain discriminator is only responsible for aligning the probability distribution of its corresponding category and enhances the domain adaptability by increasing the discriminator’s attention on each category. In addition, Sankaranarayanan [7] proposed a method called generate to
adapt (GTA), which not only uses GAN to synthesize images but also exploits the training process of GAN to align the story distribution of the source domain and target domain. The advantage is that the domain invariant feature representation can be obtained to realize DA even if the synthetic data fails. However, the above-mentioned methods effectively solve the domain shift problem, which does not consider the refined domain shift’s composition and solves them one by one.

B. UDA for Semantic Segmentation

To solve the problems of domain shift and lack of dataset labels in practical applications of semantic segmentation, some investigators have begun to introduce UDA methods into the field of semantic segmentation. UDA methods can be roughly divided into adversarial learning-based, self-learning-based, and image-to-image translation-based [14]. For the adversarial learning-based semantic segmentation methods, Hoffman et al. [15] introduced the idea of adversarial learning into the field of semantic segmentation. They chose fully convolutional networks for semantic segmentation (FCN) [16] as the feature extractor to extract the standard features of the source and target domains through sharing network parameters. And the semantic consistency loss function [17] is introduced on the discriminator network side to continuously align global and local features. However, the UDA method based on semantic segmentation cannot stably converge during the training process. Sankaranarayanan et al. [18] introduced a classified loss function into adversarial domain adaptive network to assist the training of adversarial network generation and ensure the stability of network convergence. Inspired by MADA [13], Tsai et al. [19] proposed a multilevel adversarial learning network consisting of multiple discriminators. Among them, multiple discriminators can align the features of different layers through self-learning to ensure the generalization ability of the network. In addition, several scholars try to alleviate the domain shift by introducing the idea of image-to-image translation [14], [20]. Zhu et al. [20] proposed an end-to-end image style conversion network based on CycleGAN by mapping the source domain into a form similar to the image style of the target domain. This method aligns features on each output channel to improve the effectiveness of image style conversion [21].

C. Semantic Segmentation in RSI

The semantic segmentation of RSI is of great significance in geographical investigation and land planning. Although the method based on deep learning dramatically improves the segmentation performance, due to the diversity of RSI and the lack of pixel-level labels, scholars need to customize the segmentation algorithm according to the characteristics of RSI. Afterward, a series of semantic segmentation networks for RSI came out at the historical moment and had a profound impact on remote sensing. First of all, multibranch convolutional neural networks are proposed to extract and fuse image multistage features through cascading modules to improve the prediction accuracy of RSI [22]–[24]. Xu et al. proposed a dual-branch convolutional neural network, which employs cascaded modules to extract and fuse multisource image features to improve the accuracy of segmentation in RSI [5]. Inspired by the effectiveness of graph convolution, Yang et al. [25] apply deep convolutional neural network (DCNN) and dynamic graph convolution (DGCN) for joint classification and improve the segmentation accuracy by mining the global information and local semantics of the image. Subsequently, due to that supervised learning is too dependent on large amounts of training data, scholars gradually pay attention to the small sample learning and weakly supervised learning [26]–[30]. Meanwhile, the effectiveness of the UDA method has attracted a large number of researchers who are committed to achieving high-precision segmentation of RSI [31]–[35]. Mei et al. [3] proposed a small-sample remote sensing scene deep feature classification network based on sparse representation to fuse multilayer features and achieved excellent results in a small-sample dataset. Li et al. [36] proposed a new objective function consisting of multiple weakly supervised constraints to learn cross-domain invariant features in RSI. This new method achieves excellent results in the field of weakly supervised semantic segmentation. Ma et al. [4] used spatial filtering and overall centroid alignment preprocessing methods to improve the classification results of conventional DA and introduced MMD to improve further DA performance. Shi et al. [37] proposed a novel rotation consistency-preserved GAN (RCP-GAN) to map aerial images in the source domain to the target domain for domain adaptation.

More importantly, a few people take into account the covariate shift between two domains when studying UDA-related methods. Thus, while alleviating the domain shift, we also alleviate covariate shift as much as possible.

III. Method

In this section, we first introduce the structure and data processing of this network. Second, the CMUM is illustrated in detail. Then, the details of the MJEM are described. Finally, the PPAM is present.

A. Cov-DA Architecture

Fig. 2 shows the proposed network structure diagram and specific data processing flow. The CMUM in the Cov-DA takes the lead in unifying the colorspace of datasets of source and target domain employing mapping, realizing the normalization of the color distribution of two domains, and conducting regular DA training. In semantic segmentation tasks, the prediction results of simple subscenarios images in the target domain tend to have higher confidence [38]. After completing the first UDA training, the network retains the UDA model parameters and prediction results to assist the second intraclass DA. Among them, the prediction results of the target domain are divided by MJEM as pseudolabels to complete the second UDA training. MJEM can evaluate the complexity of subscenarios images through multiple statistical characteristics of the image (entropy of the image and GLCM at the edge of the image) and redivide the target domain into new source domains (plain) and target domains (complex). Then, the network retains the first DA segmentation results as
The nonlinear colorspace mapping matrix is shown in Fig. 3. The module calculates a unique transformation matrix based on [40], which normalizes the color distribution of white balance, we designed a colorspace mapping module to correct the input image. Since the color correction matrix is related to the color distribution of the two domains and alleviates the covariate shift caused by capture conditions. The specific processing flow of the module is shown in Fig. 3. The module calculates a unique transformation matrix \( M \) by extracting the nearest neighbors that are most similar to the input image in terms of color. Based on the retrieved similar images, the color transformation \( M \) is calculated to correct the input image.

### B. Colorspace Mapping Unified Module

White balance is an essential concept in photography, which often solves a series of problems such as color reproduction and tone processing, helping the camera present a more realistic image color distribution. Inspired by the idea of white balance, we designed a colorspace mapping module based on [40], which normalizes the color distribution of the two domains and alleviates the covariate shift caused by capture conditions. The specific processing flow of the module is shown in Fig. 3. The module calculates a unique nonlinear colorspace mapping matrix \( M \) by extracting the color distribution characteristics of the input image \( I_{in} \), which can unify the colorspace mapping of the image. Since the color correction matrix is related to the color distribution of the input image \( I_{in} \), it is first necessary to construct a unique multidimensional spatial color histogram for the input image, which is called RGB-uv histogram. The construction formula of the histogram is as follows:

\[
I_{y(i)} = \sqrt{I_{R(i)}^2 + I_{G(i)}^2 + I_{B(i)}^2}
\]

\[
I_{a1(i)} = \log(I_{R(i)}) - \log(I_{B(i)})
\]

\[
I_{a2} = -I_{a1}, I_{a2} = -I_{a1} + I_{a1}
\]

\[
I_{a3} = -I_{a1}, I_{a3} = -I_{a1} + I_{a1}
\]

\[
H(I)_{(u,v,C)} = \sum_i I_y(i)^2 \left| I_{aC(i)} - \mu \right| \leq \frac{\varepsilon}{2} \wedge \left| I_{aC(i)} - \mu \right| \leq \frac{\varepsilon}{2}
\]

\[
h(I)_{(u,v,C)} = \frac{H(I)_{(u,v,C)}}{\sum_{u'} \sum_{v'} H(I)_{(u',v',C)}}
\]

\( R, G, \) and \( B \) represent the color channels in \( I \), \( C \in \{1, 2, 3\} \) each color channel in the histogram, and \( \varepsilon \) is the width of histogram bin. In order to highlight the characteristics of the projection histogram, the normalized \( H \) has also undergone a square root operation.
In order to reduce the search time of the database, the CMUM calculates compact principal component analysis (PCA) features and searches the database for the \( k \)-nearest neighbor transformation matrix that is the most similar to the input image in terms of color distribution. The weighted linear combination of the correction matrix \( M \) is calculated using the color correction matrix \( M_i \), with similar PCA characteristics of the \( k \)-pair and the weight factor \( \alpha_j \), as shown in the following:

\[
M = \sum_{j=1}^{k} \alpha_j M_i^{(j)}
\]

\( \alpha \) is a weighted vector expressed as a radial basis, and the function is

\[
\alpha_j = \frac{\exp\left(-d_j^2/2\sigma^2\right)}{\sum_{k=1}^{k} \exp\left(-d_k^2/2\sigma^2\right)}, \quad j \in [1, \ldots, k]
\]

where \( \sigma \) is the radial attenuation factor, and \( d \) is the vector, which contains the \( L_2 \) distance between the given input features and similar \( k \) training features \( I_{\text{corr}} \)

\[
I_{\text{corr}} = M\Phi(I_{\text{in}})
\]

\( \Phi \) is a kernel function that projects the RGB triplet to a high-dimensional space. Finally, the calculated \( M \) unified colorspace mapping the input image. Through CMUM, we get images of two domains under the same colorspace mapping.

**C. Multistatistics Joint Evaluation Module**

The image content of the subscenarios in the target domain data is quite different, which makes the prediction result of the complex subscenarios very confusing. Therefore, the UDA method must pay more attention to the complex subscenarios in the target domain [38]. As shown in Fig. 4, the MJEM evaluates the complexity of the image according to the GLCM of the image edge information and the entropy information of the image and divides the target domain data according to the complexity.

The entropy of an image is a feature representation that reflects the average amount of information about an image. Based on 1-D entropy, we introduce a feature that can reflect the spatial characteristics of the distribution to form the 2-D entropy of the image:

\[
P_{i,j} = \frac{f(i, j)}{W \cdot H}
\]

\[
\text{Com}_{\text{CE}} = -\sum_{i=0}^{255} P_{i,j} \log P_{i,j}
\]

\( W \) and \( H \) represent the width and height of the image, \( (i, j) \) represents a two-tuple, \( i \) is the gray value of the center in the sliding window, and \( j \) is the average gray value of the pixels in the window except for the center pixel. \( f(i, j) \) represents the number of times \( (i, j) \) this two-tuple appears in the image.

The edge of the image is produced by the sudden change of the attributes of the image area, which effectively reflects the uncertain factors contained in the image information [41]. The GLCM was proposed by [42] to describe the texture characteristics of the image:

\[
\text{CON} = \sum_{n=0}^{k-1} n^2 \left\{ \sum_{(i-j)=n} P_{i,j} \right\}
\]

\[
\text{ASM} = \sum_{i=1}^{k} \sum_{j=1}^{k} (P_{i,j})^2
\]

\[
\text{Com}_{\text{M}} = \text{CON} + \sqrt{\text{ASM}}
\]

\( P(i, j) \) means the same as earlier. Contrast (CON) represents the sharpness of the image and the depth of texture grooves, and arc-second moment (ASM) reflects the uniformity of the gray image distribution and the thickness of the texture. We set the calculated complex evaluation index based on the image edge as \( \text{Com}_{\text{M}} \), and the complex evaluation index based on the image entropy as \( \text{Com}_{\text{CE}} \).

The complex index of subscenarios is derived from a linear combination of two statistical features. The parameter \( \lambda \) is set to 0.37 empirically

\[
\text{CI} = \text{Com}_{\text{M}} + \lambda \text{Com}_{\text{CE}}.
\]

This module effectively divides the target domain into plain and complex scenes by scoring the complexity index of the subscenarios.

**D. Pyramid Perceptual Attention Module**

In the adversarial DA method, the feature extractor and the discriminator continuously perform feature alignment through negative training to alleviate the domain shift. The feature extractor’s function extracts the source domain and the target domain features. It makes the feature distribution of the target domain and the source domain image as similar as possible through adversarial training. Therefore, whether the feature extractor fully extracts the multiscale feature information of the two domains has a crucial influence on the final prediction accuracy. Conventional feature extractors cannot capture the multiscale feature information in the RSI dataset with complex background. We design a perceptual attention module that captures multiscale context feature information and applies it in segmented networks to achieve this goal. The structure is...
shown in the PPAM structure at the end of the feature extractor in Fig. 2. This module is based on the image pyramid method to capture multiscale spatial features, combined with the unique ability of CA [39] to establish the length dependence relationship between various channel attention mechanisms, designed an omnidirectional perceptual attention mechanism. In addition, we also compress the number of channels in the feature graph to ensure that the module does not bring too much computational load to the model while providing strong feature capability.

The conventional attention module often brings significant performance improvement to the model. In contrast, the channel attention usually loses the position information on the attention feature map of the generated space, so we introduce the CA through the precise position information. The specific structure is shown in Fig. 5. Channel relationship and long-range dependence are encoded, similar to the SE module, and divided into two steps: coordinate information embedding and CA generation. Global pooling is often applied to channel attention to globally encoded spatial information as channel descriptors in the coordinate information embedding part. The global pooling is decomposed into a 1-D feature encoding operation performed along the horizontal coordinate direction and a convolution kernel along the vertical coordinate direction to promote the attention module to capture the long-range spatial dependence on the accurate position information

\[
\begin{align*}
    z_c^h(h) &= \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i) \\
    z_c^w(w) &= \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w).
\end{align*}
\]  

(13)  

The two transformations perform feature aggregation along with two spatial directions and return a pair of direction-aware attention feature maps that preserve precise location information.

In the attention generation part, the feature map generated by the embedding part is transformed F1 using a shared 1 × 1 convolution, and the aggregation feature is obtained through conventional nonlinear normalization

\[
f = \delta(\hat{F}_1([x^h, z^w])).
\]  

(15)

We cut f into two separate tensors along the spatial dimension \(f^h\) and \(f^w\), transforming it to the same number of channels as the input using 1 × 1 convolution to obtain the result following:

\[
g_c^h = \sigma(F_h(f^h)) \quad \text{and} \quad g_c^w = \sigma(F_w(f^w)).
\]  

(16)

The final output of the CA module can be seen as the following formula:

\[
y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j).
\]  

(17)

Since multiscale segmentation objects often appear in RSI, we choose a spatial pyramid structure to extract the length-dependent information about feature images of different scales [43]. In addition, the increase in the number of feature map channels will significantly increase the amount of calculation of the model and bring a nonnegligible computational burden to the network. For this reason, we add a convolution layer after the feature graph of each scale, reduce the number of channels of all convolution kernels and introduce a short connection residual structure to fuse the input and output of PPAM.

E. Total Loss Function

The input image \(X_s\) from the source domain dataset outputs the prediction result \(P_s = \hat{F}(X_s)\) at the end of the feature extractor. Under the condition that the ground truth \(Y_t\) corresponding to the source domain input image is available, there is a supervised branch in the feature extractor of the \(i\)th stage. The cross entropy loss function of can be written as

\[
L_{\text{seg}}^i(X_s, Y_s) = -\sum_{h,w} \sum_{c} Y_s^{(h,w,c)} \log(P_s^{(h,w,c)})
\]  

(18)

where \(C\) is the number of categories in the dataset, \(h\) and \(w\) represent the height and width of the graph, respectively.

In order to alleviate the domain-shift between the source and target domains, the discriminator needs to discriminate whether the input entropy feature map \(I\) is from the source domain or the target domain. Therefore, the loss function of the domain adaptation part can be written as follows:

\[
L_{\text{adv}}^i(X_s, X_t) = \sum_{h,w} \log(1 - D_i(I_s^{(h,w)})) + \log(D_i(I_t^{(h,w)}))
\]  

(19)

where \(I_s\) and \(I_t\) are the entropy map of \(X_s\) and \(X_t\), respectively. The loss function \(L_{\text{seg}}^i\) and \(L_{\text{adv}}^i\) are optimized to align the distribution shift between the two domains.

Finally, the total loss function of this network is the sum of the two-stage loss functions

\[
L_{\text{full}} = L_{\text{seg}}^1 + L_{\text{adv}}^1 + L_{\text{seg}}^2 + L_{\text{adv}}^2
\]  

(20)
and the optimization goal of the total loss function is to obtain a feature extractor of the target domain
\[
F^* = \arg \min_{G_1, G_{1b2}, F_{1b2}} \min_{\mathcal{L}_{full}}.
\] (21)

IV. EXPERIMENT

A. Datasets

This article conducts UDA experiments on two sets of RSI datasets and a set of self-made Martian terrain datasets. Among them, the two groups of RSI datasets come from the 2-D benchmark dataset of ISPRS and the Aerial dataset, respectively [44], [45]. In addition, to better study the performance of this algorithm, we also created a geomorphic segmentation dataset named “Mars-Seg” based on the publicly available Mars32k dataset and the publicly available data on NASA’s official website [46], [47].

1) ISPRS Dataset: ISPRS provides two state-of-the-art airborne image datasets for city classification and 3-D building reconstruction test projects. The dataset employs a digital surface model (DSM) generated from high-resolution orthophotographs and corresponding dense image matching techniques. These two datasets cover urban scenes: Postdam is a typical historical city with large buildings, narrow streets, and dense settlement structures. Vaihingen is a relatively small village with many independent buildings and small multistory buildings. The RSI comprises three near-infrared bands, red, and green, and we choose it as the target domain dataset. Each dataset has been manually classified into the six most common land cover categories.

2) Aerial Dataset: The ground truth of Berlin, Chicago, Paris, Potsdam, and Zurich consists of aerial images from Google Maps and pixelwise building, road, and background labels from OpenStreetMap. The real image of the feature includes an aerial image of the Tokyo area from Google Maps and manually generated pixel-level buildings, roads, and background labels. Here, we choose the Aerial image of the Tokyo area as the target domain of the evaluation. In this dataset, the ground truth for the Potsdam area covers the same area as the ISPRS dataset. And the manual annotations involved are consistent.

3) Mars-Seg Dataset: The Mars-Seg dataset contains high-resolution images with rich Mars scenes, which helps researchers understand the real Mars landscape. All single-channel grayscale images of this data are from the planetary data system (PDS) [47], covering 1024 high-definition images taken by the navigation camera (NAVCAM) and panoramic camera (PANCAM) of opportunity and the spirit Mars rovers (MER). The RGB images are collected by Mars 32 k [46], all from the mast camera (MastCam) of the Curiosity Rover (MSL), a total of 4155 images. We named the data of the two types of the rover as MER-Seg and MSL-Seg, respectively. It should be noted that the spatial resolution of the grayscale image in MER-Seg is 1024 × 1024, while the color image from MSL-Seg is downsampled to 560 × 500 by bilinear interpolation.

The original land covers on the surface of Mars has high texture similarity, complex scenes, and a lack of prior information. In 2021, the first large-scale Martian terrain dataset named AI4Mars is released by NASA’s Mars Science Laboratory [48] containing four types of landforms. To better comprehend of Mars, only four categories cannot perceive Mars in detail. By analyzing the difficulties that the rover has suffered during the exploration process and the high-risk problems [49], [50] that may be encountered with the execution of the mission, we divide the terrain in the dataset into nine categories. This dataset not only contains the four terrain categories (martian soil, sands, bedrock, and rocks) and unknown categories that AI4Mars has already proposed but also adds gravel, tracks, shadow, and background categories to enrich the Martian scene [51]–[55]. Among them, gravel is defined as smaller than rock particles but rougher than sand. This category is often formed by weathering and denudation of rocks. The tracks category is the trace left by the Mars rover after traveling. The shadow category indicates the area projected by the rover and the rock under the light. The background category refers to the blurred sky, mountains, and so on, in the distance. We selected some samples of the dataset as a reference, as shown in Fig. 6. Among them, the color bar displays the color of each terrain category and counts the number of images in each category in the dataset.

a) Labeling process follows: Twenty-four professional labelers were selected from hundreds of volunteers and received several months of training on Martian terrain knowledge. During the labeling process, all images are randomly assigned to multiple labelers for multiple labeling, and the average boundary value is extracted as the final labeling result to alleviate subjective factors. After finishing the labeling, we hand over the data to professional geological experts for review and adjustment. In the future, we plan to add the annotation results of the Zhurong from China and Perseverance rover data from the United States to enrich the dataset.

B. Network and Training Details

The EfficientNet [56] point out that convolutional neural networks usually improve the prediction accuracy by increasing the width of the network. Res2Net expresses multiscale features in the granularity level and increases the receptive field of each layer [57]. In the original residual unit structure, a small residual block is added to increase the size of the receptive field of each layer. Therefore, we choose Deeplab-V2 architecture as the basic framework to extract features from a pretrained Res2Net backbone. In order to prevent features from disappearing in the deep network, the multilevel method is applied to obtain the feature maps of Res2Net at different depths at stages 3 and 4 and added PPAM at the end of the feature maps to improve the prediction accuracy. Finally, the up-sampling layer and softmax operator assist the network to obtain segmentation results that match the size of the input image. In addition, all the comparison methods in this manuscript adopt the Deeplab-V2 framework with Res2Net as the backbone and the multilevel feature extraction to maintain consistency with our method.

The details of our experiment settings are as follows: in the ISPRS series dataset, we choose RGB images of Postdam
Fig. 6. Sample image of the Mars-Seg dataset. The dataset contains nine geomorphic categories and has been carefully labeled. The color labels corresponding to each geomorphic category and the number of images in the dataset have been marked in the color bar.

as the source domain and Vaihingen as the target domain. In the Aerial series dataset, consistent with the author of the original paper [45], we chose the Berlin dataset as the source domain and the Tokyo dataset as the target domain. In the Mars-Seg series datasets, we chose the gray image of the MER-Seg dataset as the source domain and the color image of the MSL-Seg dataset as the target domain. Considering that RGB images can be directly converted into grayscale images, we set the MSL-Seg dataset as the target domain to highlight the advantages of the DA method.

All experiments were carried out on a single-card GPU of NVIDIA 2080Ti with the software environment of PyTorch. Due to GPU performance limitations, we cut all images to $512 \times 512$ dimensions (except the MSL-Seg dataset). In all GAN-based DA algorithms, we choose stochastic gradient descent (SGD) as the optimization method and ensure that all training parameters are consistent. The detailed training parameters are as follows: the batch size of the source domain and target domain are 2, the momentum is 0.9, the supervised learning rate is 2.5e-4, the maximum number of iterations for 150,000, the weight decay is 5e-4, and the adversarial learning rate is 1e-4. In addition, in the MJEM module proposed by this algorithm, we divide the target dataset into the new source domain and target domain in a ratio of 4:6. Mean intersection over union (mIOU), as the most common performance index in semantic segmentation tasks, effectively reflects the performance of a segmentation algorithm. Therefore, in this article, the IOU and mIOU of each category were selected as the evaluation index of the experimental results.

C. Comparing Methods

In order to highlight the superiority of our algorithm, representative DA methods in recent years are selected for comparison: CycleGAN [20], ColorMapGAN [58], AdaptSegNet [19], category-level adversaries for semantics consistent domain adaptation method (CLAN) [59] and baseline segmentation algorithm Intra-DA [38]. We also add supervised algorithm results trained only with the source domain dataset and domain dataset to show the performance gap between supervised and unsupervised algorithms.

1) CycleGAN [20] is a style transfer network with bidirectional cyclic generation structures. This network changes the impressionistic style of the image while keeping the content unchanged. We utilize the CycleGAN network...
to transfer the style of the source domain image to the target domain image. We train the source domain dataset by the supervised algorithm and evaluate the target domain dataset.

2) ColorMapGAN [58] is a DA semantic segmentation network that generates pseudotraining images through color mapping. In this network, the color distribution of the source domain image is mapped to a color similar to that of the target domain through a simple matrix operation, thus eliminating part of the domain shift.

3) AdaptSegNet [19] is a semantic segmentation network that is different from the DA method of feature extraction network structure by sharing model parameters to achieve DA training.

4) CLAN [59] takes into account the category differences in feature alignment and further alleviates domain shift among different categories by setting weights for the antagonism losses of different pixels in the feature graph.

5) Intra-DA [38] takes into account the domain shift within the target domain. To alleviate the domain shift, after the first DA training, the network redivides the target domain images into a new source domain and a target domain. It retains the test result of the first DA as a pseudo-label to complete the second DA, effectively eliminating the domain translation within the target domain. This DA network gradually adapts to the feature distribution of the target domain image. Therefore, we choose it as the baseline of our method.

### D. Results in Different Dataset

Table I shows the IOU indicators and mIOU indicators of each method on the ISPRS dataset (Potsdam to Vaihingen) and the Aerial dataset (Berlin to Tokyo). As can be seen from the result of ISPRS, the mIOU indicator of our method is 4.66 points higher than the second-best method CLAN [59]. Our method also achieves optimality for every single class except tree. Compared with source-only, the baseline without DA, our method performs very well, indicating the importance of UDA in semantic segmentation tasks. However, compared with target-only results, UDA algorithms still have great potential for improvement.

It’s encouraging that the Aerial dataset results show our algorithm improves the mIOU score by 10 points over the baseline. In addition, our method keeps first place in IOU in all three categories of indicators. The results of Aerial also strongly confirm the superiority of our proposed algorithm. Surprisingly, ColorMapGAN’s results on this dataset were not satisfactory and even scored lower than source-only. When comparing CLAN with baseline results, baseline performs even better. Although there are only 16 original images in the Tokyo dataset, our method achieves excellent results based on the baseline, which indicates that our method performed better results than other methods in the scenario where the target domain data is exceptionally scarce.

Table II provides the segmentation results of different methods in the Mars-Seg dataset. Similarly, our method still achieves the highest mIOU. Compared with the baseline [38], our method increases the mIOU indicator from 61.59 to 64.76 in this dataset. This time, our algorithm achieved the highest IOU across all categories. The original intention of our study on unsupervised segmentation is to try to obtain the optimal result from the Mars-Seg dataset, and the results in Table II confirm that specific results have been achieved in this study. However, compared with target-only results, there is still a large gap between the performance of unsupervised and supervised methods.
In addition, we selected several representative comparison methods to compare the visual results with our method, and the specific results are shown in Fig. 7. From the visualization results, we see that AdaptSegNet’s visualization results are the worst. There are many wrong category predictions and category confusion. In the first diagram of ISPRS, AdaptSegNet does not effectively predict the types of cars and valid building areas. This result also happens in the Aerial dataset. By comparing the visualization results of CLAN and Intra-DA, we see that the visualization results of the two methods are relatively similar. Still, the idea of DA within the target domain of Intra-DA makes the visualization results rarely appear category confusion (for example, in the Aerial dataset and Mars-Seg dataset in row four). In addition, the prediction accuracy of the car category (ISPRS) and rock category (Mars-Seg) is not satisfactory. Unlike other methods, although Cov-DA cannot predict the precise boundary of a car, the prediction result is still quite considerable. The Aerial
Let al.: STEPWISE DOMAIN ADAPTIVE SEGMENTATION NETWORK WITH COVARIATE SHIFT ALLEVIATION

TABLE III
EXPERIMENTAL RESULTS OF ABLATION IN THE ISPRS AND AERIAL DATASETS

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset: Potsdam to Vaihingen</th>
<th>Dataset: Berlin to Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-DA (baseline) 1*</td>
<td>Imp.Sur. 71.30</td>
<td>Building 26.91</td>
</tr>
<tr>
<td>Intra-DA + PPAM 1*</td>
<td>66.85 73.68</td>
<td>27.56 52.55</td>
</tr>
<tr>
<td>Our Method (CMUM + PPAM) 1*</td>
<td>62.94 76.24</td>
<td>32.12 52.05</td>
</tr>
<tr>
<td>Intra-DA (baseline) 2*</td>
<td>59.57 70.03</td>
<td>32.80 53.79</td>
</tr>
<tr>
<td>Intra-DA + PPAM 2*</td>
<td>60.29 71.19</td>
<td>32.26 56.63</td>
</tr>
<tr>
<td>Intra-DA + MJEM + PPAM 2*</td>
<td>61.20 71.80</td>
<td>32.28 57.66</td>
</tr>
<tr>
<td>Our Method (CMUM + PPAM + MJEM)2*</td>
<td>63.24</td>
<td>78.28</td>
</tr>
</tbody>
</table>

TABLE IV
EXPERIMENTAL RESULTS OF ABLATION IN THE MARS-SEG DATASETS

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset: MIR-Seg to MSL-Seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-DA (baseline) 1*</td>
<td>Martian 64.57</td>
</tr>
<tr>
<td>Intra-DA + PPAM 1*</td>
<td>70.01 72.41</td>
</tr>
<tr>
<td>Our Method (CMUM + PPAM) 1*</td>
<td>70.18 74.24</td>
</tr>
<tr>
<td>Intra-DA (baseline) 2*</td>
<td>69.97 73.23</td>
</tr>
<tr>
<td>Intra-DA + PPAM 2*</td>
<td>70.58 72.39</td>
</tr>
<tr>
<td>Intra-DA + MJEM + PPAM 2*</td>
<td>71.23 73.85</td>
</tr>
<tr>
<td>Our Method (CMUM + PPAM + MJEM)2*</td>
<td>71.55</td>
</tr>
</tbody>
</table>

dataset even accurately predicts the background gaps between building with a few category confusions. It shows that Cov-DA has better multiscale feature extraction ability than other methods and a more robust domain adaptive ability.

It can be seen from the method results of these three datasets that the performance of the ColorMapGAN [58] method is superior to CycleGAN [20] method in most of the scenarios, which indicates that image color transformation of datasets effectively alleviates domain shift. The performance of the AdaptSegNet [19] network is very stable, which suggests that the UDA method is very robust to realize the idea of model parameter sharing in feature extractor. In general, DA methods based on shared network parameters are better than those based on image space transformation (e.g., CycleGAN), and the experimental results confirm this. However, CLAN [59] and Intra-DA have similar performance in each dataset, while Intra-DA is not stable due to the existence of quadratic DA. The possible reason is that the classification ability of the classifier is insufficient in dividing the target domain, which leads to the significant fluctuation of the results of the quadratic DA training. Our method performed very well in every dataset, achieving the best indicators in most categories, especially in small-scale ones with even more dramatic improvements (e.g., cars). Our method performs better in scenarios where the target domain data is scarce (such as Aerial experiments).

E. Ablation Study

To investigate the validity of the three modules proposed in this article, we performed different ablation experiments on three datasets: experiments on PPAM, MJEM, and CMUM. Table III shows the ablation experimental results of each module on ISPRS and Aerial datasets. Table IV shows the ablation experimental results of each module on Mars-Seg datasets. As the baseline of this algorithm, Intra-DA is a new unsupervised adaptive semantic segmentation method. There are two DA stages in this network. Therefore, we conduct ablation experiments in both phases of the baseline to ensure the effectiveness of each module proposed in this algorithm. Among them, we added “1*” and “2*” in Table IV as annotations to distinguish the experimental results of the two stages.

By comparing the results of Intra-DA and “Intra-DA + PPAM” in the ISPRS dataset, we see that the network with PPAM module improved by 2.53 points and 0.95 points in two stages, respectively. In addition, the prediction accuracy of almost every category has been improved, indicating that the network using the PPAM module has a more vital extraction ability in RSIs. The improved accuracy of the car shows that the PPAM module supports the network capture more features of small objects. In the Aerial dataset, the network using PPAM improved by 4.09 points and 3.16 points in the two phases of the baseline, respectively. This is because Tokyo contains many building objects with modified scales, and the addition of PPAM makes the network more potent in extracting multiscale features and associating context information. Similarly, in the Mars-Seg dataset, the improvement of mIOU accuracy also confirms the effectiveness of PPAM. The addition of PPAM enables the network to capture multiscale feature extraction and strengthen the connection of context information, which is called feature perception ability.

MJEM is exploited to divide the target domain dataset into simple and complex scenarios in the second stage of DA training. Therefore, we compared the index results of “Intra-DA + PPAM 2*” and “Intra-DA + MJEM + PPAM” in each dataset. In the ISPRS dataset, we observe that networks using MJEM achieve higher mIOU, although only 0.60 points improvement. However, in the Aerial dataset, the effect of MJEM is pronounced, which is improved from 42.36 to 44.41. Among them, the mIOU accuracy of the building category is improved by 4.20 points based on the original network. In addition, the results in the Mars-Seg dataset show that the network of “Intra-DA + MJEM + PPAM” is improved by...
different magnitude compared to “Intra-DA + PPAM” in up to eight geomorphic categories. The above-mentioned multigroup comparison experiments confirm that the proposed MJEM has stronger robustness in the complexity division of the target domain dataset, which is realized using the joint evaluation of image edge information and entropy information.

Finally, the efficacy of CMUM was evaluated by ablation experiments. First of all, in the comparative experiment of “Intra-DA + PPAM 1*” and “Our Method 1*” in Table III, we see that the network using CMUM in the first stage has a noticeable improvement in the final results. Compared with the results without color correction, the results in the ISPRS dataset improved by 1.38 points and the Aerial dataset improved by 4.55 points. In reaching the effects of “Intra-DA + MJEM + PPAM 2*” and “Our Method 2*” in the second stage, Our Method achieved the optimal results in both datasets and improved by 3.49 points and 3.84 points, respectively, in the two groups of data. In addition, ablation
results in the Mars-Seg dataset show that CMUM improved accuracy by 2.05 points and 2.53 points in two-stage.

In order to better study the action principle of CMUM, we selected several representative DA methods for visualization analysis on ISPRS, Aerial, and Mars-Seg, respectively. The specific visualization results are shown in Fig. 8. It can be seen that in the ISPRS dataset, the CycleGAN method transforms the green vegetation in the source domain into the red vegetation with the same color as the target domain. However, after completing the transformation of vegetation categories, this DA method makes some ground and building categories also turn red. Even regular ground is transformed into pseudovegetation, thus losing too much of its original character. Although the DA method of ColorMapGAN does not lose image information, the color mapping between building and vegetation is confused due to some wrong color changes, which reduces the model accuracy. In the Aerial dataset, the DA methods of CycleGAN and ColorMapGAN make the features of vegetation categories extremely vague and chaotic, thus interfering with the judgment ability of the segmentation network. In addition, CycleGAN even generated Mosaic results in some transformations in the Mars-Seg dataset because the source domain images were grayscale. However, the ColorMapGAN network incorrectly maps the original sand to the color distribution of Martian Soil.

In terms of the results of CMUM, although CMUM could not directly transform the color of the vegetation in the ISPRS dataset, the color contrast of source domain images through unified colorspace mapping was enhanced in each category, and the identification of each class was increased. This situation can be better reflected in the comparison of vegetation processing results in the Aerial dataset. Compared with the DA transformation from the source domain to the target domain, CMUM performs color correction on the target domain so that the source domain and target domain are closer in color distribution without changing the original feature details of the image. Although CMUM cannot achieve effective transformation in the grayscale of the MER-Seg dataset, the color correction of the target domain makes the geomorphic features of the original MSL image more distinct. Domain shift caused by different cameras is somewhat relieved.

According to the comparison results of ablation experiments, we learn that the improved segmentation network has a more vital multiscale feature extraction ability. The accuracy of vegetation and building categories in RSIs is significantly improved and the effectiveness of multiple geomorphic categories in Mars images. The validity verification experiment of the MJEM module shows that the MJEM can make the classifier pay more attention to the images with high complexity in the target domain and improve the accuracy of target domain division. In the final experimental stage, we verified the effectiveness of CMUM. CMUM makes color corrections for source domain and target domain images and modifies the datasets of two different sources into approximately the same sensor image to effectively alleviate the domain shift caused by different capture conditions.

In summary, the ablation results on three different datasets above demonstrate the effectiveness of our proposed three modules at baseline and their robustness for other datasets.

V. CONCLUSION

In this article, we propose a stepwise domain adaptive segmentation network with covariate shift alleviation for RSI. The CMUM proposed effectively alleviates covariate shift under different capture conditions by unifying the color mapping space. And the MJE module exploits multistatistics information of the image to evaluate the image complexity index of the target domain and promote the model focus on the complex subscenarios in RSI through stepwise UDA. More importantly, we released Mars-Seg, a Mars terrain dataset for the study of semantic segmentation models. We hope that this dataset will work in autonomous exploration missions outside the terrestrial area and look forward to the research results on this dataset.

ACKNOWLEDGMENT

The authors would like to thank Bobo Xi, Chaoxiong Wu, Huanqing Zhang, Yinle Ma, Songcheng Du, Xue Sun, Siyao Fan, Chenxi Ji, and Qiaoyang Ren et al. for support of Mars-Seg annotations.

REFERENCES

Jiaojiao Li (Member, IEEE) received the B.E. degree in computer science and technology, the M.S. degree in software engineering, and the Ph.D. degree in communication and information systems from Xidian University, Xi’an, China, in 2009, 2012, and 2016, respectively.

She was an exchange Ph.D. Student of Mississippi State University, Starkville, MS, USA, under the supervision of Dr. Qian Du. She is currently an Associate Professor and a master’s Supervisor with the School of Telecommunication, Xidian University. Her research interests include hyperspectral remote sensing image analysis and processing, pattern recognition.

Shunyao Zi received the B.E. degree in telecommunications engineering from Xidian University, Xi’an, China, in 2020, where he is currently pursuing the master’s degree with the State Key Laboratory of Integrated Service Network.

His research interests include semantic segmentation and deep learning.

Rui Song (Member, IEEE) received the Ph.D. degree in signal and information processing from Xidian University, Xi’an, China, in 2009.

He is currently an Associate Professor and a Ph.D. Advisor with the State Key Laboratory of Integrated Service Network, School of Telecommunications, Xidian University. His research interests include image and video coding algorithms and VLSI architecture designs, intelligent image processing, and understanding and reconstruction of 3-D scenes.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunications engineering and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi’an, China, in 1999 and 2002, respectively.

He joined the School of Telecommunications Engineering, Xidian University, in 1999, where he is currently a Professor and also the Director of the State Key Laboratory of Integrated Service Networks, Image Coding and Processing Center. His research interests include image and video processing, hyperspectral image processing, and high-performance computing.

Yinlin Hu (Member, IEEE) received the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi’an, China, in 2011 and 2017, respectively.

He was a Senior Algorithm Engineer and the Technical Leader with Zienon, LLC, Chicago, IL, USA. He is currently a Post-Doctoral Researcher with the CVLab, École Polytechnique Fédérale de Lausanne, Switzerland. His research interests include optical flow, 6-D pose estimation, and geometry-based learning methods.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore County, Baltimore, MD, USA, in 2000.

She is currently a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a fellow of the SPIE—International Society for Optics and Photonics. She was a recipient of the 2010 Best Reviewer Award at the IEEE Geoscience and Remote Sensing Society (GRSS). She was the Co-Chair of the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013. She was the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She was the General Chair of the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held in Shanghai, China, in 2012. She served as an Associate Editor for the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), the Journal of Applied Remote Sensing, and the IEEE Signal Processing Letters. Since 2016, she has been the Editor-in-Chief of the IEEE JSTARS.