High-Resolution Semantically Consistent Image-to-Image Translation

Mikhail Sokolov, Christopher Henry, Joni Storie, Christopher Storie, Member, IEEE, Victor Alhassan, and Mathieu Turgeon-Pelchat

Abstract—Deep learning has become one of remote sensing scientists’ most efficient computer vision tools in recent years. However, the lack of training labels for the remote sensing datasets means that scientists need to solve the domain adaptation (DA) problem to narrow the discrepancy between satellite image datasets. As a result, image segmentation models that are then trained, could better generalize and use an existing set of labels instead of requiring new ones. This work proposes an unsupervised DA model that preserves semantic consistency and per-pixel quality for the images during the style-transferring phase. This article’s major contribution is proposing the improved architecture of the SemI2I model, which significantly boosts the proposed model’s performance and makes it competitive with the state-of-the-art CyCADA model. A second contribution is testing the CyCADA model on the remote sensing multiband datasets, such as WorldView-2 and SPOT-6. The proposed model preserves semantic consistency and per-pixel quality for the images during the style-transferring phase. Thus, the semantic segmentation model, trained on the adapted images, shows substantial performance gain compared to the SemI2I model and reaches similar results as the state-of-the-art CyCADA model. The future development of the proposed method could include ecological domain transfer, a priori evaluation of dataset quality in terms of data distribution, or exploration of the inner architecture of the DA model.

Index Terms—Deep learning (DL), remote sensing (RS), spot-6, unsupervised domain adaptation (DA), WorldView-2.

I. INTRODUCTION

L et al. [1] states that in an era of big Earth data, also called remote sensing (RS) big data, there are significant challenges associated with a high dependency on large-scale supervised land cover labels needed to generate effective map products. The use of unsupervised domain adaptation (DA) methods to align the representation of RS images taken from different satellites and geographical regions offers a solution to this challenge. Specifically, domain adaptation methods enable the transfer of knowledge represented by existing map products in one domain to be used in domains where no such knowledge exists. Examples include situations where there are no land cover labels for a new area or they are unavailable, when the sensor characteristics in the two domains are different, or when both of these situations occur at once. Moreover, these adaptation methods can be successfully incorporated into existing map production pipelines.

The DA approach presented here reduces the discrepancy between the source domain (images with corresponding semantic labels) and the target domain (images without labels) so the segmentation model which is trained afterward can work effectively with both datasets. Given the application domain is RS-based satellite imagery, it is vital to ensure the adaptations are done in a way that preserves the meaningful land content and structure of the imagery. Thus, the key feature of the presented work is that the adapted images are highly accurate in terms of semantic consistency, i.e., the objects in the adapted images preserve their original logical meaning. Semantically consistent adaptation is crucial in RS because each pixel brings certain information which must be preserved. Focusing on semantically consistent DA approaches means that the solution presented here is increasingly important for government and private industries since existing map products (i.e., training labels) in RS are limited, not publicly available, or expensive to obtain.

Land use and land cover (LULC) maps are generated products resulting from satellite imagery that relate each pixel in the satellite image to a specific class of objects, e.g., vegetation, hydro, road, etc. Land use applications generally serve to monitor the changes in human economic and cultural activity on the land (i.e., recreation, agricultural, mining, etc.). Land cover, in turn, refers to the natural (rivers, forests, snow, etc.) or human-made objects (buildings, cars, roads, etc.) that exist on the ground [2]. It is land cover features that are detected using reflected energy recorded per pixel while the land use is inferred based on land cover elements.

Governments and commercial organizations involved in land management widely use LULC maps because they can provide valuable and accurate information if generated on a regular basis. For example, LULC maps are used for emergency response to efficiently deploy restoration forces after significant flooding or landslides. Demand for frequent and up-to-date LULC maps is increasing due to increased satellite platforms and corresponding sensors, thereby increasing the availability and volume of RS data. This big RS data has inevitably led to new variations of its use and application [1]. Furthermore, new constellations that...
provide data at higher temporal frequencies (i.e., shorter revisit periods) and broader area coverage are coming online [3].

The process of generating LULC maps involves many trained specialists. Before deep learning (DL) was introduced for RS data, the process of LULC map production was semiautomated. It required a human in the loop, but also made use of existing tools and algorithms [4]. To label one satellite image, a knowledgeable person needed to visually assess each image pixel and assign a corresponding label. This time-consuming and expensive labour, coupled with the desire for increased temporal frequency of LULC maps, drove the search for fully automated solutions. This need and the rise of DL algorithms [5] led to the development of segmentation algorithms for LULC map generation.

Semantic segmentation is a process of assigning each pixel a logical label, e.g., vegetation, road, background, etc. Recent achievements in image classification problems using DL convolutional neural networks (CNNs) led to significant advances in per-pixel segmentation tasks as well [6], [7], [8]. This, in turn, was used in a wide spectrum of computer vision applications [9], [10]. The success of CNN-based segmentation algorithms led to their use in automated development of LULC maps, with very good results [11], [12], [13]. However, the process of annotating a large number of images needed for training CNN-based semantic segmentation models is a significant bottleneck. For example, the release of each new satellite (and corresponding sensors) usually requires the creation of a new LULC labeled training dataset due to spatial and spectral differences of the new sensors. To reduce costs of developing new training datasets, the ability to develop methods to adapt labeled data from one domain (called the source domain) to another domain (called the target domain) would be very useful due to the expense of labeling new datasets. For example, different RS scene image datasets may be taken from different types of sensors, in different weather conditions, in different geographical areas, and have different resolutions and scales. Consequently, the domain distribution discrepancy may be significant from one dataset to another, which makes DL models trained on a source dataset not useful for new target domains.

To tackle this issue, researchers are investigating and have been developing DA techniques that are used to close the gap between source and target domains. Most common approaches aim to align features across two domains so that a semantic segmentation model can generalize across them [14], [15], [16], [17]. However, compared to classification tasks, feature adaptation in segmentation tasks is more complicated because the model has to encode a diversity of different visual characteristics, such as appearance, shapes, and context [15]. Another group of adaptation methods deal with a style transferring task and has shown excellent performance when applied to RS datasets [18], [19], [20].

Considering DA methods that utilize style-transferring approaches, the question becomes which method is more efficient, and thus, worth pursuing further. The dominance of CycleGAN-based methods [21] is challenged by models that contain an embedded adaptive instance normalization layer; for example, the Semi2I model proposed by Tasar et al. [19]. This model is focused on preserving the semantic consistency during the style-transferring phase, which is crucial in RS. However, the Semi2I architecture is not sufficiently sophisticated to provide high-quality image-to-image translation outputs because it was initially designed to reduced memory consumption and computation speed.

This article presents a new unsupervised DA method called high-resolution semantically consistent image-to-image translation (HRSemI2I) that employs the AdaIN [22] layer and aims to transfer the target domain’s style to the source images preserving semantic consistency and per-pixel image quality. This research proposes to significantly improve the architecture of the Semi2I model to boost model performance and make it competitive with the state-of-the-art CyCADA model [23]. We also test the CyCADA model on the RS multiband datasets which has not been applied to this application to date. CyCADA was initially trained on the synthetic photorealistic datasets, such as GTA5 and Synthia, and validated on the Cityscapes dataset of vehicle-egocentric real-world images.

The next section summarizes the research related to unsupervised DA methods, in general and for RS data. This is followed by a discussion of proposed network architecture and training methods including details of the source and target datasets as well as the experiments conducted.

II. RELATED WORKS

DA techniques are used to close the gap between source and target domains. They became popular after Ganin and Lempitsky [24], Ganin [25] proposed unsupervised DA through back-propagation. Since then, many variants and model architectures that fit the DA task have been proposed. The most common approaches aim to align features across two domains by using a semantic segmentation model that can generalize the two domains. For instance, a feature-level adaptation method was proposed by Tzeng et al. [26] that uses the so-called domain confusion loss, which directly minimizes the distance between source and target representations, thus initializing domain invariance. Also, there is a classification loss which solves the image classification task.

Hong et al. [27] proposed using a conditional generative adversarial network (GAN) for structured DA. The central part of the proposed model is a conditional generator which is aimed to enhance source domain features to have a similar distribution as the target. The conditional generator consists of several residual blocks. There is also a CNN encoder with five convolutional layers that extracts target features. The target domain features and enhanced source domain features are then passed through the discriminator represented by a multilayer perceptron. Thus, the DA task is resolved via adversarial learning. At the same time, the semantic segmentation task is solved after the cross-entropy loss is calculated by passing the enhanced source features through the deconvolutional layer.

Another feature-level adaptation method was proposed by Tzeng et al. [28], which combines adversarial learning with discriminative feature learning. In particular, during the training, the model must learn a discriminative mapping of target images
to the source feature space. The model consists of a CNN feature encoder and an FCC discriminator, and the DA is solved via adversarial learning. The methods mentioned above are limited by a common issue: they do not enforce semantic consistency while aligning feature representation of both domains. However, this is crucial when the final goal is the semantic segmentation of target RS samples. The following pixel-level DA methods were developed to resolve this issue.

Originally based on CycleGAN [21] and feature-level adaptation methods proposed in [24] and [28], the CyCADA model was proposed by Hoffman et al. [23] to preserve semantic consistency through pixel-level unsupervised DA. The model uses a noisy labeler, which is a semantic segmentation model trained on source data and applied to target data without adaptation. After that, a pixel-space adaptation process is performed using two generators and two discriminators. The first generator and discriminator work to translate the target domain feature representation to source samples and the second pair of generator and discriminator work as a whole to translate the source domain feature representation to target samples. The noisy labeler, trained previously, is integrated into the training process and encourages an image to be classified in the same way after translation as it was before translation, according to this classifier. Finally, the semantic segmentation model with feature-level adaptation is trained on source samples stylized as target ones.

Another pixel-level adaptation method that is aimed to preserve semantic consistency, which is called bidirectional learning (BDL), was proposed by Li et al. [29]. The authors explore a model where two separate modules cooperate. There is an image-to-image translation model and segmentation adaptation model, both similar to the ones proposed in [23]. The learning process involves two directions: “translation-to-segmentation” and “segmentation-to-translation,” and that is why the whole process is a closed-loop cycle. Moreover, a self-supervised learning (SSL) approach is incorporated into a segmentation adaptation module. This way, different from segmentation models trained only on source data, the segmentation adaptation model uses both source and target images for training. Iteratively predicting labels for the target domain, they are considered as the approximation of the ground truth labels. Thus, only those with high probability are used in training over and over again. After the segmentation model is trained on the source and target labels, it is used as the noisy labeler discussed above, thus making the training process looped. Even though these two methods are both focused on the pixel-level adaptation task, where semantic consistency is preserved, the source and target datasets they were trained on are represented by first-view car driving samples. The following pixel-space adaptation methods were applied to the RS imagery and are claimed to be effective.

In [17], Liu et al. proposed using a curve feature extractor to represent each pixel of the input image as a curve. A semantic segmentation model DeepLab v3+ [30], pretrained on the source images, is used to extract deep features for both domains. After the deep features are extracted, each pixel in the input image is converted into a feature curve. Then, the conditional generative adversarial network (cGAN) aligns the representation of the curves so that they are indistinguishable for the discriminator model regarding the source and target domains. This way, the discrepancy between the domains is reduced.

Another method was proposed by Liu et al. in [16], where the authors propose using the Kullback–Leibler constraint in their DA framework (KL-ADDA). The model consists of generator and discriminator parts. The generator is represented by the DeepLab v3+ framework, and the discriminator is a fully CNN, similar to [31]. First, the generator is pretrained on the source images with corresponding labels. Then, the images from both domains are passed through the generator, and the semantic labels are acquired, where a semantic loss for the source domain is calculated. After this, predicted labels are passed to the discriminator, which decides what domain they belong to; thus, the discriminator loss is calculated. Afterward, a KL divergence loss is calculated using the intermediate features extracted by the discriminator. Finally, the adversarial loss is calculated based on the discriminator loss and the KL divergence loss and is forwarded as a constraint for the generator. Even though this method represents the improved [15], it still lacks semantic consistency preservation.

In 2021, Shao and Zhang [32] proposed a statistical feature-based image-to-image translation method (SPatchGAN). The crucial part of this model is the specific architecture of a discriminator, which focuses on the statistical features instead of individual patches. The model’s architecture is represented by a forward generator, a SPatchGAN discriminator, and a backward generator that works with low-resolution images. The generator aims to produce outputs indistinguishable from the target domain images, and the discriminator’s goal is to assess the quality of the generated samples. The backward generator stabilizes the network during the training process and ensures the similarity between the source and stylized images. The discriminator in this model takes inputs from the generator and produces multiple outputs based on the number of scales and the number of features. Each scaling is conducted by a downsampling block, an adaptation block, a statistical feature calculation block, and several multilayer perceptrons. Opposite to the patch-based discriminators, the SPatchGAN derives the output from a group of statistical features of the picture. The model training includes optimization of the generator and discriminator through the adversarial loss. Along with this loss, the authors employ the so-called “weak cycle loss,” which is a reduced version of the cross-reconstruction loss, similar to the one in the CycleGAN model. This loss has only a forward cycle constraint which compares low-resolution original and translated images. The model became highly popular and proved the high quality of image-to-image translation in such applications as selfie-to-anime and male-to-female translation. However, unlike CyCADA, the SPatchGAN model lacks semantic consistency, and, thus, is inoperable for RS applications that require consistent representation of contextual features.

The whole group of the pixel-level adaptation methods which were validated on RS imagery were proposed by Tasar et al. [18], [19], [20], where the authors put special attention to the semantic consistency preservation. The first method is called ColorMapGAN and is aimed to linearly shift each source sample band distribution to match the target domain distribution. At first, a
The discriminator part of the GAN consists of a generator and a discriminator. The generator is an architecturally simple construction that is represented by only scaling and shifting matrices. Thus, each source sample band (red, green, and blue) is passed through the scaling and shifting operations. The output of such transformations is then passed through the discriminator, which is similar to [31]. The discriminator decides how close the transformed source sample is to the target distribution. The generator’s goal is to fool the discriminator by faking the source images. It is crucial to notice here that the semantic consistency is preserved during such transformations because there are neither convolutional nor pooling layers. The final step of the training process is to fine-tune the initial classifier with faked source samples and their corresponding labels.

The drawback of the ColorMapGAN is that it processes each band separately, which generates slightly noisy outputs. Another model proposed by these authors is called DAugNet. There is only one image-to-image translation part for the source-to-target and target-to-source style directions. There is also only one discriminator which estimates how accurate the style translation was. It can be used to evaluate both direction translations because it has domain-specific output layers. The scaling and shifting of feature representation are performed by constant predefined target-style vectors for the mean and variance, which do not change during the training. The same losses enforce the semantic consistency of the transformed samples as in the previously discussed method. After the style transferring part is done, the target-like source samples are used for semantic segmentation model training and validation on the target dataset.

The last method that was developed with application to RS data is called semantically consistent image-to-image translation (Semi2I). The idea of this method is close to [21] but has some specific differences. First, the image-to-image translation module operates using the AdaIN layer between the encoder and decoder part. During training, this layer scales and shifts the input source sample feature representation to match the target domain’s accumulated mean and variance. Also, the image-to-image part consists of relatively shallow convolutional models; thus, it can be trained quickly. The following losses enforce semantic consistency of the translated source images: cross-reconstruction loss, self-reconstruction loss, and image gradient loss computed for the original source image and its translated to target domain representation copy. In addition, the authors resize the low-level features extracted by the first convolutional encoder layer and concatenate them to each deconvolution layer in the corresponding decoder. After the image-to-image part is trained, a semantic segmentation model is trained on translated source images with corresponding labels and validated on the target dataset. Inspired by [19], this article proposes another unsupervised DA model which utilizes an AdaIN layer and is highly semantically consistent while performing the style-transferring task. The best building blocks from [19] and [33] are taken and combined, so the resulting model achieves the state-of-the-art performance in the adaptation of the RS imagery.
of the $G_{T \rightarrow S}$. After the deep features for both domain inputs are extracted, the mean $\mu$ and variance $\sigma$ are calculated for each domain input. Since the mean and variance can significantly vary from input to input even within the same domain, the global mean and variance are used. This approach finds balanced values for each parameter through the accumulation process using the formulas

$$\mu_{\text{glob}} = \text{decay} \_\text{rate} \times \mu_c + (1 - \text{decay} \_\text{rate}) \times \mu_c$$

$$\sigma_{\text{glob}} = \text{decay} \_\text{rate} \times \sigma_c + (1 - \text{decay} \_\text{rate}) \times \sigma_c$$  \hspace{1cm} (3)

where $\mu_{\text{glob}}$ and $\sigma_{\text{glob}}$ are the global mean and variance, and $\mu_c$ and $\sigma_c$ are mean and variance for the current input image batch. After current the $\mu_c$ and $\sigma_c$ are calculated for both domains, the extracted features from each domain encoder are scaled and shifted and the resulting features are passed through the opposite domain decoder. Using the following formulas, fake$_B$ and fake$_A$

\begin{align*}
\text{fake}_A &= \text{Dec}_A(S(\text{Enc}_B(I_s), \text{Enc}_A(I_t))) \\
\text{fake}_B &= \text{Dec}_B(S(\text{Enc}_A(I_s), \text{Enc}_B(I_t)))
\end{align*}  \hspace{1cm} (4)

where $I_s$ and $I_t$ are the source and target samples, respectively, and $S()$ denotes the AdaIN layer. After that, assigned discriminators are used to evaluate how close the fake$_A$ and fake$_B$ are to the source and target domain distribution, respectively, the total adversarial loss for both generators is calculated, and the weights are optimized. Then, the discriminators’ weights are also updated.

**Semantic Consistency:** The proposed method employs several constraints to enhance the semantic consistency of the style-transferring operation. It is a cross-reconstruction loss, which is an L1-norm of the original image and its reconstructed version. After the fake output is acquired, it is passed through the opposite generator to get a fake version of the faked original image. Ideally, the reconstructed image and the original one must be the same. The cross-reconstruction loss can be expressed as

$$L_{\text{cross}} = |I_s - \text{Dec}_A(S(\text{Enc}_B(\text{fake}_B), \text{Enc}_A(\text{fake}_A)))|$$

$$+ |I_t - \text{Dec}_B(S(\text{Enc}_A(\text{fake}_A), \text{Enc}_B(\text{fake}_B))))|.$$  \hspace{1cm} (5)

Another constraint is known as self-reconstruction loss. After the embedding is extracted by the related encoder (before the AdaIN layer), it is passed through the same-domain decoder, and the L1-norm is calculated for the original image and its self-reconstructed version. The loss function can be expressed as

$$L_{\text{self}} = |I_s - \text{Dec}_A(\text{Enc}_A(I_s))|$$

$$+ |I_t - \text{Dec}_B(\text{Enc}_B(I_t))|.$$  \hspace{1cm} (6)

Also, a gradient loss is used as an additional semantic consistency constraint. After the fake version of the original image is generated, the first-order image derivative is calculated for both using a Sobel operator [34]. Having the difference between them as small as possible, the model is forced to preserve the edges of the objects in the training images. The gradient loss can
Fig. 3. Graphs depicting dataset label distributions. Left: Label distribution of the source dataset. Right: Label distribution of the target dataset.

Fig. 4. Histogram of pixel value distribution of the original target domain dataset before and after applying a smoothing filter. The blue, red, green, and black lines represent a number of pixel values in the blue, red, green, and near IR channels of the entire dataset, respectively.

| TABLE I | MODELS’ PERFORMANCE (IoU) |
|---------|---------------------------|
| **Model** | **mIoU** | **background** | **vegetation** | **hydro** | **roads** | **buildings** |
| **No Adaptation** | | | | | | |
| Baseline | 53.02 | 69.48 | 68.77 | 73.78 | 36.63 | 16.43 |
| **Unsupervised Domain Adaptation** | | | | | | |
| Sem2I | 60.25 | 74.80 | 76.81 | 77.34 | 42.18 | 30.11 |
| CyCADA | 63.92 | 75.47 | 75.60 | 81.52 | 46.67 | 40.35 |
| HRSem2I (ablated) | 52.13 | 70.18 | 80.39 | 70.32 | 29.00 | 22.90 |
| HRSem2I | 63.99 | 76.46 | 80.39 | 80.35 | 44.96 | 37.38 |

be expressed as

\[
L_{\text{grad}} = |\text{grad}(I_s) - \text{grad}(\text{fake}_B)| \\
+ |\text{grad}(I_t) - \text{grad}(\text{fake}_A)| \quad (7)
\]

where grad(·) is a spatial gradient operator.

Training: Given the source domain images designated as \(I_s\), and the target domain images denoted as \(I_t\), the training process for the DA network can be expressed in the following iterations.

1) The global mean and variance variables are initiated with zeros.

2) The inputs \(I_s\) and \(I_t\) are passed through the corresponding modules of \(G_A\) and \(G_B\), where \(G_A\) and \(G_B\) denote the generators assigned to the source-to-target and target-to-source transformations, respectively.

3) Freeze the weights of the discriminators. Then, fake \(A\) and fake \(B\) are passed through the corresponding discriminator network, and the probabilities are acquired.

4) After the outputs are generated, the subsequent losses of the generators are calculated: the cross-reconstruction loss, the self-reconstruction loss, the gradient loss, and the adversarial loss of the generators.
5) The weights of the generators are updated.
6) Unfreeze (or activate) the discriminator weights. Then, 

\[ \text{fake}_A \text{ and } \text{fake}_B \text{ are passed through the corresponding discriminator networks, the probabilities are acquired, and the adversarial loss of the discriminators is calculated.} \]

7) The weights of the discriminators are updated.
8) The global mean and variance are updated.
9) After the training process is done, all generators are saved along with the global mean and variance vectors.

**Testing:** The testing stage uses only certain modules from each of the generators. To get the \( \text{fake}_B \) output, which is a source domain image, transferred to the target domain style, the encoder part of the \( G_A \) and the decoder part of the \( G_B \) are needed together with the resulting global mean and variance parameters. The inference of the trained DA model can be expressed as

\[
\text{fake}_B = \text{Dec}_B \left( St \left( \text{Enc}_A \left( I_s \right), \left( \mu_{\text{target}}^{\text{glob}}, \sigma_{\text{target}}^{\text{glob}} \right) \right) \right)
\]  

(8)

where \( \mu_{\text{target}}^{\text{glob}} \) and \( \sigma_{\text{target}}^{\text{glob}} \) are the global mean and variance, calculated for the target domain inputs during the training stage.

**IV. EXPERIMENTS**

**A. Datasets**

The source and target datasets used in this work were acquired by WoldView-2 and SPOT-6 satellites, respectively. Both were automatically annotated by five semantic classes: 1) background, 2) vegetation, 3) hydrology, 4) roads, and 4) buildings. The original satellite images were provided by Natural Resources Canada [2] as preprocessed rasters (GeoTIFFs) with corresponding labels in vector format (GeoPackage).

The domain with known labels (source) for this project is represented by the WoldView-2 imagery dataset, which is a preprocessed 0.5-m spatial resolution multiband 8-bit (resampled) set of images taken in the spring, summer, and fall months across Canada. The images were further downsampled to 1.5-m spatial resolution to match the resolution of the target images. The GeoTIFF raster images and corresponding label geopackages were cropped into samples of size \( 4 \times 512 \times 512 \) and \( 1 \times 512 \times 512 \), respectively, where 4 represents the number of bands (blue, green, red, and near-infrared) in the image file and 1 represents a single band of the corresponding label file which was rasterized and saved in GeoTIFF format. The total number of source samples is 5560, with the label distribution represented in Fig. 3.

SPOT-6 imagery was used as a target domain or the domain where labels do not participate in training the model and which style must be transferred to the source images. This dataset is a preprocessed 1.5-m 8-bit set of images taken in the spring, summer, and fall months across Canada. The GeoTIFF raster images and corresponding label geopackages also were cropped into samples of size \( 4 \times 512 \times 512 \) and \( 1 \times 512 \times 512 \), respectively, the same as for the source dataset. Since the originally
Fig. 6. Original source image (left column), its stylized version by CyCADA (central column), and its stylized version by the proposed model (right column).

provided SPOT-6 imagery had a spiky pixel value distribution (as shown in Fig. 4), the whole dataset was smoothed using a Gaussian filter [34] with the sigma parameter equal to (1, 1, 0) and the order to 0. The original pixel value distribution and the resulting distribution are depicted in Fig. 4. The total number of target samples is 4735, with the label distribution given in Fig. 3.

B. Training Parameters

Domain Adaptation Model: The DA model was trained with the following parameters. The generator of the model was optimized by the Adam optimizer with parameters beta1 and beta2 equal to 0.5 and 0.999, respectively. The initial learning rate was equal to $10^{-4}$ and was decayed over training steps with the formula

$$lr_c = lr_b \times \frac{\text{iter}_m - \text{iter}_c}{\text{iter}_m - \text{iter}_d}$$

(9)

where $lr_c$ is the current learning rate, $\text{iter}_c$ is the current training iteration, $\text{iter}_m$ is the maximum number of training iterations, and $\text{iter}_d$ is the number of iterations where decaying starts. The total number of training steps was 100 000, and the learning rate decaying step was 75 000. The discriminator of the model was optimized using the Adam optimizer with the same parameters as were used for the generator’s optimizer except for the initial learning rate, which was set to $10^{-5}$. Since the style-transferring part of the objective function has many composing losses, the following coefficients were assigned to each of them. The adversarial losses for the generator (in both directions) are multiplied by 1, the cross-reconstruction loss is multiplied by 20, the self-reconstruction loss is multiplied by 10, and the edge loss is multiplied by 25. The images from both domains were normalized from $(-1, 1)$ and then packed in batches of size 1.

Segmentation Model: After the DA model is trained, all source domain images are transferred to the style of the target domain. A segmentation model was used to evaluate the quality of the translated images. It was represented by the DeepLab v2 framework [6] with a modified number of input channels equal to 4. Same as for the SemI2I method, the original source images were mixed with their stylized versions and used as a training dataset. Training the model using only stylized images did not improve the segmentation performance significantly. As an optimizer, the Adam method was chosen. The initial learning rate and weight decay were set to $1 \times 10^{-4}$ and $5 \times 10^{-4}$, respectively. During training, the learning rate was decreased using polynomial decay with a power of 0.9 using

$$lr_c = lr_b \times \left(1 - \frac{\text{iter}_c}{\text{iter}_m}\right)^p$$

(10)

where $lr_c$ is the current learning rate, $\text{iter}_c$ is a current training iteration, $\text{iter}_m$ is a maximum number of training iterations, and $p$ is a decaying power. The model was trained with eight images in a batch, over 90 000 steps. The validation was performed over the whole target dataset. Prior to the segmentation model training, the mean of each band for all datasets (training and
RESULTS

Results of the proposed DA model were compared with SemiI2I, CyCADA models, and with the baseline model. Similarly, results of an ablated version of the proposed model are given to show the effect of the self-reconstruction and gradient losses as an essential component for preserving semantic consistency in DA models. Comparisons with other state-of-the-art image-to-image translation models, such as SPatchGAN, CycleGAN, or BDL are not given due to the lack of semantic consistency mechanisms in these approaches. For example, SPatchGAN has a statistical feature-based discriminator, which expects similar features in both domains to be located at the same place in the satellite images. In remote sensing, however, source and target domain samples represent a wide variety of possible combinations of the on-ground objects. Thus, using this model for RS data does not make sense since translated images would show little semantic consistency here consistency with the original images. The CycleGAN model, however, has a self-reconstruction loss that provides some level of stability in the image translation. However, it is not enough for artifact-free translation. For instance, a translated satellite image can be populated by buildings which were not presented in the source image. The BDL model utilizes a noisy labeler, similar to the CyCADA, along with the SSL steps. Experimenting with this model showed that the SSL iterations resulted in some less represented classes (such as buildings and roads) were removed in favor of more represented classes, such as hydro and vegetation. Clearly, this led to significantly deteriorated overall accuracy of the segmentation task, and, as a result, only the models with adequate semantic consistency mechanisms were considered for comparison. The baseline model is a segmentation model trained on data that was not adapted. In this case, the baseline model was trained strictly on source data and then validated on the target dataset. It has the same architecture and the same training parameters as the model discussed in Section IV-B.

The numerical results of the style transferring phase are presented in Fig. 5. The original source dataset pixel values are primarily grouped between 0 and 50, with the near IR band values peak at 100. The original (smoothed) target pixel values are between 50 and 150, with the near IR band values peaking at around 175. The bottommost plot shows the stylized source dataset where pixel values look similar to the target pixel value distribution.

As can be seen from Table I, the proposed HRSemiI2I method demonstrated performance improvement, both overall (∼3% to 10%) and per-class compared to the SemiI2I model and the baseline model. Moreover, the performance is comparable with the CyCADA model, which is considered the state-of-the-art DA method.

Even though, numerically, the proposed and CyCADA models produce similar metric values, visually, the proposed model has a significant advantage. As can be seen in Fig. 6, some of the style-transferred images generated by the CyCADA model have a noticeable pattern structure. In the uppermost example, the water body was replaced by square patches, which look like a green field. This could be explained due to the original source image’s green appearance and a noisy labeler marked the water body as a green field, which led to semantically inconsistent style translation. Similar pattern structure can be seen in the middle and bottommost examples. The outputs of the proposed model, however, demonstrate a free-of-patterns structure with preserved semantic meaning for the objects. The ablated HRSemiI2I model, instead, shows very low performance because of a lack of semantic consistency between the original and stylized imagery. Some additional visual results of the proposed method are also given in Fig. 7.

The role of semantic consistency becomes even more evident when applied to satellite images with high spatial resolution. The application of high spatial resolution images to classify fine
objects, such as cars, roads, individual trees, etc., will require
perfect style translation, where all semantic objects preserve
their original meaning. With the growing fleet of high-resolution
satellites, this solution becomes even more relevant.

VI. CONCLUSION

Big Earth data generated from a number of satellites is now
available; however, the labeling process for these images is
expensive and time-consuming. Moreover, multigenerational
satellites, such as the Landsat constellation may have multiple
years of labeled data where more recent satellites may not
currently have any labeled data publicly available data. There-
fore, unsupervised DA methods may be used to facilitate the
annotating of the images. One of the most recent DA approaches
is style-transferring when the style of the target domain trans-
sfers to the source domain images. However, it is essential to
provide semantic consistency and per-pixel accuracy during the
style-transferring process because each pixel in the RS image is
meaningful.

The proposed model presented here improves the previously
developed idea of using an adaptive instance normalization layer,
maintaining semantic consistency and per-pixel accuracy for the
style transferred images. The results were compared to the state-
of-the-art CyCADA model, never applied to RS applications,
using the WorldView-2 and SPOT-6 datasets. The results of the
proposed model are comparable to the results of the CyCADA
model, however, our model is better at preserving semantic
consistency which is essential for RS applications. The ablation
of the semantic consistency losses in the proposed method
showed how models without consistency in spatial features will
not perform as well in RS applications.

The future development of the proposed method could include
ecological domain transfer, a priori evaluation of dataset quality
in terms of data distribution, or exploration of residual blocks
in the encoder. First, both datasets (the source and the target)
are composed of satellite images taken in different ecological
regions, and, thus, represent different style characteristics.
Potentially, there could be a situation when the target domain is
represented by so many samples taken from different regions
that the task of acquiring its common style becomes meaningless
because of the increased complexity of the style-transferring
task. A region-to-region DA approach can lower the complexity
and increase the accuracy of the style-transferring task. Sec-
ond, the target dataset was smoothed by the Gaussian filter to
overcome its spiky pixel value distribution. The next step
could be replacing the target dataset with one where the pixel
value distribution is initially smoothed and evaluating the seg-
mentation performance in this case. Third, the proposed DA
model distinguishes from the Sem2I by different encoder and
decoder structures of the generator. However, the last three
residual blocks of the encoder remain unchanged throughout
the methods. As was explored by Hong et al. [27], the number of
the residual blocks in the generator can substantially influence
the DA results. Therefore, changing the number of the residual
blocks is worth further exploration.

REFERENCES

[1] Y. Li, J. Ma, and Y. Zhang, “Image retrieval from remote sensing Big Data:
a survey,” Inf. Fusion, vol. 67, pp. 94–115, 2021.
[2] Natural Resources Canada, Ottawa, ON, Canada, “Natural resources
Canada,” Jul. 2016. Accessed: 2022. [Online]. Available: https://www.
nrccan.gc.ca/home
[3] D. Dobrevic, “Space calendar 2022: Rocket launches, sky events, mis-
sions & more! space,” 2022. [Online]. Available: https://www.space.com/
32286-space-calendar.html
[4] D. Lu and Q. Weng, “A survey of image classification methods and
techniques for improving classification performance,” Int. J. Remote Sens.,
vol. 28, no. 5, pp. 823–870, Mar. 2007.
[5] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521,
no. 7553, pp. 436–444, 2015.
[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille,
“DeepLab: Semantic image segmentation with deep convolutional nets,
atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal.
Mach. Intell., vol. 40, no. 4, pp. 834–845, Apr. 2018.
[7] G. Lin, C. Shen, A. van den Hengel, and I. Reid, “Efficient piecewise
training of deep structured models for semantic segmentation,” in Proc.
IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3194–3203.
[8] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with
recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern
Recognit., 2017, pp. 1377–1385.
[9] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving?
the KITTI vision benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern
Recognit., 2012, pp. 3354–3361.
[10] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang,
“Deep image harmonization,” in Proc. IEEE Conf. Comput. Vis. Pattern
Recognit., 2017, pp. 3789–3797.
[11] V. Alhassan, C. Henry, S. Ramanna, and C. Storlie, “A deep learning frame-
work for land-use/land-cover mapping and analysis using multispectral
satellite imagery,” Neurol. Comput. Appl., vol. 32, no. 12, pp. 8529–8544,
Jun. 2020.
[12] C. D. Storie and C. J. Henry, “Deep learning neural networks for land
use land cover mapping,” in Proc. IEEE Int. Geosci. Remote Sens. Symp.,
2018, pp. 3445–3448.
[13] C. J. Henry et al., “Automated LULC map production using deep neural
networks,” Int. J. Remote Sens., vol. 40, pp. 4416–4440, 2019.
[14] I. Goodfellow, “Generative adversarial nets,” in Proc. Adv. Neural Inf.
Process. Syst., vol. 27, 2014, pp. 1–9.
[15] Y.-H. Tsai, W.-C. Hung, S. Schulte, K. Sohn, M.-H. Yang, and M.
Chandraker, “Learning to adapt structured output space for semantic
segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,
2018, pp. 7472–7481.
[16] M. Liu, P. Zhang, Q. Shi, and M. Liu, “An adversarial domain adaptation
framework with KL-constraint for remote sensing land cover classifica-
tion,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 3002305.
[17] W. Liu and F. Su, “Unsupervised adversarial domain adaptation network
for semantic segmentation,” IEEE Geosci. Remote Sens. Lett., vol. 17,
no. 11, pp. 1978–1982, Nov. 2020.
[18] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, “ColorMapGAN:
Unsupervised domain adaptation for semantic segmentation using color
mapping generative adversarial networks,” IEEE Trans. Geosci. Remote
Sens., vol. 58, no. 10, pp. 7178–7193, Oct. 2020.
[19] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, “SEMI2I: Semantically
consistent image-to-image translation for domain adaptation of remote
sensing data,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2020,
pp. 1837–1840.
[20] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, “DAugNet:
Unsupervised, multisource, multitarget, and life-long domain adaptation for
semantic segmentation of satellite images,” IEEE Trans. Geosci. Remote
Sens., vol. 59, no. 2, pp. 1067–1081, Feb. 2021.
[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image
translation using cycle-consistent adversarial networks,” in Proc. IEEE
Int. Conf. Comput. Vis., 2017, pp. 2223–2232.
[22] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with
adaptive instance normalization,” in Proc. IEEE Int. Conf. Comput. Vis.,
2017, pp. 1501–1510.
[23] J. Hoffman et al., “CyCADA: Cycle-consistent adversarial domain adap-
tation,” in Proc. 35th Int. Conf. Mach. Learn., (ser. Proc. Mach. Learn.
Workshops), vol. 80, 2018, pp. 1989–1998.
[24] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by back-
propagation,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.
Mikhail Sokolov received the bachelor’s degree in applied mathematics and the master’s degree in mathematics from the Ural State University, Yekaterinburg, Russia, and the master’s degree in applied computer science from the University of Winnipeg, Winnipeg, MB, Canada.

He is currently a Physical Science Specialist with Natural Resources Canada, Ottawa, ON, Canada. His research interests include solving semantic segmentation and domain adaptation problems in remote sensing.

Christopher Henry (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2004, 2006, and 2011, respectively.

He is currently a Professor of Applied Computer Science with the University of Winnipeg, Winnipeg, MB, Canada. He has also pioneered approaches to classify pixels obtained from satellite images using deep neural networks developed for semantic segmentation for the creation of land-use/land-classification maps. He has also been collaborating for many years on GPU-based computing initiatives, he has worked to establish a provincial consortium of researchers in high-performance computing. He cofounded the Applied Parallel Computing and Collaborative Research Laboratory and co-led the process to establish the University of Winnipeg as an NVIDIA GPU Education Centre. His research interests include the theory and application of machine learning, such as his work in developing machine learning data sets for digital agricultural applications.

Victor Alhassan received the M.Sc. degree in applied computer science and society from the University of Winnipeg, Winnipeg, MB, Canada, in 2018.

He is currently a Geospatial Analyst with Natural Resources Canada, Ottawa, ON, Canada. His research interests include machine learning, computer vision, geospatial analysis and natural language processing. His current work focuses on research and application of deep learning models on remotely sensed data.

Mathieu Turgeon-Pelchat received the bachelor’s degree in applied geomatics and the master’s degree in remote sensing from the Université de Sherbrooke, Sherbrooke, QC, Canada.

He has been with the Canada Centre for Mapping and Earth Observation, Natural Resources Canada, Ottawa, ON, Canada, since 2016. His research interests include applying machine/deep learning methods to remote sensing data for the automation of land use and land cover data creation.