

Improving Land Cover Segmentation Across Satellites Using Domain Adaptation

Nadir Bengana  and Janne Heikkilä , *Senior Member, IEEE*

Abstract—Land use and land cover mapping is essential to various fields of study, such as forestry, agriculture, and urban management. Generally, earth observation satellites facilitate and accelerate the mapping process. Subsequently, deep learning methods have been proven to be excellent in automating the mapping via semantic image segmentation. However, because deep neural networks require large amounts of labeled data, it is not easy to exploit the full potential of satellite imagery. Additionally, land cover tends to differ in appearance from one region to another; therefore, having labeled data from one location does not necessarily help map others. Furthermore, satellite images come in various multispectral bands, which range from RGB to over 12 bands. In this study, our aim is to use domain adaptation (DA) to solve the aforementioned problems. We applied a well-performing DA approach on the DeepGlobe land cover dataset as well as datasets that we built using RGB images from Sentinel-2, WorldView-2, and Pleiades-1B satellites with CORINE Land Cover as ground truth (GT) labels. The experiments revealed significant improvements over the results obtained without using DA. In some cases, an improvement of over 20% mean intersection over union was obtained. Sometimes, our model manages to correct errors in the GT labels.

Index Terms—Domain adaptation (DA), image segmentation, land cover segmentation.

I. INTRODUCTION

OVER the last few years, remote sensing (RS) data became easily obtainable thanks to the surge of open data provided by earth observation (EO) satellites. Notably, the data from satellites, such as Sentinel-2 [1] and Landsat [2], are available to the public free of charge. These satellites provide high-resolution multispectral imagery (up to 10 m), which facilitates the application of multiple methods in data processing.

Land cover represents the (bio)physical cover on the earth's surface, whereas land use is the cover resulting from human action (e.g., a wheat field is an example of land use, whereas an ocean is strictly land cover). Land use and land cover (LULC) mapping is among the most crucial RS applications that help monitor forests, agricultural areas, and oceans, among others. The mapping can be performed manually by looking through satellite images [3]. However, this approach is both costly and

time consuming. Additionally, no fine global land cover map exists. CORINE Land Cover (CLC) [4] provides land cover mapping with a pixel resolution of 100 m/px that covers Europe only, and is updated roughly once every six years. Moderate Resolution Imaging Spectroradiometer [5] provides a global land cover map that is updated annually with a pixel resolution of 500 m/px, which might be too coarse for many applications, such as urban cover monitoring.

Several methods exist for performing LULC mapping, depending on the available data and desired accuracy. The simplest approach is land cover classification, which labels patches on the basis of the majority land cover type. Semantic pixel segmentation, which, as the name suggests, labels each pixel, is an approach that is considered more challenging than classification. Recently, in semantic segmentation, including LULC segmentation, classical machine learning (ML) tools have fallen out of favor. Penatti *et al.* [6] showed that convolutional neural networks vastly outperform the classical ML methods in terms of land cover classification. In the land cover segmentation section of the DeepGlobe challenge [7], the leaderboards were completely dominated by deep neural networks (DNNs) [8]–[10].

Although deep learning (DL) methods have a good performance, they require a huge amount of data to show their true potential. As mentioned previously, although EO data are available free of charge, fine ground truth (GT) labels are rare. Moreover, although CLC has a resolution of 100 m/px, a finer version exists with a resolution of 20 m/px covering Finland [11], as well as a 10 m/px version covering Germany [12]. Although these versions cover a limited area and exhibit some inaccuracies, they can still be used to train a DNN for land cover segmentation.

Because land cover types vary depending on the location, having a DNN model that performs land cover segmentation for one area does not guarantee that it will perform equally as well on other areas. Additionally, the images captured by various satellites are different because of the mismatch in the capture time, pixel resolution, radiometric resolution, and other properties. These variables result in a domain shift between the datasets acquired from one satellite covering a particular region and another satellite covering either the same region or a different one. Therefore, to obtain consistent results, a model needs to be trained on each data variation, which requires massive labeled datasets from each satellite.

Removing the domain shift between different datasets is called domain adaptation (DA). In semantic segmentation, DA is used to segment images from a target dataset with the help of a source dataset. Generally, two types of DA approaches

Manuscript received June 22, 2020; revised September 6, 2020, October 14, 2020, and November 14, 2020; accepted November 26, 2020. Date of publication December 22, 2020; date of current version January 6, 2021. This work was supported by the Business Finland under Grant 1259/31/2018. (*Corresponding author: Nadir Bengana.*)

The authors are with the Faculty of Information Technology and Electrical Engineering, University of Oulu, 90014 Oulu, Finland (e-mail: mohamed.bengana@oulu.fi; janne.heikkila@oulu.fi).

Digital Object Identifier 10.1109/JSTARS.2020.3042887

exist: supervised, in which some or all of the target data are labeled, and unsupervised, in which the target data are unlabeled. Usually, when the labeled data are scarce, transfer learning is used. Transfer learning involves using a pretrained model on a labeled dataset and then training it with the scarce dataset using the full model or freezing part of it. DA can improve the overall results achieved with simple transfer learning as it can be unsupervised and, thus, requires no labels.

In this study, we focused on applying deep DA to segment RGB satellite data from different satellites and locations semantically. Our contributions are as follows.

- 1) To the best of our knowledge, this is the first study in which DA is applied to achieve LULC mapping from widely different areas around the globe using RGB bands only.
- 2) We built RGB datasets from the Sentinel-2, WorldView-2, and Pleiades-1B satellites using CLC as labels.
- 3) We customize and improved upon an existing DA method to fit satellite imagery.

The rest of this article is organized as follows. Section II presents some of the related background research. Section III introduces the materials used, including the satellite images and labels. Section IV describes the methods used. Section V outlines the performed experiments and the obtained results. Finally, Section VI is the conclusion, which also contains our reflection regarding the results.

II. RELATED WORK

A. LULC Segmentation

Image segmentation in LULC is performed differently when compared to other fields, such as street-view images. This is because satellite imagery comes in multispectral forms, including active sensing images, such as synthetic-aperture radar or LIDAR.¹

Soon after the first launch of the first publicly available RS satellite in 1972, computer vision started to be used to map LULC. Methods, such as histogram thresholding [13], provided acceptable results but exhibited problems associated with the variations in satellite images. Research on other methods, mainly statistical ones based on maximum *a posteriori*, had varying degrees of success. In 2000, the commercial software eCognition [14] combined classification methods, edge detection, and segmentation into one solution. Neural network methods at this point were unfavorable because of their high computational complexity.

In the early 2000s, the primary methods applied for LULC mapping were based on classical ML methods, such as support vector machines and decision trees [15], [16]. However, by the 2010s, DNNs began to emerge but were still not used much in RS imagery because of the lack of labeled data required to train the DNNs. In particular, using DL to tackle land cover segmentation has not been adequately addressed in the literature so far. For example, although state-of-the-art semantic segmentation methods have been translated to satellite imagery, the results were not as good [8]–[10]. This is because land cover

types have random shapes, such as forest stands and bodies of water, whereas objects in ordinary photos have consistent shapes that are easier to learn. Kuo *et al.* [10] proposed a method that provides one of the leading results in the DeepGlobe challenge, in which improving the performance depended on a variation of DeepLabV3+ [17] wherein atrous spatial pyramid pooling replaces the fully connected layers of the ResNet backbone. Additionally, DeepLabV3+ uses an encoder–decoder architecture to reduce the effect of resolution loss due to pooling and strided convolution. The encoder–decoder method is a common approach in land cover segmentation because it preserves high-resolution features, such as texture and color, which play a significant role in distinguishing land covers. Arief *et al.* [18] used a state-of-the-art semantic segmentation method and added LIDAR data, which slightly improved the results in comparison to other methods.

In LULC, especially when working with vegetation, the near infrared (NIR) band is usually applied directly, or as the normalized difference vegetation index (NDVI).² Generally, vegetation is highly reflective on the NIR band, which makes it useful for detecting forestry and plant types on the surface. NDVI is considered as an easy and quick way to segment vegetation from nonvegetation [19].

B. Domain Adaptation

Generally, DA reduces the domain shift between two different sets (source and target) with different distributions, which is achieved by aligning the distribution of one set to match that of another set or mapping both sets to a common space.

Mainly there are three forms of DA approaches [20], [21]. The first method is to minimize the distance between the source and the target data. Maximum mean discrepancy is an example of minimization aimed at achieving a domain-invariant feature representation that performs well in both source and target domains. The second method is adversarial DA that uses generative adversarial networks (GANs) [22] to make one set appear similar to another. Tzeng *et al.* [23] outlined an example of this adversarial method, in which the target data are translated to the source data by a discriminator to differentiate between the two. The third method involves creating a shared representation for both domains by translating them into a common space. CycleGAN [24] is an example of this third method, in which two discriminators are used to map images from the source data to the target data, and vice versa.

C. DA for Semantic Segmentation

Mostly, DA is a perfect fit for semantic segmentation since the latter requires pixel annotations, which, as mentioned previously, might not be available.

Generally, DA methods used for classification do not translate well to semantic segmentation [25]. Therefore, adversarial and reconstruction methods are preferred. Architectures, such as FCNs in the Wild [26] and No More Discrimination [27], are

¹LIDAR uses light in the form of a pulsed laser to measure ranges.

²NDVI is obtained by dividing the difference between the NIR band and red band over the sum of the two.

examples of adversarial DA whose aim is to use a GAN to generate sourcelike images and then segment theme images using a network trained on the source data. Notably, the reconstruction approach has been tested in many methods with different variations [28]–[31]. The datasets used were almost exclusively street-view image datasets, including Cityscape [32], GTA5 [33], and SYNTHIA [34]. Chang *et al.* [29] proposed a domain-invariant structure extraction framework to disentangle images into domain-invariant structure and domain-specific texture representations. Li *et al.* [35] proposed adding an extra step called bidirectional learning (BDL). The principle behind this step is to alternate between segmentation learning and image translation, which is supervised using the segmentation model. Generally, BDL prevents the translation model from converging to a point at which the discriminator views the images as being from the same distribution while not aligning the classes correctly, causing the segmentation to fail.

D. Domain Adaptation in RS

Various studies on RS use the term “domain adaptation.” However, because RS is a vast field, it is tricky to classify such studies. As described by Tuia *et al.* [36], some methods are based on selecting invariant features on the training data [37], [38], whereas others are based on the adaptation of the data distribution [39]–[41], and lastly building the adaptation in the classifier. The first type of method may be time consuming because of the difficulty of selecting invariant data, which might require building a new dataset for each target data point. The second type is what usually comes to mind when DA is mentioned. In this type, two distributions are translated to remove the domain shift between them, and the methods applied range from basic ones, such as principal component analysis, to adversarial ones. Finally, the third type is based on building a model that can process both data distributions in the same way. Usually, this requires semisupervised learning or is accompanied by an adaptation of the data distribution, which is what this study tackles.

Although the work of Benjdira *et al.* [40] is similar to ours, in their article, they only tackled the adaptation of data distributions. However, in our study, our goal was to adapt both the data and the model, and we also did not limit our work to urban environments.

III. MATERIALS

A. Satellite Data

The satellite data used in this study were obtained from four EO satellites: Sentinel-2, WorldView-2, Pleiades-1B, and WorldView-3.³

The Sentinel-2 constellation is composed of two polar-orbiting satellites (Sentinel-2 A and Sentinel-2B) placed in the same Sun-synchronous orbit, phased at 180° to each other. Each of them has a multispectral sensor with 12 bands with a resolution ranging from 60 up to 10 m/px, of which we used the RGB bands with a resolution of 10 m/px, as shown in Table I.

TABLE I
SENTINEL2-B PROPERTIES

Bands	Bandwidth (nm)	Central wavelength (nm)	Spatial resolution (m)
B2: Blue	66	492.1	10
B3: Green	36	559	10
B4: Red	31	665	10

TABLE II
WORLDVIEW-2B PROPERTIES

Bands	Bandwidth (nm)	Central wavelength (nm)	Spatial resolution (m)
B2: Blue	73	478	1.84
B3: Green	70	546	1.84
B5: Red	60	659	1.84
B7: NIR	136	831	1.84

TABLE III
PLEIADES-1B PROPERTIES

Bands	Bandwidth (nm)	Central wavelength (nm)	Spatial resolution (m)
B2: Blue	80	490	2
B3: Green	80	550	2
B4: Red	80	660	2
B5: NIR	140	845	2

From the European space agency’s third party mission, data from the WorldView-2 satellite are available as WorldView-2 European Cities [42]. WorldView-2 is a very-high-resolution satellite that has an eight-band multispectral sensor with a resolution of 1.8 m/px (see Table II) and a 0.46-m/px panchromatic sensor. Similar to what we have done with the data of Sentinel-2, we only used the RGB bands.

Pleiades-1B is part of the Pleiades-1 constellation of satellites whose data are not available for the public free of charge. Nevertheless, we had access to a few rasters from which we extracted the RGB bands. The properties of this satellite are similar to those of WorldView-2 (see Table III), which makes it interesting to investigate how a model trained on WorldView-2 would perform on the data of Pleiades-1B.

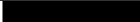
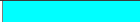





The data obtained from satellite imagery generally have a raster image form in a floating-point format. Both Sentinel-2 and WorldView-2 have a 12-b radiometric resolution encoded in a 32-b floating-point representation. These are encoded in an 8-b unsigned integer format using QGIS, an open-source geographic information system software program. The main steps consist of merging the rasters, translating the format into 8-b images, and extracting the RGB channels before normalizing them to have uniform illumination. To form the datasets, we divided the obtained RGB rasters to patches of PNG images with a resolution of 224 × 224 for Sentinel-2, 512 × 512 for WorldView-2, and 448 × 448 for Pleiades-1B. As a result, the number of images obtained from Sentinel-2 was 37 706, the number of WorldView-2 images was 3570, and the number of images from the Pleiades-1B dataset was 500, which is the least number of images obtained. The ratios for the training,

³The DeepGlobe land cover dataset was built from WorldView-3’s vivid data.

TABLE IV
WORLDVIEW-3 PROPERTIES

Bands	Bandwidth (nm)	Central wavelength (nm)	Spatial res- olution (m)
B2: Blue	60	480	1.24
B3: Green	70	545	1.24
B4: Red	60	660	1.24

TABLE V
CLASSES IN THE LABEL DATA

Class name	Class code	Color
Unknown	0	
Urban	1	
Agriculture	2	
Rangeland	3	
Forestry	4	
Water	5	
Barren	6	

validation, and test sets for all the datasets were 80%, 15%, and 5%, respectively.

Additionally, to address the possibility of having clouds in the data, we acquired a set of Sentinel-2 and WorldView-2 rasters with cloud coverage. The data of Sentinel-2 include level 1-C preprocessing, which includes cloud masks. However, the data of WorldView-2 have no GT cloud masks.

We introduced data augmentation during training as a mixture of random rotation and cropping for all the datasets, and we avoided using any band other than the RGB ones because our goal was to obtain as much compatibility with any satellite image dataset as possible. However, to test the NIR band's effect on DA, we extracted it from both Pleiades-1B and WorldView-2. Then, we applied to the NIR data the same processing as with the RGB data.

B. Labeled DeepGlobe Data

The DeepGlobe land cover segmentation dataset [7] is used for comparison with available methods. This dataset contains 1146 images with a resolution of 2000×2000 , of which only 803 are labeled. The dataset is built from WorldView-3's vivid+ images [43], and is readily available without the need for preparation. As seen in Table IV. The dataset used in this study contains 12 847 images with a size of 612×612 pixels, which we cropped from the full size, since using the full resolution images would require an excessive amount of GPU memory. The image format is an 8-b JPG with labels in PNG format. We divided the set into train, validation, and test subsets with a ratio of 70%, 20%, and 10%, respectively.

DeepGlobe dataset comes with its labels made by human annotators. The labels are the same ones shown in Table V and have been defined in [7] as follows.

- 1) Urban land: man-made, built up areas with human artifacts.
- 2) Agriculture land: farms, any planned (i.e., regular) plantation, cropland, orchards, vineyards, nurseries, and ornamental horticultural areas, confined feeding operations.
- 3) Rangeland: any nonforest, nonfarm, green land, grass.

- 4) Forest land: any land with at least 20% tree crown density plus clear cuts.
- 5) Water: rivers, oceans, lakes, wetland, ponds.
- 6) Barren land: mountain, rock, desert, beach, land with no vegetation.
- 7) Unknown: clouds and others.

C. Label Data

The label data that we used for the satellite imagery were obtained from CLC by Copernicus [4]. Generally, CLC is a manually annotated pixel-based map of Europe with five major classes divided into 44 subclasses ranging from natural covers, such as forests and water surfaces, to man-made covers, such as buildings and crops. CORINE's technical document [4] provides a full description of each subclass. Roughly, since 2000, a new version of CLC has been released every six years. Although the pixel resolution of CLC is 100 m/px, there exists a version with a resolution of 20 m/px covering the whole area of Finland [11] as well as a 10 m/px version covering Germany. For the rasters captured between 2010 and 2012, we used CLC2012, whereas for the rasters captured between 2015 and 2017, we used CLC2018.

We applied preprocessing on the labels by merging some classes to narrow them down to seven classes instead of the original 44 to match the label data in the DeepGlobe dataset (see Table V). The details of the classes are as follows.

- 1) Urban land: urban, industrial, mine.
- 2) Agriculture land: arable land, perma crops, pastures, heteroagriculture.
- 3) Rangeland: shrubs, inland wetland, artificial nonagriculture green land.
- 4) Forest land: forests.
- 5) Water: inland water and marine water.
- 6) Barren land: open spaces.
- 7) Unknown: clouds and others.

It should be pointed out that the CORINE versions that we used exhibited some inaccuracies, mainly in the German version. These inaccuracies include missing houses or small forest stands.

We aligned the label data to the same coordinate reference system as the corresponding satellite rasters. Then, we upsampled the CLC rasters to match the pixel resolution of the satellite rasters that they cover. Finally, we divided them into patches and converted them into single-channel 8-b PNG images, with each pixel's value ranging from 0 to 6, representing the corresponding class at that pixel.

D. Study Area

The study area varies depending on the satellite. Sentinel-2 and WorldView-2 cover parts of Finland and Germany, with no overlap between the areas covered by the satellites. WorldView-2 covers 1520.17 km² in Finland and 1310.18 km² in Germany (see Fig. 1). We carefully chose the rasters to avoid any cloud coverage that may compromise the training efficiency. Compared to WorldView-2, Sentinel-2 covers a far larger area in Finland of around 128 320.21 km². The area covered in Germany is also larger, at around 74 361.98 km² (see Fig. 2). More data

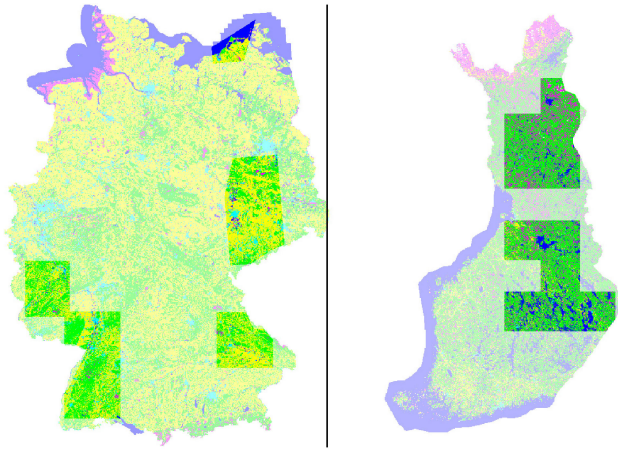


Fig. 1. Area covered by Worldview-2. Left: Area covered in Germany. Right: Area covered in Finland.

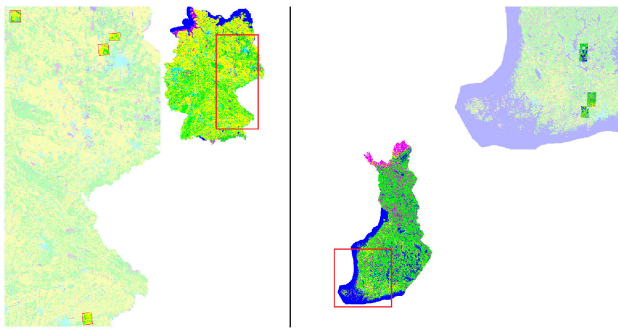


Fig. 2. Area covered by Sentinel-2. Left: Area covered in Germany. Right: Area covered in Finland.

were used from Sentinel-2 because all of its data are available free of charge, whereas only a limited amount of the WorldView-2 data is freely available. The data of Pleiades-1B cover a small area of Finland of around 519.67 km², with no overlap with the data of WorldView-2. Finally, the DeepGlobe dataset covers 1716.9 km² from India, Indonesia, and Thailand.

Finland and Germany do not share much of the land cover distribution. For instance, the tree species are quite different. Finland has more lakes and forests, whereas Germany has more urban and agricultural areas.

IV. METHOD

In this study, we used DA to semantically segment unlabeled satellite images by land cover using a different labeled dataset.

A. Network Architecture

The model used in this work is based on BDL [35], as illustrated in Fig. 3. This model is divided into two parts: a translation part \mathbf{F} , which transforms source images into target images, and the segmentation part \mathbf{F} , which assigns labels to the input images. The segmentation part is accompanied by a domain discriminator (\mathbf{D}_M) that distinguishes between the

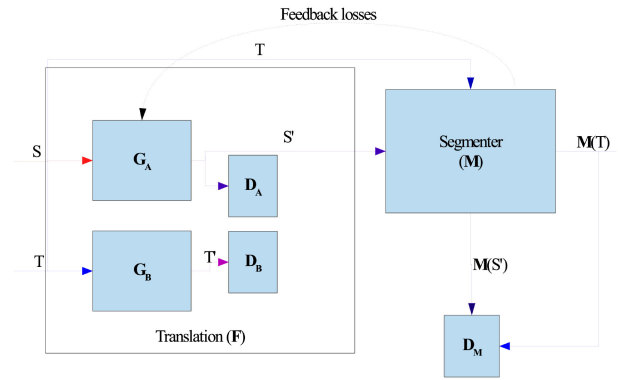


Fig. 3. BDL architecture, S represents the source data, T represents the target data, \mathbf{F} denotes the translation network with \mathbf{G}_A and \mathbf{G}_B as the generators, and \mathbf{D}_A and \mathbf{D}_B as the discriminators, \mathbf{M} represents the segmentation network, and \mathbf{D}_M is the domain discriminator.

labels generated from the target dataset and translated source dataset.

1) *Translation Network*: The translation network \mathbf{F} is a CycleGAN [24] that consists of two nine-blocks ResNet generators and two discriminators containing three fully connected layers. The model essentially contains two GANs working together: one of them translates images from the source dataset to the target dataset, whereas the other does the opposite. The link between the two networks is the cycle loss. If an image from any dataset was to go through both generators in series, it should theoretically be the same at the output.

2) *Segmentation Network*: The segmentation network \mathbf{M} is based on DeepLabV2 with ResNet101 as its backbone. The model is pretrained on the ImageNet dataset.

As mentioned earlier, the segmentation network is accompanied by a domain discriminator (\mathbf{D}_M). This network is composed of four fully connected layers, whose task is to recognize whether the labels are from the source or the target dataset. The ultimate goal here is to encourage the segmentation network to generate similar labels for both datasets and, thus, reduce the domain shift on the label space.

B. Cloud Masking Network

The cloud masking network we used for the cloud covered data is Cloud-Net [44]. This network has a U-net-like architecture that contains six convolution blocks and five deconvolution blocks (see Fig. 4). Cloud-Net is originally trained on four spectral layers from Landsat rasters (red, green, blue, and NIR).

V. EXPERIMENTS

A. Cloud Masking

We first attempted to use the pretrained Cloud-Net [44] model on the WorldView-2 data, but the results were not satisfactory. Thus, we trained the network using the data of Sentinel-2, which has GT cloud masks, and obtained a good performance (see Fig. 5). To avoid adding a new class in the label data, we merged the cloud masks with the unknown class (see Figs. 6 and 7).

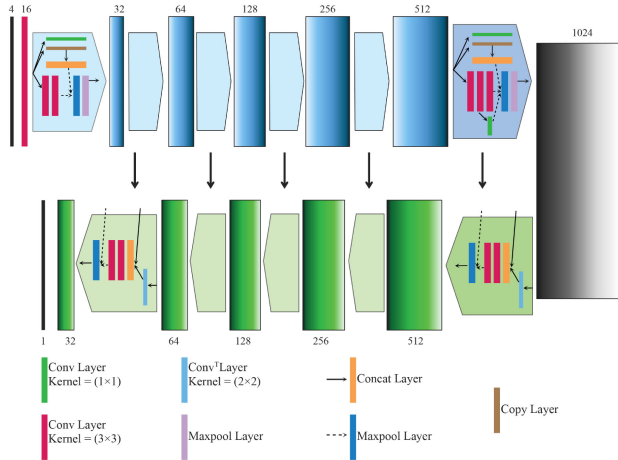


Fig. 4. Architecture of the cloud masking network used [44] (©[2019] IEEE).

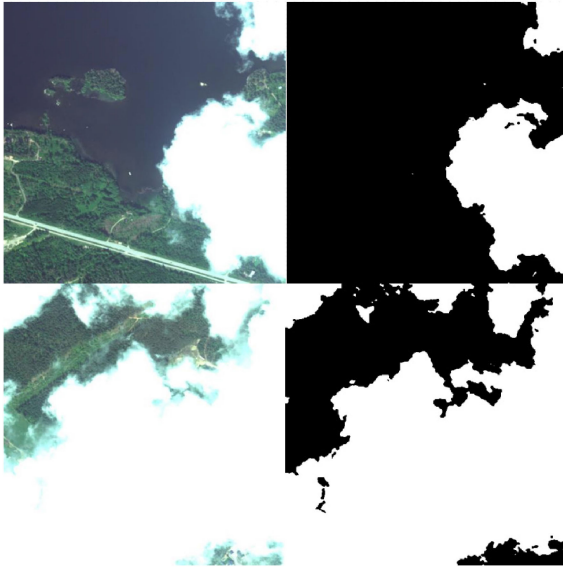


Fig. 5. Cloud masking of WorldView-2 images.

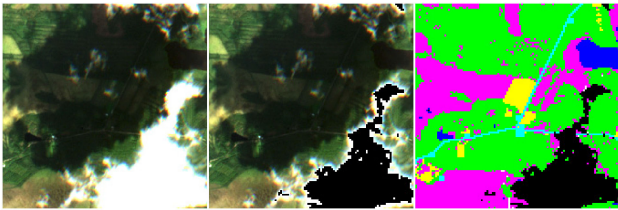


Fig. 6. Masked clouds from the GT Sentinel dataset. Left: Satellite image before cloud masking. Middle: Satellite image after cloud masking. Right: Label image after cloud masking.

B. Training

We performed the training on Nvidia GPUs (Tesla V100, Tesla P100, and Tesla T4) [45] with 16 GB of video memory for about 250 000 iterations with a batch size of 4. The batches were randomized for every iteration in the epoch. We resized



Fig. 7. Masked clouds from the WorldView-2 dataset. Left: Satellite image before cloud masking. Right: Satellite image after cloud masking.

the images to the lowest resolution between the source and the target dataset during the image translation phase.

The training of the BDL network uses the following loss functions (l_M) to train the segmentation network:

$$l_M = \lambda_{adv} l_{adv}(\mathbf{M}(S'), \mathbf{M}(T)) + l_{seg}(\mathbf{M}(S'), Y_s) \quad (1)$$

where S is the source data and T is the target data. S' or $F(S)$ is the translated source to target data. $\mathbf{M}(S')$ and $\mathbf{M}(T)$ are the prediction labels for the translated source to target data and target data, respectively. Y_s is the GT label for the source data. The adversarial loss (l_{adv}) of the domain discriminator is calculated as follows:

$$l_{adv}(\mathbf{M}(S'), \mathbf{M}(T)) = \mathbb{E}_T[\mathbf{D}_M(\mathbf{M}(T))] + \mathbb{E}_S[1 - \mathbf{D}_M(\mathbf{M}(S))] \quad (2)$$

where λ_{adv} is the coefficient for the adversarial loss. The cross-entropy loss (l_{seg}) between the GT labels and the predictions is represented as follows:

$$l_{seg}(\mathbf{M}(S'), Y_s) = -\frac{1}{HW} \sum_{H,W} \sum_{c=1}^C 1_{[c=y_s^{hw}]} \log P_S^{hw} \quad (3)$$

where C is the number of classes in the labels. H and W are the height and width of the label images, respectively. P_S is the segmentation probability of the translated image and is the output of $M(S')$ before the softmax layer.

The corresponding loss for the translation network l_F is

$$l_F = \lambda_{GAN}(\mathbb{E}[\lambda_D \mathbf{D}(T)] + \mathbb{E}[1 - \lambda_D \mathbf{D}(S')]) + \mathbb{E}[\lambda_D \mathbf{D}(T')] + \mathbb{E}[1 - \lambda_D \mathbf{D}(S)] + \lambda_{recon}[\mathbb{E}[\|\mathbf{F}^{-1}(S') - S\|_1] + \mathbb{E}[\|\mathbf{F}^{-1}(T') - T\|_1]] + \lambda_{perA} \mathbb{E}[\|\mathbf{M}(S) - \mathbf{M}(S')\|_1] + \lambda_{per_recon} \mathbb{E}[\|\mathbf{M}(\mathbf{F}^{-1}(S')) - \mathbf{M}(S)\|_1] + \lambda_{perB} \mathbb{E}[\|\mathbf{M}(T) - \mathbf{M}(T')\|_1] + \lambda_{per_recon} \mathbb{E}[\|\mathbf{M}(\mathbf{F}^{-1}(T')) - \mathbf{M}(T)\|_1] \quad (4)$$

where T' or $\mathbf{F}(T)$ is the translated target to source data. $\mathbf{M}(S)$ and $\mathbf{M}(T')$ are the prediction labels for the source data and the translated target to source data, respectively. \mathbf{F}^{-1} is the inverse function of \mathbf{F} .



Fig. 8. Results of translation from WorldView-2 to DeepGlobe. Right: Result with $\lambda_{\text{per}A} = 1$. Center: Result with $\lambda_{\text{per}A} = 0.1$. Left: WorldView-2 image.

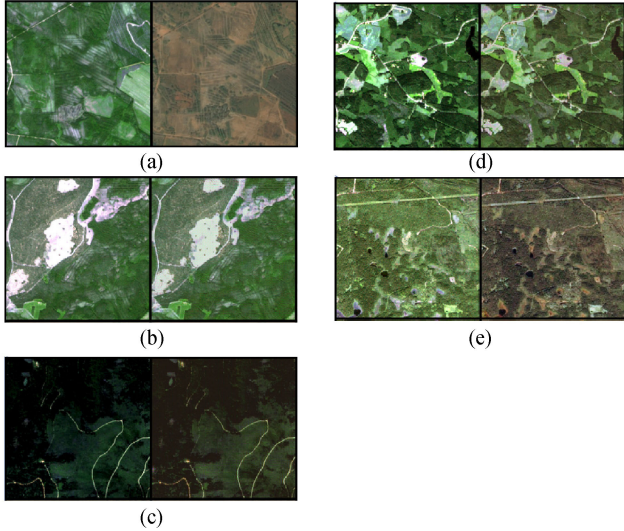


Fig. 9. Examples of multiple combinations of coefficients for the Sentinel to DeepGlobe translation test. In each example, the left image is the Sentinel source and the right image is the translation to DeepGlobe. (a) $\lambda_D = 1$, $\lambda_{\text{per}A} = 0.1$, and $\lambda_{\text{per}B} = 0.1$. (b) $\lambda_D = 1$, $\lambda_{\text{per}A} = 10$, and $\lambda_{\text{per}B} = 10$. (c) $\lambda_D = 10$, $\lambda_{\text{per}A} = 2$, and $\lambda_{\text{per}B} = 0.5$. (d) $\lambda_D = 50$, $\lambda_{\text{per}A} = 2$, and $\lambda_{\text{per}B} = 0.5$. (e) $\lambda_D = 100$, $\lambda_{\text{per}A} = 2$, and $\lambda_{\text{per}B} = 0.5$.

λ_{GAN} is the coefficient for the GAN loss. λ_D is the coefficient for the discriminator loss, whereas λ_{recon} is the coefficient for the reconstruction loss or cyclic loss. $\lambda_{\text{per}A}$ signifies the coefficient scaling of the perceptual loss of the source data, whereas $\lambda_{\text{per}B}$ is the coefficient for the target data’s perceptual loss. $\lambda_{\text{per}^{\text{recon}}}$ denotes the coefficient for the perceptual reconstruction loss. Those coefficients help guide the translation network using the segmentation network.

Notably, the coefficients presented above differentiate the original BDL network from what we have used. Fig. 8 shows an example in which we compared $\lambda_{\text{per}A} = 0.1$ with $\lambda_{\text{per}A} = 1$, where the first case resulted in trees from WorldView-2 being replaced by the barren class in the DeepGlobe domain, whereas in the second case, we obtained a more accurate translation. Each experiment required its own set of coefficients, which we obtained through trial and error. Another example is illustrated in Fig. 9.

C. Metrics

In the literature on semantic segmentation, various metrics are used to measure the accuracy compared to the GT data.

These metrics include the mean intersection over union (MIoU), average precision, pixel accuracy, and boundary F1 score. In our experiments, we used MIoU, which computes the mean of the rate of overlap between the GT segments and the resulting segmentation

$$\text{MIoU} = \frac{1}{n} \sum_{i=1}^n \frac{\text{GT}_i \cap \text{Output}_i}{\text{GT}_i \cup \text{Output}_i} \quad (5)$$

where n is the number of classes. This formula can also be written as follows:

$$\text{MIoU} = \frac{1}{n} \sum_{i=1}^n \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (6)$$

where TP stands for true positive, FP for false positive, and FN for false negative.

In our experiments, since the unknown class (seventh class) was unbalanced between the datasets, which caused the results to be skewed, we masked it while calculating the MIoU. The only exception made was to the tests that included clouds.

D. Baseline Results

Since all datasets are labeled, it is possible to know the upper-bound results to compare them with the DA ones. To obtain these results, we trained and tested the segmentation network on the target dataset. In this part, we used the datasets DeepGlobe, WorldView-2, Pleiades-1B, and WorldView-2 FI, which refers to the version of WorldView-2 that covers Finland. The results are shown in Table VI.

E. Results

The DA test from WorldView-2 (Finland and Germany) to DeepGlobe is referred to as “WV2 to DG,” whereas that from Sentinel-2 to DeepGlobe is referred to as “Sen to DG.” To test how well DA performs between satellites when the location is similar, we implemented a test between Sentinel-2 and WorldView-2, referred to as “Sen to WV2.” The next experiment performed was “WV2FI to PLFI,” which applies DA between two similar satellites (WorldView-2 and Pleiades-1B) covering the same location (Finland). The aim of the final experiment is to investigate how well DA works for different locations when the sensor used is the same. Therefore, we implemented WorldView-2 Germany to WorldView-2 Finland, also referred to as “WV2GR to WV2FI.”

To test the improvements obtained compared to when no DA was used, we performed a separate experiment with only the segmentation network enabled and ran the model on the target dataset’s validation subset. We also tested the effect of the backbone segmentation network on DA by replacing DeepLabV2 with DeepLabV3+.

Additionally, to test whether including the NIR band results in better DA across satellites, we ran a couple of tests with RGB and NIR bands.

1) *WV2 to DG*: The results of *WV2 to DG* without DA are presented in Table VII. Because the images of WorldView-2 are

TABLE VI
UPPER-BOUND MIOU RESULTS

Dataset	Urban	Agriculture	Rangeland	Forest	Water	Barren	Average
DeepGlobe	65.04	80.33	40.15	78.33	70.56	58.12	65.42
WorldView-2	52.56	69.81	32.55	77.37	93.09	8.82	55.59
WorldView-2 FI	40.14	76.01	29.36	75.36	87.79	0.48	51.59
Pleiades-1B FI	51.5	81.45	23.36	79.57	57.68	3.11	49.44
WorldView-2 FI NIR	42.83	76.77	28.43	77.27	92.35	2.24	53.32
Pleiades-1B FI NIR	51.82	82.37	26.05	79.58	51.4	6.51	49.62

TABLE VII
EXPERIMENTAL MIOU RESULTS

Experiment	Mode	Urban	Agriculture	Rangeland	Forest	Water	Barren	Average
WV2 to DG	No DA	0.02	57.55	0.0	0.01	0.56	0.0	9.69
	DeepLabV2 DA	43.99	63.81	4.99	39.76	55.79	0.0	33.89
	DeepLabV2 DA 2itr	45.03	63.65	5.12	39.54	56.21	0.0	34.93
	DeepLabV3+ DA	48.47	60.67	2.39	36.79	30.42	0.0	29.79
Sen to DG	No DA	9.29	52.19	8.02	19.66	28.74	0.11	19.62
	DeepLabV2 DA	29.78	40.42	9.73	23.3	62.67	0.58	27.47
	DeepLabV2 DA 2itr	31.27	43.18	10.66	24.11	65.24	0.36	29.13
	DeepLabV3+ DA	34.76	32.93	7.93	28.55	49.61	0.92	25.78
Sen to WV2	No DA	11.33	41.58	2.92	50.28	58.21	1.2	27.58
	DeepLabV2 DA	34.65	79.87	6.16	76.27	77.06	0.0	45.66
	DeepLabV2 DA 2itr	31.5	81.32	7.31	76.11	80.79	0.0	46.17
	DeepLabV3+ DA	36.94	80.27	9.79	75.78	76.38	0.0	46.52
WV2FI to PLFI	No DA	42.12	68.59	18.06	50.25	10.17	0.07	31.54
	DeepLabV2 DA	49.64	80.85	21.14	76.71	6.81	0.04	39.2
	DeepLabV2 DA 2itr	51.45	80.40	22.82	75.5	6.98	0.8	39.66
	DeepLabV3+ DA	49.9	79.82	21.61	76.39	5.76	0.7	39.03
WV2FI to PLFI NIR	No DA	47.41	71.62	19.43	58.56	9.29	0.06	34.56
	DeepLabV2 DA	50.93	79.61	22.67	77.82	8.94	0.18	40.03
WV2GR to WV2FI	No DA	21.47	33.51	3.24	21.74	31.51	0.0	18.57
	DeepLabV2 DA	23.9	39.91	2.68	71.07	51.84	0.0	31.57
	DeepLabV2 DA 2itr	23.79	40.02	2.55	70.64	51.88	0.0	31.48
	DeepLabV3+ DA	24.92	39.62	3.94	70.56	57.63	0.0	32.78
WV2GR to WV2FI NIR	No DA	25.58	29.83	9.93	35.83	30.96	0.0	22.02
	DeepLabV2 DA	31.06	39.32	4.11	76.12	55.32	0.09	34.37



Fig. 10. Sample images from WorldView-2 dataset and DeepGlobe dataset. Right: DeepGlobe dataset image. Left: WorldView-2 dataset image.

different from those of DeepGlobe in terms of both sensor properties and location (see Fig. 10), the results were unsatisfactory. Fig. 12 shows an example of a few test images and the model output without DA. However, the network considers everything to be agriculture, which makes it very unreliable.

Table VII shows results of using DA for *WV2 to DG*. Although the results are not very impressive numerically, there is a large difference between these results and the MIOU results without DA, ranging from less than 10% to almost 35%. Fig. 11 shows

an example of model output after DA on a few test images, manifesting very similar labeling to the GT. Additionally, in some cases, the results were better than the GT, which implies that the GT annotation is imperfect. As an example, it is unclear what is considered a forest and what is considered a rangeland. Moreover, some small villages have been completely ignored in the GT while parts of them have not. This can be observed in Fig. 12.

2) *Sen to DG*: Unlike *WV2 to DG*, in which WorldView-2 has similar properties to those of WorldView-3, *Sen to DG* attempts to perform DA between two very different satellites.

It should be noted that the results obtained without DA are not reliable, although they are better than those obtained with *WV2 to DG*, since Sentinel-2 has considerably more data. However, even a massive dataset is not enough to obtain acceptable results, as illustrated in Fig. 11.

Using DA improved the results from an MIOU of 19% up to 29%, as shown in Table VII. When compared to the upper-bound result on DeepGlobe, which is 52.24%, the DA results were found to be modest. However, they were visually still good, as illustrated in Fig. 11 with a few examples. However, in contrast to *WV2 to DG*, the results were not as good. This can be explained by the large pixel resolution difference between

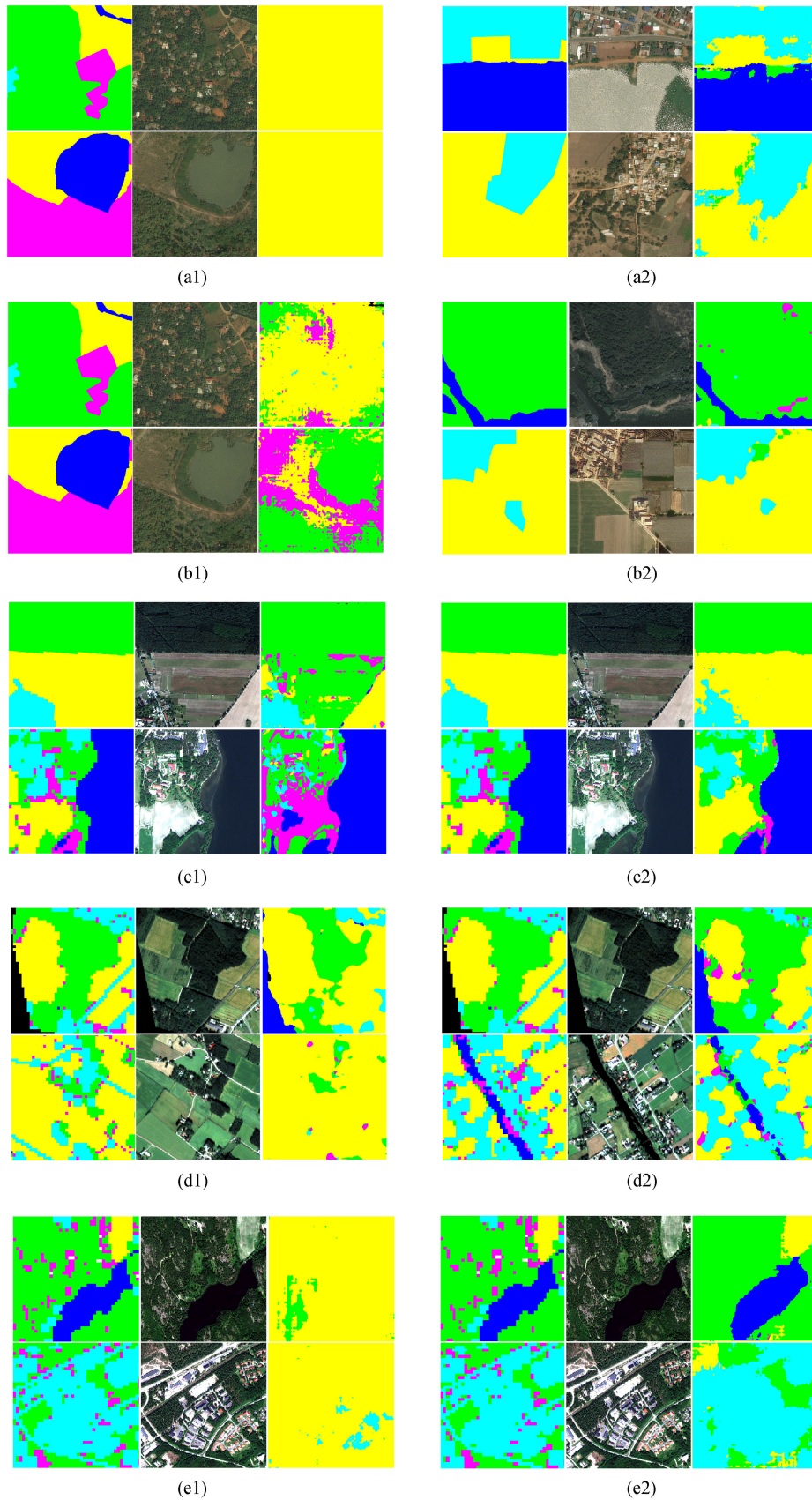


Fig. 11. Experimental results. Left: GT image. Middle: Satellite image from the target dataset. Right: Model output. (a)-1) WV2 to DG without DA. (a)-2) WV2 to DG with DA. (b)-1) Sen to DG without DA. (b)-2) Sen to DG with DA. (c)-1) Sen to WV2 without DA. (c)-2) Sen to WV2 with DA. (d)-1) WV2FI to PLFI without DA. (d)-2) WV2FI to PLFI with DA. (e)-1) WV2GR to WV2FI without DA. (e)-2) WV2GR to WV2FI with DA.

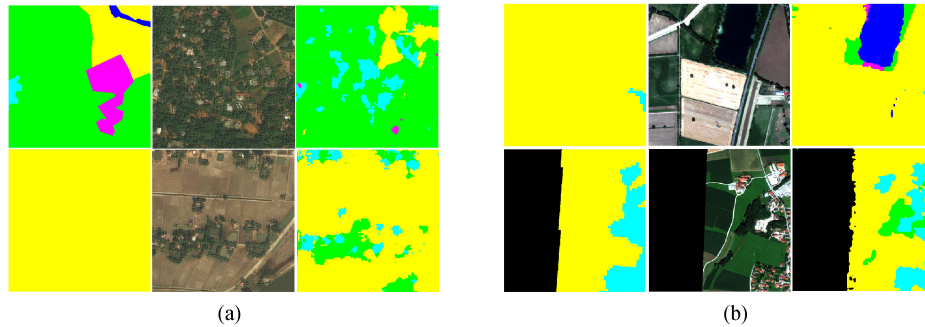


Fig. 12. Experimental results in which the output improves upon the GT. Left: GT image, middle: satellite image from target dataset, and right: model output. (a) WV2 to DG with DA. (b) Sen to WV2 with DA.

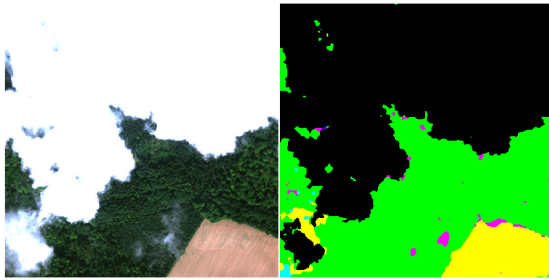


Fig. 13. Results of training with DA from Sentinel-2 with clouds masked to WorldView-2 with clouds masked. Right: Model output. Left: Test images from the WorldView-2 dataset.

Sentinel-2 and WorldView-3. However, this is still considered a good step forward as the Sentinel-2 data are free of charge, whereas the WorldView-3 data are not.

3) *Sen to WV2*: Table VII shows the results of *Sen to WV2*. Samples from the results can be seen in Fig. 11. Inaccuracies are present in the results without DA, particularly mixing up similarly looking classes, such as forestry and agriculture. As mentioned in Section III-C, the GT labels for Germany lack precision, which limits the accuracy of the mapping even when training on that specific dataset. However, having more accurate labels covering Finland helps correct the errors in the German ones, yielding better results than those obtained with the GT in a few cases, as depicted in Fig. 12.

There are no GT that could mask for the WorldView-2 images in the cloud masking test, meaning it is not possible to have a correct MIoU value. However, by overlaying the cloud masks generated by the network in Section IV-B on the GT CLC labels, we obtained an MIoU of 43.7%. An example of the results is shown in Fig. 13.

4) *WV2FI to PLFI*: In the *WV2FI to PLFI* experiment, even though the sensors in the satellites were similar, the results obtained without using DA were not as good as expected. Fig. 11 shows an example of the results of *WV2FI to PLFI* without DA.

Training a DNN on a limited amount of data, such as in the case with the Pleiades-1B dataset, would lead to overfitting. Therefore, applying DA is an excellent way to perform land cover mapping on a small dataset. *WV2FI to PLFI* shows better results with an increase of over 8% in MIoU, as illustrated in Fig. 11, which is considered encouraging given that it would be

costly to perform training on the Pleiades-1B data. The detailed results are presented in Table VII.

5) *WV2GR to WV2FI*: The results of *WV2GR to WV2FI* without DA were surprisingly weaker than expected considering the images are captured from the same satellite. Table VII shows an MIoU of only 18% MIoU. A sample of those results is shown in Fig. 11.

The results obtained from *WV2GR to WV2FI* using DA were significantly better compared to the previous case. As seen in Table VII, the MIoU score increased to 32%. Fig. 11 shows a sample from this experiment.

6) *NIR Results*: As expected in both experiments, in which we included the NIR band, the results that we obtained were better than those obtained without including an NIR band. However, the improvement was minor. The MIoU without DA improved upon the experiment in which we considered RGB bands only between 3% to 4% MIoU, as seen in Table VII. The same effect applies to the results obtained after DA, in which the improvement was also minuscule. These results were somewhat unexpected but also understandable. While the NIR band is essential for detecting vegetation, the fact that we merged many vegetation species into the same class made the best feature of NIR potentially useless. Moreover, because the model was very deep, it could extract enough features from the RGB bands only.

7) *DeepLabV3+ Results*: In addition to the previous experiments, we tested DeepLabV3+ as a segmentation network. However, as shown in Table VII, replacing the backbone with DeepLabV3+ did not bring about significant improvements as compared to DeepLabV2. This is because of the lack of accurate labels, which may be a bottleneck for the performance. In fact, the results diverged after a few epochs because the network learned to mimic the errors in the GT images. Thus, to make use of the full potential of DL methods, it is important to have very good and precise labels in land cover mapping.

8) *Comparison With Other Methods*: Since our work mostly uses datasets that we have built, it is tricky to compare it with other works. However, using Potsdam [51] and Vaihingen [52] datasets, we can have an idea on how well it performs. Table VIII presents results of DA from other works as well as ours. The other works' results were taken from the original papers. With the right coefficients, our results are very good in comparison with the other works.

TABLE VIII
MIOU RESULTS FOR DA FROM POTSDAM TO VAIHINGEN

Method	Impervious surface	Building	Low vegetation	Tree	Car	Clutter/background	Average
Benjdira's [40]	53.2	59.1	15.5	29.7	30.3	4.8	32.2
CyCADA [46]	55.0	67.2	22.7	32.4	36.2	4.1	36.3
CBST [47]	59.8	71.6	24.8	31.2	36.7	8.6	38.8
Deng's [48]	NA	NA	NA	NA	NA	NA	40.04
BDL(vanilla)	61.7	70.3	27.0	40.5	38.1	5.9	40.6
CLAN [49]	64.2	75.1	23.5	43.9	42.3	6.7	42.6
FSDAN [50]	57.4	57.8	41.7	58.4	37.0	10.0	43.7
Ours	60.36	75.81	31.38	42.51	34.0	20.09	44.03

Note: NA refers to not available.

The best results of each section are shown in bold characters.

F. Discussion

The general consensus regarding our experiments is that DA can either slightly or significantly improve the accuracy of land cover mapping. We found that the more data we have, the closer our results are to the baseline, as shown in the *Sen to WV2* experiment. In other cases in which the area covered during inference is significantly different from the area used during training, such as in the *Sen to DG* experiment, the results were relatively far from the baseline. However, a clear improvement was observed, which is impressive since the land cover types were quite different. Interestingly, the NIR band did not improve the results significantly when compared to DA. As mentioned previously, this may be because the vegetation classes are not diverse, which limits the benefits of using the NIR band.

VI. CONCLUSION

In this study, we addressed three problems related to LULC mapping: lack of labeled datasets, inaccessibility to some satellite imagery, and differences in the available spectral bands. The approach that we adopted employs DA on datasets that we have built from the RGB imagery of different satellites covering different areas. The experimental results showed that DA improved the mapping considerably even when only RGB images were used. Such results are generalizable to images from areas with considerably different land cover types. From our experimental results, we believe that the overall results can be improved by designing a more specialized model for satellite imagery.

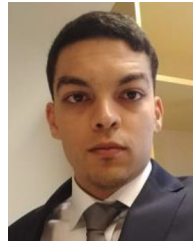
ACKNOWLEDGMENT

The authors would like to thank the project partners from VTT Technical Research Centre of Finland for providing their datasets and technical support.

REFERENCES

- [1] Copernicus, ESA, Copernicus open access hub. Accessed: Apr. 2020. [Online]. Available: <https://scihub.copernicus.eu/dhus/>
- [2] U.S. Department of the Interior, USGS. Accessed: Apr. 2020. [Online]. Available: <https://earthexplorer.usgs.gov/>
- [3] W. Ahmad, L. B. Jupp, and M. Nunez, "Land cover mapping in a rugged terrain area using Landsat MSS data," *Int. J. Remote Sens.*, vol. 13, no. 4, pp. 673–683, 1992. [Online]. Available: <https://doi.org/10.1080/01431169208904145>
- [4] B. Kosztra, G. Büttner, G. Hazeu, and S. Arnold, "Updated CLC illustrated nomenclature guidelines," Accessed: Jul. 2020. [Online]. Available: https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/docs/pdf/CLC2018_Nomenclature_illustrated_guide_20170930.pdf
- [5] NASA, "MODIS Land Cover Type/Dynamics," Accessed: Mar. 2020. [Online]. Available: <https://modis.gsfc.nasa.gov/data/dataproduct/mod12.php>
- [6] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 44–51.
- [7] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, pp. 172–181.
- [8] C. Tian, C. Li, and J. Shi, "Dense fusion classmate network for land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 262–2624.
- [9] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
- [10] T.-S. Kuo, K.-S. Tseng, J. Yan, Y.-C. Liu, and Y.-C. F. Wang, "Deep aggregation net for land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 247–2474.
- [11] Suomen Ympäristökeskus, "Metatietopalvelu," Accessed: May 2020. [Online]. Available: <http://metatieto.ymparisto.fi:8080/geoportal/catalog/search/resource/details.page?uuid=%7B6833C06E-BF77-4F0B-A066-B94AE98392EA%7D>
- [12] Bundesamt für Kartographie und Geodäsie, "BKG—Corine Land Cover," Accessed: May 2020. [Online]. Available: <https://www.bkg.bund.de/DE/Ueber-das-BKG/Geoinformation/Fernerkundung/Landbedeckungsmodell/CorineLandCover/clc.html>
- [13] A. Rosenfeld and L. S. Davis, "Image segmentation and image models," *Proc. IEEE*, vol. 67, no. 5, pp. 764–772, May 1979.
- [14] Trimble Geospatial, "What is eCognition?," Accessed: Feb. 2020. [Online]. Available: <https://geospatial.trimble.com/what-is-ecognition>
- [15] M. S. Tehrani, B. Pradhan, and M. N. Jebuv, "A comparative assessment between object and pixel-based classification approaches for land use/land cover mapping using SPOT 5 imagery," *Geocarto Int.*, vol. 29, no. 4, pp. 351–369, 2014.
- [16] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sens. Environ.*, vol. 177, pp. 89–100, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425716300578>
- [17] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision - ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 833–851.
- [18] H. A. Arief, G.-H. Strand, H. Tveite, and U. G. Indahl, "Land cover segmentation of airborne LiDAR data using stochastic atrous network," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 973. [Online]. Available: <http://www.mdpi.com/2072-4292/10/6/973>
- [19] W. Emery, *Visible and Infrared Remote Sensing*, vol. 12. Hoboken, NJ, USA: Wiley, 1999.
- [20] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231218306684>

- [21] G. Wilson and D. J. Cook, "Adversarial transfer learning," 2018, *arXiv:1812.02849*.
- [22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [23] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.
- [24] J. 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. 2242–2251.
- [25] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2039–2049.
- [26] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*.
- [27] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2011–2020.
- [28] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1335–1344.
- [29] W. Chang, H. Wang, W. Peng, and W. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1900–1909.
- [30] Y. Tsai, W. Hung, S. Schuler, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.
- [31] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2512–2521.
- [32] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
- [33] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 102–118.
- [34] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3234–3243.
- [35] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6929–6938.
- [36] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.
- [37] E. Izquierdo-Verdiguier, V. Laparra, L. Gómez-Chova, and G. Camps-Valls, "Encoding invariances in remote sensing image classification with SVM," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 981–985, Sep. 2013.
- [38] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, Sep. 2009.
- [39] S. Inamdar, F. Bovolo, L. Bruzzone, and S. Chaudhuri, "Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1243–1252, Apr. 2008.
- [40] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1369. [Online]. Available: <https://www.mdpi.com/2072-4292/11/11/1369>
- [41] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised Domain adaptation for semantic segmentation using color mapping generative adversarial networks," Working Paper, Nov. 2019. [Online]. Available: <https://hal.inria.fr/hal-02382155>
- [42] Suomen ympäristökeskus, "WorldView-2 European cities—View data product—Earth online—ESA," Accessed: Nov. 2019. [Online]. Available: <https://earth.esa.int/web/guest/-/worldview-2-european-cities-dataset>
- [43] DigitalGlobe, "DigitalGlobe Basemap +Vivid," Accessed: Dec. 2019. [Online]. Available: https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/2/DG_Basemap_Vivid_DS_1.pdf
- [44] S. Mohajerani and P. Saeedi, "Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 1029–1032.
- [45] NVIDIA Corporation, "NVIDIA Tesla V100 Data Center GPU," Accessed: Oct. 2019. [Online]. Available: <https://www.nvidia.com/en-us/data-center/tesla-v100/>
- [46] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn. Res.*, 2018, pp. 1989–1998. [Online]. Available: <http://proceedings.mlr.press/v80/hoffman18a.html>
- [47] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 297–313.
- [48] X. Deng, Y. Zhu, Y.-X. Tian, and S. Newsam, "Generalizing deep models for overhead image segmentation through Getis-Ord G_i^* pooling," in *Proc. 11th Int. Conf. Geographic Inf. Sci. (GIScience 2021) - Part I, ser. Leibniz Int. Proc. Informatics (LIPIcs)*, vol. 177. Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020, pp. 3:1–3:14. [Online]. Available: <https://drops.dagstuhl.de/opus/volltexte/2020/13038>
- [49] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2502–2511.
- [50] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–13, 2020.
- [51] The International Society for Photogrammetry and Remote Sensings, "2D semantic label—Potsdam," Accessed: Oct. 2020. [Online]. Available: <https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/>
- [52] The International Society for Photogrammetry and Remote Sensings, "2D semantic label—Vaihingen," Accessed: Oct. 2020. [Online]. Available: <https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/>



Nadir Bengana received the B.S. degree in electrical and electronics engineering and the M.S. degree in computer engineering from Boumerdes University, Boumerdes, Algeria, in 2015 and 2017, respectively, and the M.S. degree in computer science and engineering, in 2019 from the University of Oulu, Oulu, Finland, where he is currently working toward the Sc.D. degree in computer science.

Since 2018, he has been working as a Researcher in the Center for Machine Vision and Signal Analysis, University of Oulu.



Janne Heikkilä (Senior Member, IEEE) received the Doctor of Science in Technology degree in information engineering from the University of Oulu, Oulu, Finland, in 1998.

He is currently a Professor of computer vision and digital video processing with the Faculty of Information Technology and Electrical Engineering, University of Oulu, and the Head of the Degree Program in computer science and engineering. He has supervised nine completed doctoral dissertations and authored/coauthored more than 160 peer-reviewed

scientific articles in international journals and conferences. His research interests include computer vision, machine learning, digital image and video processing, and biomedical image analysis.

Prof. Heikkilä has served as an Area Chair and a member of program and organizing committees of several international conferences. He is a Senior Editor for the *Journal of Electronic Imaging*, an Associate Editor for the *IET Computer Vision and Electronic Letters on Computer Vision and Image Processing*, a Guest Editor for a special issue in *Multimedia Tools and Applications*, and a member of the Governing Board of the International Association for Pattern Recognition. During 2006–2009, he was the President of the Pattern Recognition Society of Finland. He has been the Principal Investigator in numerous research projects funded by the Academy of Finland and the National Agency for Technology and Innovation.