Learning domain invariant representations of heterogeneous image data

Mihailo Obrenović[1,2*], Thomas Lampert[1], Miloš Ivanović[2] and Pierre Gançarski[1]

[1]ICube, University of Strasbourg, Strasbourg, France.
[2]University of Kragujevac, Faculty of Science, Radoja Domanovica 12, Kragujevac, 34000, Serbia.

*Corresponding author(s). E-mail(s): mobrenovic@unistra.fr;
Contributing authors: lampert@unistra.fr; mivanovic@kg.ac.rs; gancarski@unistra.fr;

Abstract

Supervised deep learning requires a huge amount of reference data, which is often difficult and expensive to obtain. Domain adaptation helps with this problem — labelled data from one dataset should help in learning on another unlabelled or scarcely labelled dataset. In remote sensing, where variety of sensors are producing images of different modalities and with a different number of channels, it would be very beneficial to develop heterogeneous domain adaptation methods being able to work with domains coming from a different input space. However, this challenging problem is rarely addressed, majority of existing works does not use image-data, or they rely on translation from one domain to the other, completely ignoring domain-invariant feature extraction approach. In this paper, we propose novel approaches for heterogeneous image domain adaptation for both semi-supervised and unsupervised setting, based on extracting domain invariant features and deep adversarial learning. For unsupervised domain adaptation case, impact of pseudo-labelling is also investigated. We evaluate on two heterogeneous remote sensing datasets, one being RGB, and the other multispectral, for the task of land-cover patch classification; but also on a standard computer vision benchmark of RGB-depth map adaptation. The results show that our domain invariant approach consistently outperforms the competing method based on image-to-image translation, and that our

method is not limited to remote sensing only, but is more general and can also successfully be applied to standard computer vision problems.

Keywords: Domain Adaptation, Remote Sensing, Deep Learning, Representation Learning

1 Introduction

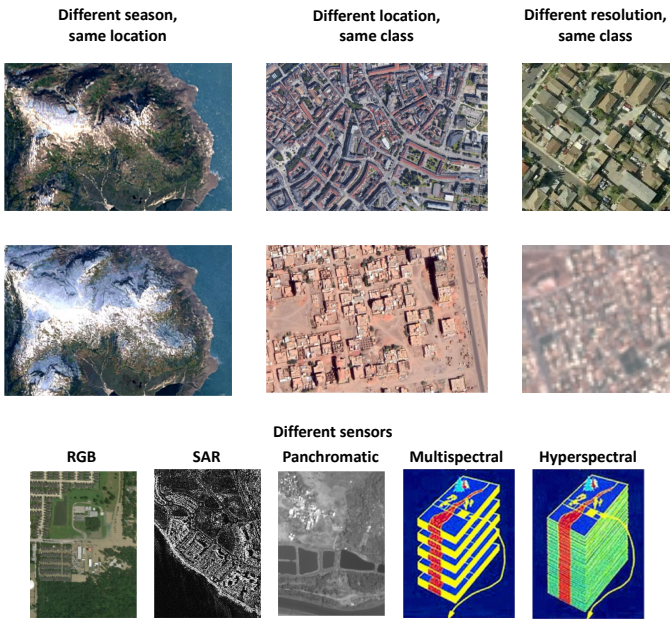


Fig. 1: Causes of domain shift in remote sensing — different seasons, location, resolution, and sensors.

Supervised deep learning (DL) methods rely heavily on the existence of large-scale labelled datasets. However, reference data is often difficult to obtain. This is especially true in the field of *remote sensing* (RS) where:

- Satellites generate a huge amount of data on a daily basis, and since labelling is a manual process, it is slow and expensive.
- The Earth's surface is constantly evolving, meaning that reference data may not be reusable for images taken at a later date.

Since DL models (and machine learning methods in general) often generalise poorly, we cannot apply existing trained models to other datasets.

To overcome this problem, focus has turned to *domain adaptation* (DA) techniques.

Some causes of domain shift in RS are shown in Figure 1: images are captured at different seasons or different places, from different height and/or with different sensors; what is vegetation in one season can be covered with snow in the other; urban areas can look very differently on different continents; depending on the height of the airborne sensor, the same objects can look bigger or smaller, in RS this is referred to as images having different spatial resolution; different sensors can capture images of different modalities, with non-corresponding channels (bands in RS), and possibly with a different number of bands. All of these cases of shift lead to very *different data distributions* between datasets.

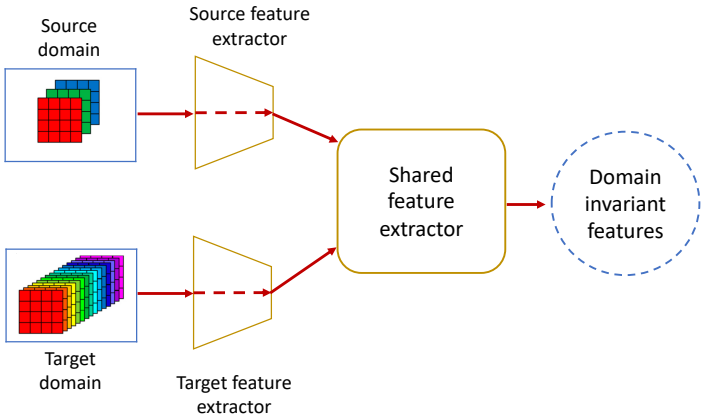


Fig. 2: The proposed approach that extracts domain invariant features and can work with images of different number of channels and/or different resolution

Domain adaptation involves learning a model on one data distribution (named *source* - typically labelled), and applying it to another, different but related data distribution (called *target* - typically with little or no reference data) by reducing the shift between domains. Most of the DA methods in computer vision (CV) assume *RGB images in both domains* (homogeneous DA). In remote sensing, such methods might be enough to solve temporal or geographical shifts. However, we can have the domains that do not lay in the same space and which possibly have a different dimensionality (heterogeneous DA). We differentiate two such cases (which can also occur simultaneously):

- Domains having different spatial resolution,
- Domains having different channels.

In the case of images having different spatial resolution, theoretically all the images could be resampled to the same resolution and homogeneous DA

used. However, it should be noted that in remote sensing resolution can be extremely different, e.g. in datasets used in this paper for evaluation, a pixel can represent the square with 10 m side of Earth surface (EuroSAT [1]), or the square with 20 cm side (RESISC45 [2]), meaning that a car can be blurred into a road background or easily visible spanning several pixels. Thus even though homogeneous DA methods could be used here, this is a very challenging problem.

However, in case of images captured by different sensors, having a *different number of channels*, homogeneous DL domain adaptation approaches cannot be applied at all because their structure (number of input neurons) is fixed, preventing images of different dimensionality being used within the pipeline.

In this paper we propose a novel semi-supervised heterogeneous domain adaptation (HDA) approach for image patch classification called SS-HIDA, and its counterpart for unsupervised HDA, an approach based on pseudo-labelling called UPL-HIDA. To the best of our knowledge this is *the first HDA method for patch classification capable of working with two image-data domains with a different number of bands*. It is also the first work on *extracting domain invariant features* from two heterogeneous image-data domains. The schema of the approach is shown in Figure 2.

Existing work on different modalities in the RS community have focused on data fusion [3] where different domains have corresponding paired images. However, in the proposed work, such a constraint does not exist, and therefore *datasets with completely independent, unpaired images* can be used, possibly taken from different parts of the world therefore dealing with multiple sources of domain shift simultaneously.

The development of models like ours could be very beneficial for the RS community where a variety of different sensors are used, some of them being RGB, multispectral, hyperspectral, SAR, LiDAR, panchromatic etc. But the field of application is not limited to RS only, we can find different sensors in robotics (depth images), in medical imaging (e.g. CT and MRI) etc. We evaluate our method on one very challenging RS case (RGB and multispectral images of different resolution), and on one common CV benchmark (RGB and depth images).

This article is organised as follows: in Section 2, a review of related existing work is given, followed by a description of the proposed SS-HIDA and UPL-HIDA architectures in Section 3. Section 4 describes the experimental setup and results. Finally, the conclusions are given and future work is discussed in Section 5.

2 Literature Review

Domain adaptation methods in computer vision can be split into two big categories:

- Methods based on extraction of domain invariant features
- Methods based on translation of data

DA	homogeneous	heterogeneous				
		vectorial	image-text	raw images same number of channels	different number of channels	
					SS	patch
domain invariant	DANN	Li	Shu - DTN			**SS-HIDA (ours)**
	WDGRL	Wang - Autoencoders	Chen			**UPL-HIDA (ours)**
	DSN	Chen	Li			
translation	OT for DA	GW		ADDA	Benjdira 2020	
	JDOT	COOT		Benjdira 2019		
	Deep JDOT	TNT		CycleGAN for HDA*		
	ADDA					
	CyCADA					
	SemI2I*					

Fig. 3: Overview of existing methods and where our approach fits in the categorisation

In homogeneous DA both these approaches are widely used, and the methods are used on raw image data. On the other hand, heterogeneous DA methods for CV focus primarily on adapting between vectorial data of different size, such as SURF - DeCAF [4, 5] and DeCAF - ImageNet features [6], or on adapting from image to text data [7, 8]. There are some HDA methods capable of working with raw image data, but they assume the same number of channels in the domains [9–11], or they are designed for semantic segmentation only, and not for patch classification [12]. All of these methods for raw image data are based on the translation of data from one domain to the other, while our approach explores the potential of using methods for extracting domain invariant features in HDA. Table 3 shows the overview of existing methods and where do the models proposed in this paper, semi-supervied SS-HIDA and unsupervised UPL-HIDA, fit in the categorisation. As can be seen, our models are the only HDA domain invariant methods capable of working on raw image data, and the only method for patch classification capable of working with images with different number of channels.

The rest of this section describes existing related methods, starting from the domain-invariant methods for both homogeneous and heterogeneous DA, then describing translation methods, and finally showing the applications in remote sensing. The work on semi-supervised DA is also briefly mentioned.

2.1 Domain-Invariant Feature Extraction

2.1.1 Homogeneous Domain Adaptation

The emergence of Generative Adversarial Networks (GANs) [13, 14] inspired numerous domain adaptation techniques for computer vision [15, 16]. The idea of making real and fake data indistinguishable is naturally extended to DA where two domains should be brought to the same space.

GANs have two main components trained through an adversarial game — a generator to generate realistically looking images; and a discriminator to distinguish between real and generated images. Ganin et al. [15, 17] were one of the first to use the principle of adversarial learning in domain adaptation.

Instead of a generator and a discriminator, their method DANN (Domain-Adversarial Neural Network) uses a *feature extractor* and a *domain classifier*. While the feature extractor has the task of extracting **domain invariant** features from two domains, the domain classifier is trained to predict to which domain the extracted features belong.

The original GAN minimises the Jensen-Shannon (J-S) divergence between the spaces of real and generated images [13]. However, this formulation suffered from problems such as unstable training, mode collapse etc. These issues were addressed with the introduction of Wasserstein GAN [14] by minimising the Wasserstein distance instead of J-S divergence. Calculating this metric is computationally expensive, so it is approximated with a *domain critic*, a neural network that replaces the discriminator in the original GAN. The improvements of Wasserstein GAN found their application in domain adaptation with the WDGRL (Wasserstein Distance Guided Representation Learning) method [16]. WDGRL is similar to DANN in that it also extracts domain invariant features from two domains but instead of a domain classifier, WDGRL uses a domain critic.

2.1.2 Heterogeneous Domain Adaptation

A large majority of DA methods for CV are concerned with RGB domains [15, 16, 18]. As for heterogeneous DA methods for CV based on domain-invariant feature extraction, most of the work has been studied in domains having different features, e.g. SURF and DeCAF features [4, 5, 8], but not on the raw image data. Other than image-image DA, the problems of image-text DA [4, 7, 8], and text-text DA [4, 5] have also been tackled.

To be capable of working with inputs of different sizes, non-DL methods used matrix projection to project the original inputs to a common space of the same size [4, 19]. On the other hand, DL methods mostly use two separate input branches for this purpose.

Deep Transfer Network (DTN) [7] has a number of weakly shared layers after separate input branches. It is specifically designed for transferring from textual to image data. DTN assumes that there is a co-occurrence set of paired text and image data, which simplifies the discovery of relations between domains. DTN also needs at least a small amount of labelled target data, and it is therefore not applicable in UDA.

Another DL based approach proposed by Wang et al. [5], uses autoencoders to project the inputs to a space of the same size. Domain divergence is reduced by minimising MMD distance. Local structure of the data is preserved using a manifold alignment term that keeps neighbouring same-class samples close in the feature space. As such, this method requires a small amount of labelled data in the target domain. It does not, however, require a co-occurence set of paired data. Pseudo-labels are also employed to augment the number of target labels.

Our methods, SS-HIDA and UPL-HIDA, also have separate input branches which bring the inputs to a space of the same size. These are then followed by

the shared layers that extract domain invariant features. SS-HIDA and UPL-HIDA can work on raw input images and do not require paired data between domains.

2.2 Translation methods

2.2.1 Optimal transport

The idea of optimal transport theory is to 'transport' the source distribution into the target distribution's space, which reduces the Wasserstein distance between the distributions to zero. Optimal transport can be used in DA by training the classifier on the transported labelled source data in target domain space [20]. The transport of the joint feature/label distribution is also considered [21]. This approach is very useful in making a neural network robust to noisy labels [22]. It can also be combined with deep learning [23], in which optimal transport is performed on the feature representations from the layers of a convolutional neural network.

Optimal transport also found its application in HDA by using Gromov-Wasserstein (GW) distance to find the transport plan [24]. GW distance is convenient for heterogeneous domains, as it is only based on preserving the distance between samples of the same domain when transporting, thus domains can have different dimensionality. Redko et al. [6] improve upon GW distance by introducing co-optimal transport and the COOT measure, which not only takes into account the correspondence between samples, but also relations between features. Though very promising for HDA, these methods are shallow and lack the abstraction that can be provided by DL.

2.2.2 Feature translation

Transfer Neural Trees (TNT) [8] is a DL method for HDA based on feature translation. It consists of two input branches, which are followed by a Neural Decision Forest for label prediction. The source input branch and decision forest are first trained (on source data). They are then fixed, and the target branch is trained to adjust the extracted target features to the already trained classifier. TNT is evaluated for image-image adaptation but, as with previous methods, DeCAF and SURF features were used.

The training algorithm of Adversarial Discriminative Domain Adaptation (ADDA) [9] is very similar to TNT, except that a softmax layer is used as a classifier instead of a neural forest, and that the domain difference is reduced in an adversarial manner by using a domain classifier. ADDA is evaluated on different modalities of raw image data (RGB and depth images), but is primarily designed for homogeneous data.

2.2.3 Image-to-Image Translation

Perhaps the most promising of the existing methods for image based HDA are Image-to-Image translation GANs [25, 26], which can translate images

from one domain to another. One of the most famous architectures is CycleGAN [25]. Unlike many similar methods, CycleGAN does not require matched pairs between domains. The network consists of two generators and two discriminators and is trained to translate both from the source domain to the target, and vice-versa. The success of CycleGAN lies on its cycle consistency loss — images translated from one domain to the other have to be correctly translated back, thus the structure and salient information is preserved during translation. It is worth noting that technically, CycleGAN can work with heterogeneous datasets of different dimensionality, but in that case one of the terms in CycleGAN loss function—identity loss—needs to be removed.

Hoffman et al. successfully applied CycleGAN to DA in RGB images [27]. Their method CyCADA performs both pixel level and feature level adaptation. However, the semantic loss introduced prevents this method from being used in HDA.

2.3 Remote sensing

Transfer learning in RS is a much more difficult task when compared to CV. For many CV object classification datasets, models pre-trained on ImageNet give transferable features. However, the equivalent large-scale curated datasets in RS are only beginning to exist [28, 29]. Neumann et al. gave an interesting study [30] on transfer learning across multiple remote sensing datasets. The authors simply pre-trained a model on a source dataset and fine-tuned it on target data, this achieved competitive results. Nevertheless, the fact that heterogeneous data (i.e. data with > 3 and/or non-RGB bands) exists would prevent them from being applied in the same manner as in CV.

Tasar et al. propose an image-to-image DA architecture named SemI2I for the task of semantic segmentation [31]. Images translated from the source space are used to train a semantic segmentation model in target space. SemI2I uses the notion of cycle consistency as in CycleGAN, but is based on autoencoders and cross reconstruction. The whole approach is evaluated on two RGB domains with images from different cities.

CycleGAN and Conditional GANs have been used for translating between optical and SAR images [32–34]. Deep features can then be used from a SAR-to-optical generator to perform semantic segmentation [33] and in change detection [34]. Though all of these works agree that the problem of optical-SAR translation is suboptimal and ill-posed, it turns out to be very useful as a proxy task, as it helps with the global understanding of the image, and provides meaningful semantic features for differentiating between land cover classes.

There have been works on DA for semantic segmentation of land cover maps using data from different sensors in different domains [10, 12], but in one case, though the bands may be different, their number still has to be the same [10], while in the other case, labelled segmentation masks are needed in the target domain, and these (segmentation masks) are used as an intermediate

space during the translation from the target domain to the source domain [12]. This approach therefore does not extend to classification.

Voreiter et al. propose the most similar method to that presented herein [11], the authors use a variant of CycleGAN — they add to it a metric loss, a classification loss, and a super-resolution capability. The method is applied to two remote sensing datasets of different resolutions. The method can be used for both unsupervised and semi-supervised HDA.

2.4 Semi-Supervised Domain Adaptation

In the literature, unsupervised domain adaptation (UDA) is addressed more often than semi-supervised domain adaptation (SSDA). However it was shown that existing UDA methods do not scale well to the semi-supervised setting, and that a method that specifically aims to use the fact that few target labels exist easily outperforms UDA methods [35]. This shows that there is a need for methods specifically tailored for SSDA, which motivated us to have separate variants of our method for SSDA and UDA.

2.5 Our Contribution

Most of the existing HDA methods are based on the idea of translating data from one domain to the other, either in pixel space using image-to-image methods [25, 26], or in feature space, e.g. ADDA [9]. When trained in this manner, however, the resulting models are only applicable to the target domain. They are therefore bound to either simplify or invent the difference between domains during the translation, since the target data distribution must be made to match the source's distribution. Instead, we propose a method that extracts domain invariant features. The extracted features are neither in the source, nor target data space, but in a learnt common latent space. The hypothesis being that this will allow the model to enhance the latent representation using information from both domains. Our method is inspired by homogeneous DA methods such as DANN [15], WDGRL [16], and DSN [18] which also extract domain invariant features, but are limited to working with homogeneous domains only.

3 Methodology

In this section we will present two models, one for semi-supervised heterogenous domain adaptation (SS-HIDA), and another for unsupervised heterogeneous domain adaptation (UPL-HIDA).

3.1 Semi-Supervised Heterogeneous Image Domain Adaptation (SS-HIDA)

We extend the homogeneous, unsupervised domain adaptation approach Wasserstein Distance Guided Representation Learning (WDGRL) [16] to the case of heterogeneous image data.

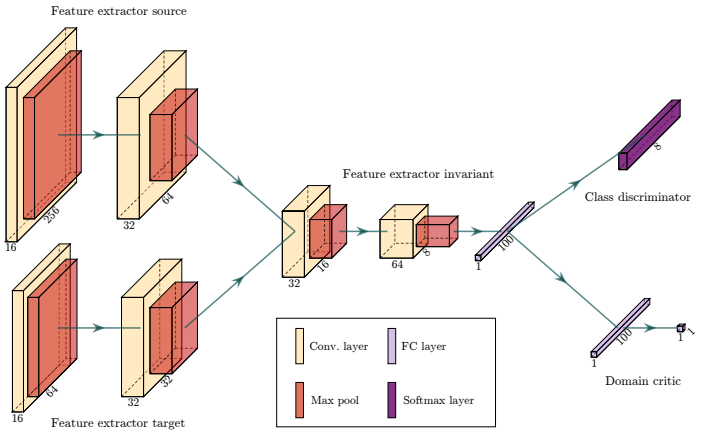


Fig. 4: The proposed heterogeneous semi-supervised domain adaptation model. The specific architecture presented is used for the case when the source dataset is RESISC45 and target dataset is EuroSAT. The kernel size of all convolutional layers is 5×5.

Let $X^s = \left\{(x_i^s, y_i^s)\right\}_{i=1}^{n^s}$ be a labelled source dataset of n^s samples from the domain $\mathcal{D}_s$ following the data distribution $\mathbb{P}_{x^s}$. SS-HIDA uses a small amount of target labels, so let us define two separate sets of target data, one being labelled $X^{tl} = \left\{(x_j^{tl}, y_j^t)\right\}_{j=1}^{n^{tl}}$, and the other being unlabelled $X^{tu} = \left\{x_k^{tu}\right\}_{k=1}^{n^{tu}}$, $n^{tl} << n^{tu}$, where target samples $x^t \in \left\{x_j^{tl}\right\}_{j=1}^{n^{tl}} \cup \left\{x_k^{tu}\right\}_{k=1}^{n^{tu}}$ come from the domain $\mathcal{D}_t$ and follow the data distribution $\mathbb{P}_{x^t}$. Unlike WDGRL, SS-HIDA is able to work with heterogeneous domains, i.e. $x^s \in \mathcal{X}^s$, $x^t \in \mathcal{X}^t$, $\mathcal{X}^s \neq \mathcal{X}^t$ where the dimensions d^s and d^t of spaces $\mathcal{X}^s$ and $\mathcal{X}^t$ may or may not differ.

SS-HIDA's architecture is presented in Figure 4, and consists of 5 neural network components: 3 feature extractors, a domain critic, and a class discriminator. To be able to work with the data coming from two different spaces, possibly of different input sizes, two different input branches are needed. Therefore, SS-HIDA has two separate feature extractors — $FE_s : \mathcal{X}^s \to \mathbb{R}^{d_1}$ and $FE_t : \mathcal{X}^t \to \mathbb{R}^{d_1}$ — these have the task to bring the data to a feature space of the same size — $g^s = FE_s(x^s)$ and $g^t = FE_t(x^t)$. Furthermore, another invariant feature extractor $FE_i : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is employed to model the similarity of the data domains, and to extract domain invariant features — $h^s = FE_i(g^s)$ and $h^t = FE_i(g^t)$. Note that in Figure 4, the specific architecture presented is for use on RESISC45 and EuroSAT datasets, which can be adapted to other datasets.

Wasserstein distance is used to measure the distance between domains. This metric is calculated by solving the optimal transport problem. Given two probability distributions μ and ν, it is necessary to find the best possible plan of

transport to transform μ to ν. The optimal transport plan can be represented as a joint probability distribution of marginals μ and ν. Let $\Pi(\mu, \nu)$ be the space of all such joint probability distributions. The optimal transport plan $P^* \in \Pi(\mu, \nu)$ is calculated such that

$$\begin{aligned} P^* &= \underset{P \in \Pi(\mu,\nu)}{\arg\min} \iint c(a,b) P(\mathrm{d}a, \mathrm{d}b), \\ s.t. &\int P(a,b)\,\mathrm{d}a = \nu(b), \int P(a,b)\,\mathrm{d}b = \mu(a), \end{aligned} \tag{1}$$

where $c(a, b)$ is a cost of transport (usually Euclidean), and $P(\mathrm{d}a, \mathrm{d}b)$ is the amount to be transported. The Wasserstein distance between distributions μ and ν is the total price of the transport using the optimal plan, and it is defined such that

$$W[\mu, \nu] = \iint c(a,b) P^*(a,b)\,\mathrm{d}a, \mathrm{d}b. \tag{2}$$

The dual formulation of the Wasserstein distance, equivalent to Eq. (2), is expressed as

$$\begin{aligned} W[\mu, \nu] &= \max_f \left(\int f(a)\mu(a)\,\mathrm{d}a - \int f(b)\nu(b)\,\mathrm{d}b \right), \\ s.t.\ f &\in L_C = \{ f : R \to R \mid f(a) - f(b) \leq c(a,b) \}, \end{aligned} \tag{3}$$

which is the difference between the mathematical expectations of function f under μ and under ν, with the Lipschitz contraint L_C that bounds the growth of f by c.

Since finding f is computationally expensive, the domain critic $DC : \mathbb{R}^{d_2} \to \mathbb{R}$ is trained to approximate it instead [14, 16], accelerating the training process. The domain critic uses the whole target dataset x^t including the unlabelled part, i.e. a total of $n^t = n^{tl} + n^{tu}$ samples. The loss of this component is defined such that

$$\mathcal{L}_{wd}(h^s, h^t) = \frac{1}{n^s} \sum_{i=1}^{n^s} DC(h_i^s) - \frac{1}{n^t} \sum_{j=1}^{n^t} DC(h_j^t). \tag{4}$$

In order to calculate the empirical Wasserstein distance, Eq. (4) needs to be maximised, therefore the domain critic component is trained by solving

$$\max_{\theta_{dc}} (\mathcal{L}_{wd} - \gamma \mathcal{L}_{grad}), \tag{5}$$

where θ_{dc} are the domain critic's weights and $\gamma \mathcal{L}_{grad}$ is a regularisation term enforcing the Lipschitz constraint. In the original version of the Wasserstein GAN, the critic function f was constrained by simple weight clipping. This choice had drawbacks such as exploding/vanishing gradients, and capacity underuse — choosing only simple functions for f. Gulrajani et al. [36] proposed an improved training procedure for Wasserstein GANs. They proved

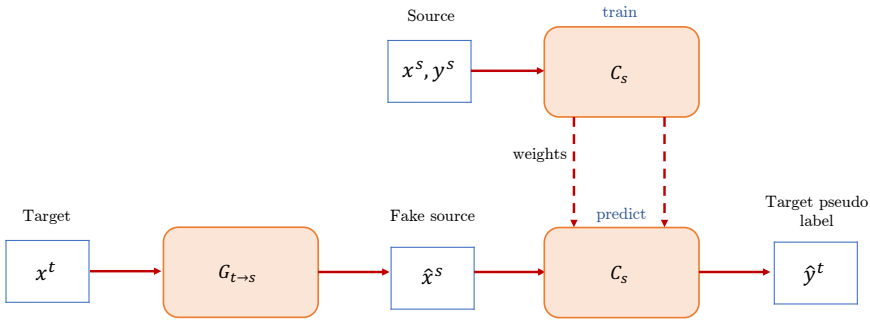


Fig. 5: Schema for obtaining pseudo-labels from CycleGAN. The generator $G_{t\to s}$ is translating target data x^t to the space of source domain. These "fake" source images $\hat{x}^s$ are then labelled by the classifier C_s which was previously trained on source data (x^s, y^s).

that the optimal choice for f has gradient norm 1 almost everywhere under two domains, making f 1-Lipschitz. Therefore, a regularisation term $\mathcal{L}_{grad}$ which penalises gradient norms different from 1 is added. When training our domain critic [16], the regularisation term amounts to

$$\mathcal{L}_{grad}(\hat{h}) = \left(\left\| \nabla_{\hat{h}} DC(\hat{h}) \right\|_2 - 1 \right)^2, \tag{6}$$

where $\hat{h}$ is the union of source and target representation points — h^s and h^t — and the points sampled from the straight lines between coupled points of h^s and h^t. This way, we are sufficiently close to enforcing the norm of 1 on the entire space of the two domains [36].

Finally, the class discriminator $C : \mathbb{R}^{d_2} \to \mathbb{R}^c$ (c being the number of classes) is trained on the extracted features of all the labelled samples $(h^l, y^l) = (h^s, y^s) \cup (h^{tl}, y^{tl})$. If labels y_i^l are one-hot encoded, cross-entropy classification loss is defined such that

$$\mathcal{L}_c(h^l, y^l) = -\frac{1}{n^s + n^{tl}} \sum_{i=1}^{n^s+n^{tl}} \sum_{k=1}^{c} y_{i,k}^l \log C(h_i^l). \tag{7}$$

If we denote the weights of the feature extractor as θ_{fe}, and the weights of the class discriminator as θ_c, the final min-max adversarial optimisation problem to be solved is

$$\min_{\theta_{fe},\theta_c} \left\{ \mathcal{L}_c + \lambda \max_{\theta_{wd}} \left[\mathcal{L}_{wd} - \gamma \mathcal{L}_{grad} \right] \right\}. \tag{8}$$

3.2 Unsupervised Pseudo-Labelled Heterogeneous Image Domain Adaptation (UPL-HIDA)

Unsupervised heterogeneous domain adaptation is a very challenging problem. With very different data modalities, and without any supervision in the target domain, it is very difficult to find correspondences between domains, limiting success.

The SS-HIDA model can be adjusted for use in unsupervised DA by simply excluding labelled target samples from the classification loss in Eq. (7), such that

$$\mathcal{L}_c(h^s, y^s) = -\frac{1}{n^s} \sum_{i=1}^{n^s} \sum_{k=1}^{c} y_{i,k}^s \log C(h_i^s). \tag{9}$$

However, preliminary experiments showed that such an approach does not work reliably, as will be explored in Section 4.1.5. The reason for this is that the classification loss, based only on the source samples, will not update the weights of the target feature extractor FE_t during backpropagation. These weights will only be updated by the loss of the domain critic $\mathcal{L}_{wd}$, which will be shown to be insufficient.

To make up for the absence of labels in the target data, we can rely on a pseudo-labelling approach instead. For this purpose, we choose the strategy employed by Voreiter et al. [11]. This specific method is chosen because it works on raw image data and is evaluated on two heterogeneous remote sensing datasets. The authors train a variant of CycleGAN and use it to translate between two image domains of different resolutions. The generator normally used in CycleGAN is here replaced by the generator from super-resolution GAN (SRGAN) [37] to handle resizing of the images during the translation. After their CycleGAN is trained, it is used to translate images from the target to the source domain. Translated images are assigned pseudo-labels by a pre-trained source classifier. In the original work, these pseudo-labels are then used for training the final classifier in the target space. Instead, we can use the pseudo-labels to replace the now missing target labels and perform training in the same manner as with SS-HIDA. We will name this approach UPL-HIDA (Unsupervised Pseudo Labelled HIDA). An overview is presented in Figure 5.

If we denote the generator for translating the target to source domain as $G_{t\to s}$, and the pre-trained source classifier as C_s, we can express the formula for calculating pseudo-labels as

$$\hat{y}^t = C_s\left(G_{t\to s}\left(x^t\right)\right). \tag{10}$$

In order to eliminate unreliable pseudo-labels we filter them and use only the most confident predictions. The usual approach would be to set a threshold for the probability given at the output of the softmax layer of C_s, and use only those target samples that exceed this threshold. However, in order to keep the dataset balanced and to not have any under-represented classes, we set this threshold per class, and ensure a balanced class representation. Let us

denote the filtered target samples as x^{tf}, their extracted features as h^{tf}, their pseudo-labels as $\hat{y}^{tf}$, and their number as n^{tf}.

The class discriminator of UPL-HIDA is trained on the union of extracted features of labelled source samples and filtered pseudo-labelled target samples $(h^l, \hat{y}^l) = (h^s, y^s) \cup (h^{tf}, \hat{y}^{tf})$. The classification loss is therefore defined such that

$$\mathcal{L}_c(h^l, \hat{y}^l) = -\frac{1}{n^s + n^{tf}} \sum_{i=1}^{n^s+n^{tf}} \sum_{k=1}^{c} \hat{y}^l_{i,k} \log C(h^l_i). \tag{11}$$

4 Experimental results

4.1 Remote sensing

In remote sensing, the data used is very different than in standard computer vision. In CV, normally the task is to recognise natural object(s) on images, while in RS images the land covers should be distinguished. Land cover classification requires understanding the complete image to determine what kind of area is shown, where in CV we pay attention to the specific objects on images. Therefore, the CV methods often cannot be applied successfully to RS. Another problem, as mentioned, is lack of reference data in RS, making it difficult to train supervised deep learning models. In CV on the other hand, the existence of large-scale ImageNet dataset allowed for training huge models with very high performance. Furthermore, ImageNet models provide good-quality features for other CV datasets as well, hence they could be used for a broad spectrum of tasks with transfer learning. But ImageNet features of RS data are not as good as for common CV benchmarks, making transfer learning in RS much more difficult task than in CV.

4.1.1 Data

The proposed approach is evaluated on the following eight corresponding classes from two heterogeneous remote sensing datasets (details given in Table 1 and examples of classes given in Figure 6):

- NWPU-RESISC45 [2] (high resolution aerial RGB images extracted from Google Earth) — dense residential, forest, freeway, industrial area, lake, meadow, rectangular farmland, and river.
- EuroSAT [1] (low resolution multi-spectral images from the Sentinel-2A satellite) — residential, forest, highway, industrial, sealake, pasture, annual crop and permanent crop (two classes merged into one), river.

The reference data is given as a single label per patch, the problem to be solved is therefore patch classification.

The RESISC45 dataset is composed of images taken from 100 countries and regions all over the world, throughout all seasons and all kinds of weather. The EuroSAT dataset covers 34 European countries, and also consists data from all over the year. Both datasets therefore have in-domain temporal and

Name	Source	Image Size	Patches	Classes	Resolution
RESISC45	Aerial	$256 \times 256 \times 3$	31,500	45	0.2 m−30 m
EuroSAT	Satellite	$64 \times 64 \times 13$	27,000	10	10 m

Table 1: Characteristics of NWPU-RESISC45 and EuroSAT datasets

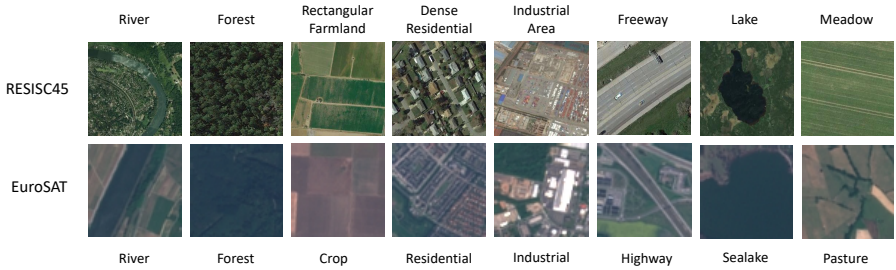


Fig. 6: Examples of chosen corresponding classes from RESISC45 and EuroSAT datasets. For EuroSAT, RGB version of the dataset is shown.

geographical variability. This variability, when intra-class, can make even the in-domain problem of classification very difficult. Consider the Figure 7a with examples of images from the same class. It can be seen that the variability can be huge, resolution can vary a lot in RESISC45, ranging from car clearly visible on the road to the road itself barely visible. The apperance of Crop class in Eurosat dataset can also vary greatly.

The problem becomes even worse with intra-class similarity on top of inter-class variability. Industrial and residential area are both images of buildings which can frequently cause mix-up. In RESISC45, dense forest can resemble meadows having similar colour and texture on aerial images. In low-resolution EuroSAT dataset, forest sometimes look just like flat green patch without texture, which can resemble the patches of green water from lake/sea. Pasture images can have separate areas of vegetation and soil, much like crops (Figure 7b).

When performing patch classification, one issue is that multiple classes of land cover can be present on a patch. The experts who annotated images chose the label that is dominant or central to the patch, however the presence of other classes is sometimes significant and can be misleading when the model is learning. This especially holds e. g. when river or road are passing though the residential or agricultural area, such as in Figure 7c.

As in-domain classification already turns to be a challenging problem in remote sensing, transfer learning brings another level of difficulty, especially with the huge domain shift like in our case. As seen in Figure 8, images from some classes might tend to be aligned with the wrong class in the other domain. Lake class in RESISC45 shows the entire lake with surrounding area, whereas

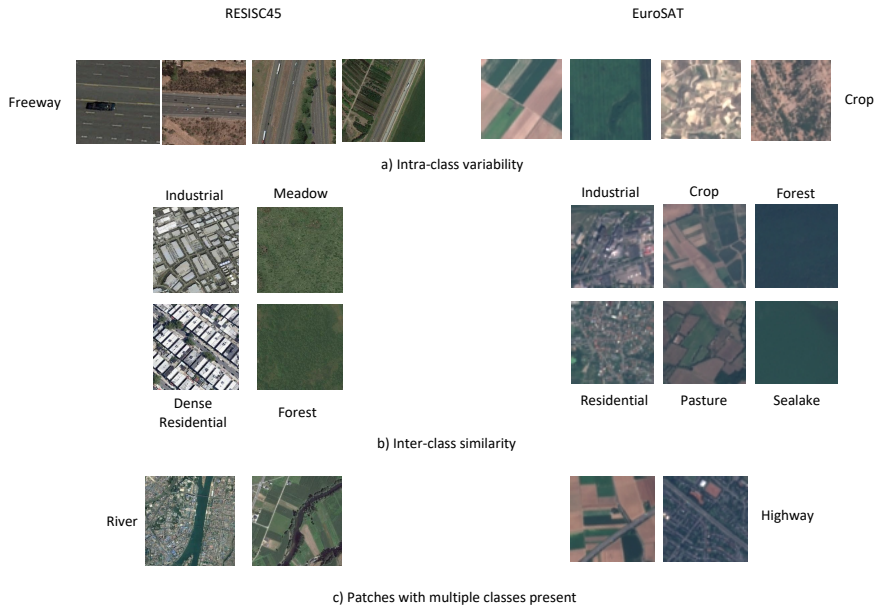


Fig. 7: Examples of issues when classifying remote sensing datasets: a) intra-class variability, b)inter-class similarity, c) patches with multiple classes present

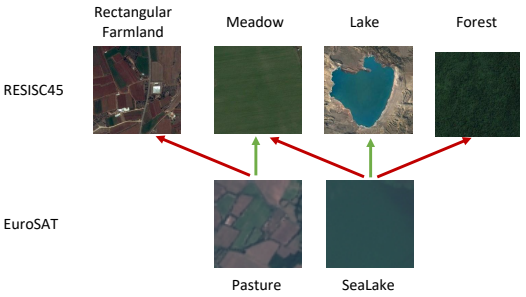


Fig. 8: Examples of issues when doing transfer learning between RESISC45 and EuroSAT

in EuroSAT only a patch of water is shown, making it more similar to other classes with single-color images like meadow and forest.

One advantage of our heterogeneous approach is that we can utilise information from all the channels. The additional information provided by non-RGB channels can be very useful in discriminating different classes, however it is usually neglected in other works. Images from multispectral EuroSAT

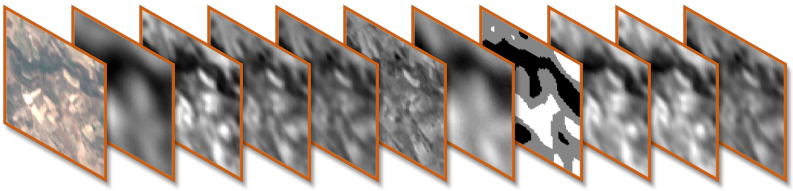


Fig. 9: All bands of a multispectral image. Red, green, and blue channels are shown together as an RGB image, while other channels are shown as greyscale images.

dataset, aside from visible RGB band, also have near infrared band (NIR) band, short-wave infrared (SWIR) and red edge bands etc. The 13 channels of an EuroSAT image are shown in Figure 9.

The datasets are split into train, validation, and test sets with the proportion of 60:20:20 while keeping the classes balanced in all sets. The test set was set aside during development and only used for the final experiments presented herein.

4.1.2 Implementation details

Unlike WDGRL, whose components are fully connected neural networks, SS-HIDA and UPL-HIDA are convolutional architectures (see Figure 4 for details). The feature extractor for RESISC45 consists of two convolutional layers with 16 and 32 filters respectively. Each conv. layer is followed by 4×4 max-pooling. The feature extractor for EuroSAT is the same, except that it has 2×2 max-pooling after every conv. layer. The shared invariant feature extractor has two convolutional layers with 32 and 64 filters respectively, and one fully-connected (FC) layer of 100 nodes. All of the kernels have size 5×5. The class discriminator has one FC layer with softmax activation. The domain critic (DC) is identical to that in WDGRL — it has an FC layer with 100 nodes followed by an FC layer with 1 node.

In each training step, the DC is trained for 10 iterations with a learning rate of 10^{-3}, the DC is then frozen and the rest of the model is trained for 1 iteration with a learning rate of 10^{-4}. The DC loss' weight λ is 0.1. The Adam optimiser is used.

The input data is standardised per channel so that each channel has mean 0 and standard deviation 1. The following augmentation transformations are used: flipping with a probability of 0.45, rotation with a probability of 0.75 for 90°, 180°, or 270°, changing contrast with the probability of 0.33 by multiplying the values of the pixels with the coefficient ranging between 0.5 and 1.5, changing brightness with the probability of 0.33 by adding the coefficient ranging between -0.3 and 0.3 scaled by the mean of pixel values per channel before standardisation, blurring with the probability of 0.33 with Gaussian filter with σ parameter values ranging from 1.5 to 1.8, and finally adding Gaussian noise with mean 0 and standard deviation between 10 and 15 with the probability

R → E	**6.25%**	**2.5%**	**1.25%**
Target classifier	76.07 (1.75)	64.95 (3.25)	58.70 (4.19)
CycleGAN for HDA	63.29 (3.80)	56.39 (6.70)	41.57 (9.20)
SS-HIDA	**81.34** (**1.24**)	**70.91** (**2.13**)	**66.14** (**2.92**)

E → R	**6.25%**	**2.5%**	**1.25%**
Target classifier	66.25 (2.90)	58.41 (1.73)	47.55 (5.11)
CycleGAN for HDA	58.75 (6.91)	50.79 (5.40)	47.29 (1.53)
SS-HIDA	**73.57** (**2.64**)	**68.34** (**2.59**)	**62.68** (**3.24**)

Table 2: Top — Accuracy of domain adaptation with RESISC45 as source and multispectral EuroSAT as target (R → E). Bottom — Accuracy of domain adaptation with multispectral EuroSAT as source and RESISC45 as target (E → R). Standard deviation is shown in parentheses. In both RESISC45 and EuroSAT, 6.25%, 2.5%, and 1.25% labelled data is 25, 10, and 5 images per class respectively.

of 0.33. The batch size is 32, and in each iteration, half of the training batch (16) comes from the source, and the other half from the target domain. The model is trained for 40 epochs.

4.1.3 Compared approaches

To the best of our knowledge, there are no other HDA methods created for working with the raw images from two domains of unpaired data with different number of channels. The only method we found possible to compare SS-HIDA and UPL-HIDA with is CycleGAN for HDA by Voreiter et al. [11]. This method is specifically tailored for data with different spatial resolution, but technically could be applied to data with different number of channels as well. CycleGAN for HDA is representative of an image-to-image translation model (contrary to our domain invariant SS-HIDA and UPL-HIDA), and can be used in both SSDA and UDA setting. The results are also compared with the performance of a simple target baseline, i.e. a classifier trained on the same amount of labelled target data as our semi-supervised DA model. The same architecture is used: the same layers as the target FE, invariant FE, and class discriminator. The same augmentation transformations (described above) are also used.

The semi-supervised DA models are evaluated on different amounts of labelled target data — 25 (6.25%), 10 (2.5%), and only 5 labelled samples per class (1.25%). The unsupervised models are evaluated without using any labelled target data.

4.1.4 Semi-supervised DA Results

The accuracy of the proposed and comparison models are presented in Table 2. Two cases are demonstrated, one when the RESISC45 is used as a source

R → E-RGB	**6.25%**	**2.5%**	**1.25%**
Target classifier	65.10 (2.10)	55.57 (1.29)	44.55 (3.85)
CycleGAN for HDA	54.54 (3.53)	46.00 (7.84)	49.59 (9.33)
SS-HIDA	**69.07** (**1.01**)	**62.50** (**3.52**)	**56.52** (**2.39**)

E-RGB → R	**6.25%**	**2.5%**	**1.25%**
Target classifier	66.25 (2.90)	58.41 (1.73)	47.55 (5.11)
CycleGAN for HDA	51.11 (7.64)	52.70 (5.45)	38.38 (3.91)
SS-HIDA	**73.98** (**0.99**)	**65.39** (**3.38**)	**62.34** (**4.82**)

Table 3: Results when using only RGB bands. Top — Accuracy of domain adaptation with RESISC45 as source and RGB EuroSAT as target (R → E-RGB). Bottom — Accuracy of domain adaptation with RGB EuroSAT as source and RESISC45 as target (E-RGB → R). Standard deviation is shown in parentheses. In both RESISC45 and EuroSAT, 6.25%, 2.5%, and 1.25% labelled data is 25, 10, and 5 images per class respectively.

domain, and vica-versa. All 13 bands from the EuroSAT dataset were used throughout.

The results show that SS-HIDA outperforms the competing method CycleGAN for HDA by a large margin in all cases. With RESISC45 as source and EuroSAT as target (R → E), SS-HIDA gains around 14–24% in accuracy, the most for the case with 1.25% labelled data. With EuroSAT as source and RESISC45 as target (E → R), the difference in favor of SS-HIDA is around 15–18%. It should also be noted that the results of CycleGAN for HDA almost always have much higher standard deviation than those of SS-HIDA. As with all GAN architectures, training CycleGAN is unstable, so the result may vary for different runs. Nevertheless, in our experiments, the maximal result that CycleGAN achieved was lower than the minimal result of SS-HIDA in all cases.

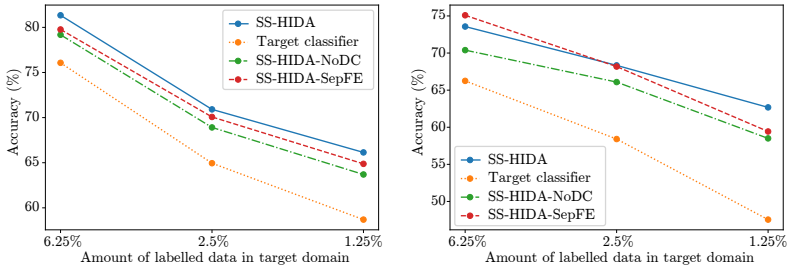
The baseline classifier performs better than CycleGAN for HDA. For R → E, SS-HIDA is stronger than the baseline by around 5–7%. For E → R, the gain of SS-HIDA is 7–15%, the highest gain being for the case of 1.25% labelled data. The reason for the higher gap in E → R is the fact that RESISC45 is more difficult to solve than EuroSAT, hence the baseline classifier for RESISC-45 cannot perform as well as that of EuroSAT, while SS-HIDA is not as affected and retains strong performance with RESISC45 as the target domain.

It is worth noting that the baseline performs surprisingly well with such few labelled images, achieving almost 60% for R → E, and almost 50% for E → R when only five labelled images per class are given. Keeping in mind that the specific architecture used is not optimised to achieve state-of-the-art performance, this indicates that the classification problem is relatively easy, especially for the EuroSAT dataset (which is backed up by other findings in the literature [30, 38]) and perhaps more pronounced improvements could be found in more difficult applied problems.

Fig. 10: Comparing translations of CycleGAN models trained to translate from Multispectral EuroSAT to RGB RESISC45 (second column) and RGB EuroSAT to RGB RESISC45 (third column). The first column shows original EuroSAT images and the fourth column original RESISC45 images of the corresponding class.

Results with RGB bands only. To assess the impact of using non-RGB bands, this section presents the DA results of SS-HIDA and the comparison models using the RESISC45 dataset, and RGB-only bands of the EuroSAT dataset. CycleGAN is known to have difficulties when translating between domains of different spectral bands. Figure 10 presents the translations when using multispectral and RGB-only versions of EuroSAT. It is clear that this could degrade DA performance. On the other hand, using additional information from non-RGB bands could improve both CycleGAN for HDA and SS-HIDA classification performance, especially when EuroSAT is the target domain.

In the following experiments the images do not undergo re-sampling and therefore, although both domains are now RGB, they have different resolutions and image sizes. The results are shown in Table 3. As can be seen, SS-HIDA trained on RGB-only data still outperforms both CycleGAN for HDA and the baseline classifier, gaining 7–16% over CycleGAN for HDA, and 4–12% over the baseline in R $\rightarrow$ E-RGB. The advantage of SS-HIDA is even more pronounced for the opposite case (E-RGB $\rightarrow$ R), gaining 13–24% over CycleGAN for HDA (again with reduced standard deviation), and 7–15% over baseline.



(a) Source: RESISC45, Target: EuroSAT

(b) Source: EuroSAT, Target: RESISC45

Fig. 11: Ablation study of SS-HIDA, comparison with the model without domain critic and without shared layers with varying numbers of labelled training images. The numbers are expressed in percentages of labelled images.

It should be noted that in all cases when EuroSAT is the target, the result of SS-HIDA is higher when using all of the bands compared to using RGB-only bands. Even when EuroSAT is used as a source domain, the results on the target domain are either comparable or higher when using all of the bands of EuroSAT. This proves that SS-HIDA is taking advantage of the additional information in the multispectral bands. As for CycleGAN for HDA, using non-RGB bands is sometimes more and sometimes less successful than using RGB-only data. While it can certainly make use of non-RGB information, the limitation is that CycleGAN translations between RGB and multispectral spaces are not of very good visual quality (Figure 10). This shows the clear advantage of a domain invariant method over translation, it can only benefit from additional bands.

Ablation Study. In order to uncover the impact of each of the model's components, an ablation study is performed. One comparison model is created in which the domain critic is removed (SS-HIDA-NoDC), thus removing the domain adaptation component. A second comparison model is obtained by separating all of the layers of source and target architecture so that only the classifier is shared between them (SS-HIDA-SepFE), thus reducing the capacity of learning a general representation. The multispectral version of the EuroSAT dataset is used.

The results are shown in Figure 11. In both cases we can confirm that removing the domain critic leads to a significant drop in performance. In this case, there is no requirement for the model to learn overlapping distributions, therefore reducing the classifier's ability to generalise between domains. On the other hand, separating the source and target layers has less effect on performance. When RESISC45 is the source and EuroSAT is the target, SS-HIDA-SepFE is a little worse than SS-HIDA. But when EuroSAT is the source and RESISC45 is the target, SS-HIDA-SepFE even outperforms SS-HIDA when there is 6.25% labelled data in target domain. The two models

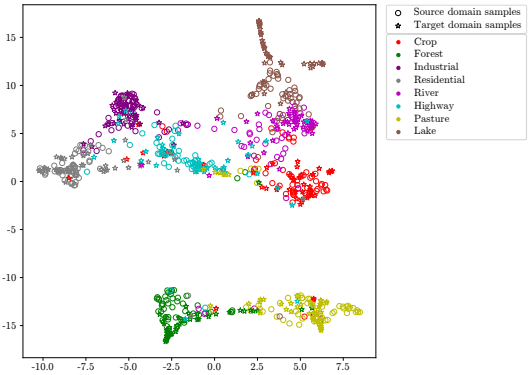


Fig. 12: PaCMAP visualisation of SS-HIDA features.

fare about the same for 2.5%. But SS-HIDA remains significantly better when there is only 1.25% target labelled data. From this we can conclude that, when there are sufficient labels in the target domain, the domain critic is able to compensate for the separation of source and target layers, and still forces the model to extract domain invariant features. It is also worth noting that none of these variations result in performance worse than the baseline.

Feature visualisation. A PaCMAP [39] visualisation of the features from the penultimate layer of SS-HIDA is shown in Figure 12. It illustrates that, although these two datasets have different dimensionalities/modalities and pass through two separate feature extractors, SS-HIDA successfully learns a latent space in which their distributions overlap and their classes are well-matched.

4.1.5 Unsupervised DA results

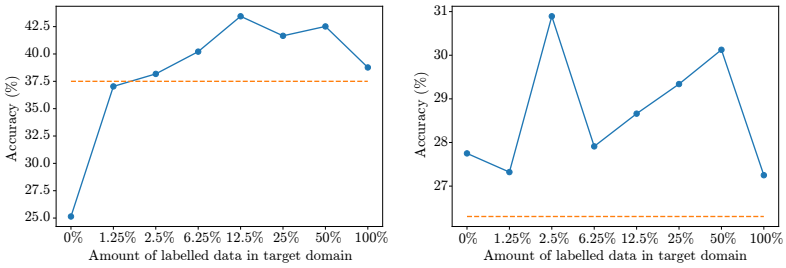
Our unsupervised pseudo-labelled solution, UPL-HIDA, depends upon a threshold value to define how many samples should be pseudo-labelled. The threshold of 12.5% was found to be a good choice in preliminary experiments, meaning that the most confident 50 images per class are pseudo-labelled. The following experiments were performed using this threshold value, and the algorithm's sensitivity to this parameter will be explored at the end of this section.

UPL-HIDA is trained and compared with CycleGAN for HDA. The results can be seen in Table 4 and these show that heterogeneous UDA is indeed a challenging problem. Nevertheless, UPL-HIDA manages to outperform the competing method in almost all of the cases. The best results can be obtained in the R $\rightarrow$ E case using RGB EuroSAT where UPL-HIDA achieves an average accuracy of 43.64%, outperforming CycleGAN for HDA for more than 3%. The standard deviation, however, is very high for both models. The reason being that CycleGAN occasionally fails to learn meaningful translations, giving completely wrong pseudo-labels, which directly affects the performance of

R → E	**RGB**	**Multispectral**
CycleGAN for HDA	39.98 (16.25)	18.48 (8.00)
UPL-HIDA	**43.64** (**16.15**)	18.84 (7.34)

E → R	**RGB**	**Multispectral**
CycleGAN for HDA	24.96 (11.94)	16.82 (5.74)
UPL-HIDA	**27.39** (**15.23**)	**21.14** (**5.33**)

Table 4: Results of unsupervised domain adaptation models. Top — Accuracy of unsupervised domain adaptation with RESISC45 as source and EuroSAT as target (R → E). Bottom — Accuracy of unsupervised domain adaptation with EuroSAT as source and RESISC45 as target (E → R). Standard deviations are shown in parentheses next to the accuracy.



(a) Source: RESISC45, Target: EuroSAT

(b) Source: EuroSAT, Target: RESISC45

Fig. 13: Accuracy of the unsupervised pseudo-labelled solution with varying thresholds for choosing the most confident pseudo labels. Horizontal dashed line shows the performance of CycleGAN for HDA. The numbers are expressed in percentages of labelled images.

UPL-HIDA. But in most of the cases, the accuracy of both models is between 40% and 55%, with UPL-HIDA improving over the result of CycleGAN for HDA in all of the experimental runs.

The difficulty of using RESISC45 as the target domain, limits the performance of UDA methods (including increased standard deviations). Still, UPL-HIDA is more successful, gaining more than 2% over CycleGAN for HDA.

The UDA models particularly struggle when multispectral data is used. Without supervision, adding non-RGB channels appears to confuse the models rather than providing additional information. As shown before, CycleGAN for HDA does not translate very well between RGB and multispectral data,

and UPL-HIDA fails to find correspondences between features in such heterogeneous domains. For R → E, the performance of two models is very similar. In the opposite E → R case, UPL-HIDA is stronger by more than 4%.

Ablation study. In order to asses the impact of using different thresholds for pseudo-labelling, we show the results of UPL-HIDA on different amounts of pseudo-labelled target data, ranging from 100% (whole dataset) to 1.25% (the most confident 5 images per class are pseudo-labelled). The results without using any pseudo-labels (0%) are also included. The performance of CycleGAN for HDA is given as a horizontal line. The comparison is shown in Figure 13. These results are obtained using the RGB version of EuroSAT. As CycleGAN for HDA does not translate very well between RGB and multispectral data, the potential of UPL-HIDA using pseudo-labels given by CycleGAN is better seen on RGB-only data. N.B. These results are not comparable to those above as the ablation study is performed on the validation set of the target domain (rather than the test set).

When RESISC45 is source and EuroSAT is target domain, starting with 1.25%, the accuracy grows as more pseudo-labelled data is added, outperforming CycleGAN for HDA from 2.5%, and reaching its peak with a threshold of 12.5%. Afterwards, additional pseudo-labels become less reliable and harm performance, but the accuracy remains higher than that of CycleGAN for HDA.

The situation is not as clear when adapting in the opposite direction, when EuroSAT is source and RESISC45 is target. As stated before, the RESISC45 dataset is more difficult to solve than EuroSAT, so the quality of pseudo-labels given by CycleGAN is not very high to begin with. Regardless, the model without using any pseudo-labels (0%) already outperforms CycleGAN for HDA, and remains higher in all the cases. As can be seen, better performances could have been obtained using threshold values different than 12.5% (notably 2.5%), however optimising this parameter would require using a certain amount of additionally labelled target data, so it was not done for these experiments.

It is worth mentioning that the pseudo-labelling strategy used in UPL-HIDA does not provide any improvement when used with SS-HIDA. SS-HIDA already makes a good use of available target labels and vastly outperforms CycleGAN for HDA, therefore the pseudo-labels provided by CycleGAN are not helpful. Filtering does not help either, as the most confident samples are usually very similar to the available labelled images, thus not bringing any new information to the model.

Feature visualisation. A PaCMAP visualisation of the features from the penultimate layer of UPL-HIDA is shown in Figure 14. UPL-HIDA successfully learns a latent space in which the distributions of the datasets overlap. However, as there is no supervision for the target domain, label flipping can occur — "river" target points (magenta) are matched with "lake" source points (brown). Also, target points of certain classes are more dispersed than was observed with SS-HIDA, notably "crop" (red points) and "highway" (cyan

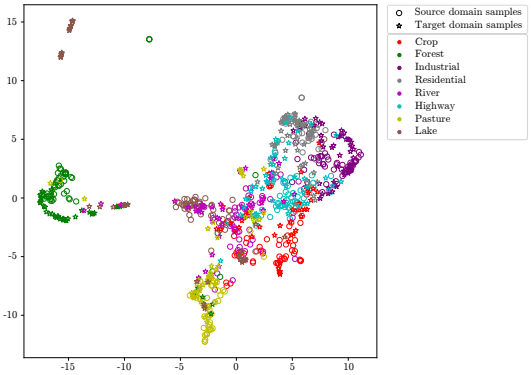


Fig. 14: PaCMAP visualisation of UPL-HIDA features.

R → E	
Source classifier	86.16 (0.63)
UPL-HIDA	**87.71** (**1.33**)

E → R	
Source classifier	93.07 (0.95)
UPL-HIDA	93.05 (0.74)

Table 5: Source domain performance of the baseline source classifier and UPL-HIDA using multispectral EuroSAT without pseudo-labels. Top — Accuracy on the source domain with RESISC45 as source and EuroSAT as target (R → E). Bottom — Accuracy on the source domain with EuroSAT as source and RESISC45 as target (E → R). Standard deviations are shown in parentheses next to the accuracy.

points). Sebag et al. [40] point out that in adversarial domain adaptation, unlabelled samples are only constrained by the domain discriminator, as opposed to labelled samples, therefore domain alignment can shuffle unlabelled samples and lead to negative transfer.

Results on the source domain. An advantage of domain invariance over translation is that the final model can work in both (or more) domains. SS-HIDA and UPL-HIDA's performance on the source domain is comparable to the source baseline classifier in all the cases. UPL-HIDA using multispectral EuroSAT and not using pseudo-labels even outperforms the baseline classifier trained on the source domain in R → E and has the same performance as the baseline in E → R, see Table 5.

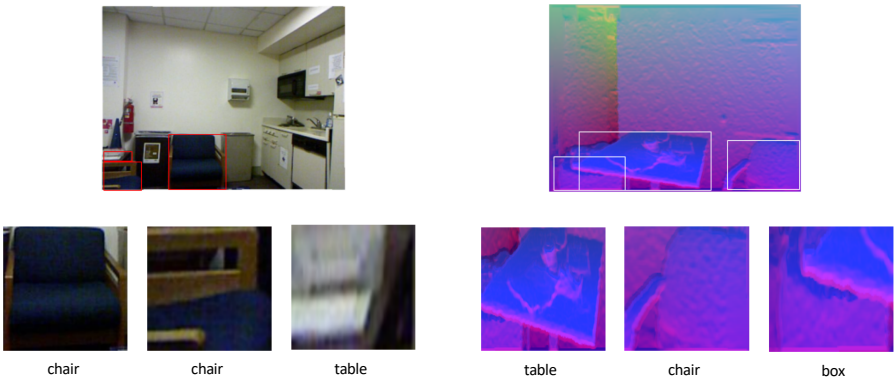


Fig. 15: Visualisation of RGB-D data from NYU depth V2 dataset. Shown are two whole scenes — one RGB and one depth map — and several patches extracted from them.

4.2 RGB-Depth adaptation

Apart from application to the remote sensing, we also evaluate our method on a common computer vision benchmark of adaptation between RGB and depth images with the goal to show that we have developed a general-purpose model which could be applied in various fields. Robots are often equipped with a depth sensor in addition to an RGB camera, to be able to measure the distance to the objects and have better orientation in space. One possible benefit of doing RGB-depth adaptation is being able to recognise objects when the visibility is bad, i. e. during the night. In that case only depth images are available, but there are much less labelled depth datasets in comparison to abundance of RGB labelled data, so it is a logical choice to use the knowledge from related RGB datasets.

4.2.1 Data

We evaluate our method on NYU depth V2 dataset [41]. This dataset consists of paired RGB and depth images (RGB-D) of indoor scenes captured by cameras from the Microsoft Kinect. Since our method is a single-label patch classification model, we cut the patches based on tight bounding boxes around the objects of interest, as is done in other works on object classification of NYU dataset [9, 42]. The objects are categorised into 19 classes. The dataset is highly imbalanced, with some classes having very few examples. Figure 15 shows the original images and extracted patches, both RGB and depth. The problem is obviously very challenging even for in-domain classification in both domains, and more so for domain adaptation. The patches are sometimes of very low resolution and they are often blurry, so it is difficult even for a human to do the correct classification.

The original NYU depth V2 dataset has corresponding paired RGB and depth images showing the same scene. As our method does not require paired

data, we split the training set in two equal halves, of which from one we take RGB images for source domain, and from the other we use the depth images as a target domain. In the end, both our source and target domain consists of 1958 patches each. Validation and test patches are taken from the images from the validation and data test as given in original dataset, there are 775 RGB-D pairs of patches in validation set, and 3859 in test set.

Raw depth values are encoded as one channel (greyscale) image. Instead of raw depth images, we use HHA encoding [43] which converts raw depth maps to a three-channel image, the channels being horizontal disparity, height above ground, and the angle the pixel's local surface normal makes with the inferred gravity direction.

4.2.2 Compared approaches

We compare our approach with Adversarial Discriminative Domain Adaptation — ADDA method [9]. This method is based on translation, contrary to our domain-invariant method. The translation is done in a feature spaces, unlike CycleGAN which translates on the level of pixels. Even though ADDA is a several years old method and not a state-of-the-art in homogeneous DA anymore, it is the only general-purpose DA method we found that is evaluated on the cross-modal adaptation between unpaired RGB and depth domains for object (patch) classification task.

Having two separate feature extractors for source and target data, ADDA is able to work with heterogeneous domains who have different channels, like RGB and HHA-depth. But the limitation of the algorithm is that two domains need to have the same number of channels. HHA encoding has 3 channels as RGB, so ADDA can be used here, but cannot be used on domains with different number of channels such as we saw in remote sensing. For this adaptation task, ADDA is using VGG-16 architecture [44] with pretrained weights from ImageNet.

CycleGAN for HDA is specifically made for remote sensing application, and is not evaluated on general CV benchmark, so we do not include it here.

We also compare with the baseline — a classifier trained on source data only, and then evaluated on target data without performing any adaptation.

We show only unsupervised DA results, because ADDA cannot be used in semi-supervised setting, but we note that, as shown before, our method can fully take advantage of available labels in target data.

4.2.3 Implementation details

To have a fair comparison with ADDA method, we also use VGG-16 architecture as a basis for our model. We separate first two convolutional layers into source and target FE; the rest of the layers stays in the common FE. The convolutional layers are initialised with pretrained ImageNet weights, with source and target FE being initialised identically; the fully-connected layers are initialised randomly. At the first phase of the training, convolutional part of the network is frozen, and only FC layers are trained with the learning rate 10^{-4}.

	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	television	toilet	overall
baseline	0	0.1281	0.0042	0.2685	0.3422	0.1885	0.0114	0.5841	0	0.0205	0.598	0.0176	0.0198	0.5654	0.0481	0.0412	0.066	0.0115	0.0057	0.2630
ADDA	0	0.146	0.046	0.229	0.344	0.447	0.025	0.023	0	0.018	0.292	0.081	0.02	0.297	0.021	0.116	0.143	0.091	0	0.2110
UPL-HIDA	0	0.0311	0.0408	0.1165	0.4335	0.2594	0.0114	0.5146	0.0316	0.1103	0.1537	0.0274	0	0.8306	0.0648	0.0543	0.1315	0.0115	0.0171	0.2900
U-HIDA	0	0.2443	0.1	0.1913	0.6659	0.5146	0.0263	0.7414	0	0.0667	0.2966	0.0137	0.0099	0.5545	0.1667	0.1358	0.314	0.0038	0	**0.3874**

Fig. 16: Results of unsupervised DA methods on NYU depth V2 dataset, per class and overall, expressed in accuracy.

Then the whole network is fine-tuned with a smaller learning rate of 10^{-5}. We present results of two variants of our unsupervised method:

- UPL-HIDA — using CycleGAN to obtain pseudo-labels; train FC layers for 1 epoch, then finetune for 5 epochs; CycleGAN trained with ResNet generator for 50 epochs (no super-resolution); CycleGAN translated target images pseudo-labelled with a source baseline classifer with FC layers trained for 40 epochs, then finetuned for 40 additional epochs; most confident 10 images per class (where available) are pseudo-labelled;
- U-HIDA — no pseudo-labels; train FC layers for 40 epochs, then finetune for 30 epochs.

Domain critic has similar architecture as the domain classifier in ADDA, it is a fully connected neural network consisted of 3 layers with 1024 nodes, 2048 nodes, and 1 node in the output layer. The rest of the training process is the same as with RS datasets.

The data is preprocessed as is required to use pretrained VGG network — all the patches are resized to 224×224 size, and all the channels are zero-centered without scaling. Data augmentation is not used here to be fair in comparison to ADDA which also does not use it. The batch size used is 256 — 128 per domain.

Note that VGG can be easily replaced with some other architecture like ResNet.

4.2.4 Unsupervised DA results

The results of our and comparing methods are shown in Table ?. Accuracy per class and overall accuracy are shown. Both our methods give higher overall accuracy than ADDA, showing the advantage of domain-invariant over a translation method, with U-HIDA having the stronger performance than UPL-HIDA. The pseudo-labels provided by CycleGAN are not of very good quality in this case, because the dataset used has high number of classes (19), and is highly imbalanced. The usage of pseudo-labels therefore degrades the performance of our adaptation method, which results in U-HIDA having higher overall accuracy. U-HIDA also has the best performance on 9 classes, and UPL-HIDA on 4 classes. The worst performance in general is on the smallest classes — class *bathtub* (13 images only) is never predicted by any model, and classes *toilet* (16 images) and *dresser* (31 images) are predicted very rarely.

The results presented herein prove that our method can outperform translation method and achieve SOTA performance not only on remote sensing, but also on CV benchmarks. Having in mind the high number of classes, the low-resolution patches, difficulty to classify even for humans, and a huge difference between domains, the improvement in performance that our methods bring on this challenging problems is significant.

5 Conclusions

This article has proposed a novel approaches to semi-supervised and unsupervised heterogeneous image domain adaptation called SS-HIDA and UPL-HIDA. To the best of our knowledge, these are the first such approaches to extract domain-invariant features. The results showed that our domain-invariant approach significantly outperforms the competing image-to-image translation method, especially in the semi-supervised case. SS-HIDA showed that it can make use of non-RGB data to improve performance. UPL-HIDA showed that there is a potential for combining domain-invariant and translation methods in UDA. SS-HIDA and UPL-HIDA also give comparable results on the source data, with UPL-HIDA even outperforming the source baseline under certain conditions. The models were evaluated on two very challenging cases — a remote sensing case with an aerial dataset RESISC45 and a satellite dataset EuroSAT, but also on a RGB and depth images proving that our method not limited to remote sensing applications only, and could be used for other cases of heterogeneous images.

A possible improvement to SS-HIDA and UPL-HIDA might be brought by using other pseudo-labelling techniques without CycleGAN instability, which will be investigated in future work. Using our method for the task of semantic segmentation in images of heterogeneous domains should also be considered. Adaptation between optical and SAR images would be a very useful use case e. g. for the tasks of change detection during natural catastrophes. Another interesting direction is to extend our method to time-series data of different sensors; using temporal information for land cover classification is one of the most important trends in RS recently. Future work could also include application to other fields like medical imaging.

Acknowledgments. This work was granted access to the HPC resources of IDRIS under the allocation 2021-A0111011872 made by GENCI. We thank Nvidia Corporation for donating GPUs and the Centre de Calcul de l'Université de Strasbourg for access to the GPUs used for this research. Supported by the French Government through co-tutelle PhD funding.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are

submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
- Ethics approval
- Consent to participate
- Consent for publication
- Availability of data and materials
- Code availability
- Authors' contributions

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

References

[1] Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J Sel Top Appl Earth Obs Remote Sens **12**(7), 2217–2226 (2019)

[2] Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)

[3] Rudner, T.G., Rußwurm, M., Fil, J., Pelich, R., Bischke, B., Kopačková, V., Biliński, P.: Multi3Net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 702–709 (2019)

[4] Li, J., Lu, K., Huang, Z., Zhu, L., Shen, H.T.: Heterogeneous domain adaptation through progressive alignment. IEEE Trans. Neural Netw. Learn. Syst. **30**(5), 1381–1391 (2018)

[5] Wang, X., Ma, Y., Cheng, Y., Zou, L., Rodrigues, J.J.: Heterogeneous domain adaptation network based on autoencoder. J Parallel Distrib Comput **117**, 281–291 (2018)

[6] Titouan, V., Redko, I., Flamary, R., Courty, N.: CO-Optimal Transport. In: Proceedings of the NeurIPS, vol. 33, pp. 17559–17570 (2020)

[7] Shu, X., Qi, G.-J., Tang, J., Wang, J.: Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In: ACM Multimedia, pp. 35–44 (2015)

[8] Chen, W.-Y., Hsu, T.-M.H., Tsai, Y.-H.H., Wang, Y.-C.F., Chen, M.-S.:

Transfer neural trees for heterogeneous domain adaptation. In: Proceedings of the ECCV, pp. 399–414 (2016)

[9] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE CVPR, pp. 7167–7176 (2017)

[10] Benjdira, B., Bazi, Y., Koubaa, A., Ouni, K.: Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. Remote Sensing **11**(11), 1369 (2019)

[11] Voreiter, C., Burnel, J.-C., Lassalle, P., Spigai, M., Hugues, R., Courty, N.: A Cycle GAN approach for heterogeneous domain adaptation in land use classification. In: Proceedings of the IEEE IGARSS, pp. 1961–1964 (2020)

[12] Benjdira, B., Ammar, A., Koubaa, A., Ouni, K.: Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks. Applied Sciences **10**(3), 1092 (2020)

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the NIPS, pp. 2672–2680 (2014)

[14] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the ICML, pp. 214–223 (2017)

[15] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016)

[16] Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4058–4065 (2018)

[17] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the ICML, pp. 1180–1189 (2015)

[18] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Proceedings of the NIPS, pp. 343–351 (2016)

[19] Duan, L., Xu, D., Tsang, I.: Learning with augmented features for heterogeneous domain adaptation. In: Proceedings of the ICML, pp. 667–674 (2012)

[20] Courty, N., Flamary, R., Tuia, D.: Domain adaptation with regularized

optimal transport. In: Proceedings of the ECML/PKDD, pp. 274–289 (2014)

[21] Courty, N., Flamary, R., Habrard, A., Rakotomamonjy, A.: Joint distribution optimal transportation for domain adaptation. In: Proceedings of the NIPS, vol. 30 (2017)

[22] Damodaran, B.B., Flamary, R., Seguy, V., Courty, N.: An Entropic Optimal Transport loss for learning deep neural networks under label noise in remote sensing images. Computer Vision and Image Understanding **191**, 102863 (2020)

[23] Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the ECCV, pp. 447–463 (2018)

[24] Yan, Y., Li, W., Wu, H., Min, H., Tan, M., Wu, Q.: Semi-Supervised optimal transport for heterogeneous domain adaptation. In: Proceedings of the IJCAI, vol. 7, pp. 2969–2975 (2018)

[25] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE ICCV, pp. 2223–2232 (2017)

[26] Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE ICCV, pp. 2849–2857 (2017)

[27] Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of the ICML, pp. 1989–1998 (2018)

[28] Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In: Proceedings of the IEEE IGARSS, pp. 5901–5904 (2019)

[29] Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L.: RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. Sensors **20**(6), 1594 (2020)

[30] Neumann, M., Pinto, A.S., Zhai, X., Houlsby, N.: Training general representations for remote sensing using in-domain knowledge. In: Proceedings of the IEEE IGARSS, pp. 6730–6733 (2020)

[31] Tasar, O., Happy, S., Tarabalka, Y., Alliez, P.: SemI2I: Semantically consistent image-to-image translation for domain adaptation of remote sensing data. In: Proceedings of the IEEE IGARSS, pp. 1837–1840 (2020)

[32] Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., Schmitt, M.: SAR-to-Optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. Remote Sensing **11**(17), 2067 (2019)

[33] Ley, A., Dhondt, O., Valade, S., Haensch, R., Hellwich, O.: Exploiting GAN-based SAR to optical image transcoding for improved classification via deep learning. In: Proceedings of the EUSAR 2018, pp. 1–6 (2018)

[34] Saha, S., Bovolo, F., Bruzzone, L.: Building change detection in VHR SAR images via unsupervised deep transcoding. IEEE Trans. Geosci. Remote Sens. (2020)

[35] Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE ICCV, pp. 8050–8058 (2019)

[36] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Proceedings of the NIPS, vol. 30 (2017)

[37] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., *et al.*: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE CVPR, pp. 4681–4690 (2017)

[38] Gómez, P., Meoni, G.: MSMatch: Semisupervised multispectral scene classification with few labels. IEEE J Sel Top Appl Earth Obs Remote Sens **14**, 11643–11654 (2021)

[39] Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y.: Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. JMLR **22**(201), 1–73 (2021)

[40] Sebag, A.S., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L., Altschuler, S.: Multi-domain adversarial learning. In: Proceedings of the ICLR (2019)

[41] Nathan Silberman, P.K. Derek Hoiem, Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)

[42] Mordan, T., THOME, N., Henaff, G., Cord, M.: Revisiting multi-task learning with ROCK: a deep residual auxiliary block for visual detection. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc., ??? (2018). https://proceedings.neurips.cc/paper/2018/file/

7f5d04d189dfb634e6a85bb9d9adf21e-Paper.pdf

[43] Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13, pp. 345–360 (2014). Springer

[44] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)