



HAL
open science

Constrained-HIDA: Heterogeneous Image Domain Adaptation Guided by Constraints

Mihailo Obrenovic, Thomas Lampert, Miloš Ivanović, Pierre Gançarski

► **To cite this version:**

Mihailo Obrenovic, Thomas Lampert, Miloš Ivanović, Pierre Gançarski. Constrained-HIDA: Heterogeneous Image Domain Adaptation Guided by Constraints. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Turin, Italy, janvier 2023, Jan 2023, Turin, Italy. pp.443-458, 10.1007/978-3-031-43424-2_27 . hal-04146352

HAL Id: hal-04146352


<https://hal.science/hal-04146352v1>

Submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constrained-HIDA: Heterogeneous Image Domain Adaptation Guided by Constraints

Mihailo Obrenović^{1,2}^[0000-0003-1272-7437], Thomas Lampert¹^[0000-0002-4911-3941], Miloš Ivanović²^[0000-0002-8974-2267], and Pierre Gañarski¹^[0000-0003-1230-6560]

¹ ICube, University of Strasbourg, Strasbourg, France

² University of Kragujevac, Faculty of Science, Radoja Domanovica 12, 34000 Kragujevac, Serbia
`mobrenovic@unistra.fr`

Abstract. Supervised deep learning relies heavily on the existence of a huge amount of labelled data, which in many cases is difficult to obtain. Domain adaptation deals with this problem by learning on a labelled dataset and applying that knowledge to another, unlabelled or scarcely labelled dataset, with a related but different probability distribution. Heterogeneous domain adaptation is an especially challenging area where domains lie in different input spaces. These methods are very interesting for the field of remote sensing (and indeed computer vision in general), where a variety of sensors are used, capturing images of different modalities, different spatial and spectral resolutions, and where labelling is a very expensive process. With two heterogeneous domains, however, unsupervised domain adaptation is difficult to perform, and class-flipping is frequent. At least a small amount of labelled data is therefore necessary in the target domain in many cases. This work proposes loosening the label requirement by labelling the target domain with must-link and cannot-link constraints instead of class labels. Our method Constrained-HIDA, based on constraints, contrastive loss, and learning domain invariant features, shows that a significant performance improvement can be achieved by using a very small number of constraints. This demonstrates that a reduced amount of information, in the form of constraints, is as effective as giving class labels. Moreover, this paper shows the benefits of interactive supervision — assigning constraints to the samples from classes that are known to be prone to flipping can further reduce the necessary amount of constraints.

Keywords: Deep Learning · Domain Adaptation · Constraints · Remote Sensing.

1 Introduction

Supervised deep learning (DL) models are heavily dependent on the existence of the large-scale labelled datasets. The process of obtaining the reference data is however often very slow and expensive. This is especially the case in the field

of remote sensing (RS), where acquiring the labels requires collecting data in the field from locations that may be complicated to reach due to inaccessibility, natural disasters, armed conflicts etc. Furthermore, the existing reference data may not be reusable for images taken at a later date due to constant changes of the Earth’s surface, such as seasonal and climate changes, deforestation, growth of urban area etc. Since satellites generate huge amount of data on a daily basis, these limitations put the pace of producing reference data far behind the speed of acquiring new data.

Most of the time, existing trained supervised DL models cannot be applied to other dataset as they often generalise poorly. If the conditions during data acquisition differ, there will be a domain shift — a difference between probability distributions — between the datasets. *Domain adaptation* (DA) techniques can help with overcoming this problem. Typically, DA involves learning a model on one data distribution (named *source domain*, typically labelled), and applying it to another, different but related data distribution (called *target domain* typically unlabelled or scarcely labelled) by reducing the shift between domains. Alternatively, both domains can be given to one model at training time, yet with the labels present exclusively or primarily in the source domain.

When there is a small amount of labelled data in a mainly unlabelled target domain, semi-supervised domain adaptation (SSDA) can be employed, and methods for SSDA are specifically developed to take advantage of existing target labels. When there are no labels at all in the target domain, unsupervised domain adaptation (UDA) methods are used. These methods often try to compensate for the absence of supervision in the target domain by using pseudo-labels. Another possible way to incorporate certain knowledge about the target domain, rarely addressed in DA so far, is using the *constraints*.

Constrained clustering is a type of learning where knowledge is provided in the form of constraints rather than labels. The motivation for developing such methods was to improve upon the performance of unsupervised models by providing alternative knowledge about the problem domain in the absence of exact hard labels. Constraints are most often given between the pairs of samples in the form of must-link and cannot-link constraints. There is a growing base of constrained clustering literature, the paradigm is gaining in popularity due to the fact that it does not require classes to be defined (since constraints only act upon pairs of samples) and offers a much weaker form of supervision than labelled samples. It is much easier for an expert to express their preference that two samples should be grouped together (or not), rather than defining absolute labels. This is particularly useful when samples are hard to interpret and interactive, iterative approaches are preferable.

Constraints can be very helpful in DA, especially in situations when there is a huge domain shift. Though existing DA methods are very successful in the field of computer vision (CV), most of them assume RGB images in *both domains* (homogeneous DA). In remote sensing, however, a variety of sensors are used (Figure 1), capturing images of different modalities, with:

- different spatial resolution

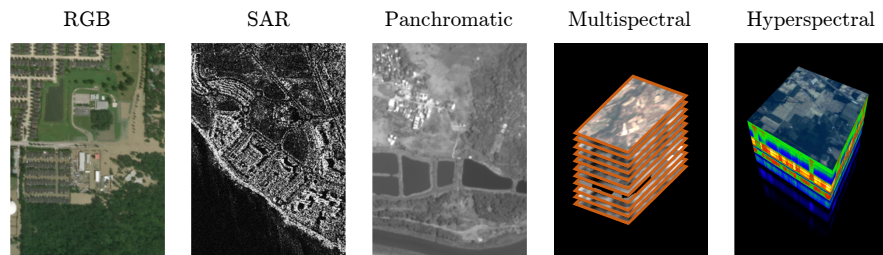


Fig. 1: Different sensors in remote sensing. Images taken from Maxar Open Data Program, Sentinel-1, WorldView-2, EuroSAT and Indian Pines datasets.

- different, non-corresponding channels (referred to as bands in RS), and possibly different numbers of bands.

The domains in RS therefore may not lay in the same space and may have a different dimensionality, increasing still the effective ‘domain shift’. Homogeneous DA approaches cannot be applied to such heterogeneous domains. Instead, heterogeneous domain adaptation (HDA) methods are used.

HDA methods show good performance in semi-supervised settings when a small amount of labelled data is also available in the target domain. The results are, however, much more limited in unsupervised HDA. The problem of class flipping occurs frequently, and it is difficult for the algorithms to associate the same-class samples between domains with such a huge domain shift without any supervision. Many works, therefore, state that a presence of at least a small amount of reference data is required to perform HDA [24, 5].

In this work, we offer a new approach to HDA by introducing constraints to the learning process to reduce the labelling requirement. We hypothesise that HDA methods may greatly benefit from just a few constraints to avoid incorrectly matching classes between domains and that hard labels are not necessary to overcome the problem of large domain shifts. We present a novel method named Constrained-HIDA, a heterogeneous image domain adaptation model for the task of patch classification, in which the knowledge of the target domain is provided in the form of constraints. Constrained-HIDA extracts domain invariant features from two heterogeneous domains, where samples are forced to respect the constraints in the learned representation space through the use of contrastive loss. We show that by using a very small number of constraints, our method can match the performance of semi-supervised HDA methods, thus reducing greatly the amount of information from the target domain needed. Interactive supervision can make the method even more efficient, by assigning constraints to the samples that are known to be difficult to solve. The results show that in this manner less constraints can be used without affecting the performance.

The developed Constrained-HIDA method could be very beneficial to the RS community since the domain adaptation problem is exacerbated by large domain shifts due to the use of different modalities, e.g. RGB, multispectral,

hyperspectral, SAR, LiDAR, panchromatic data etc. The field of application, however, is not limited to RS; different sensors having the same point of view can be found in robotics (depth images, radar), in medical imaging (e.g. CT and MRI) etc. Another benefit is facilitating the labelling process, Hsu et al. state that in many cases, it may be an easier task for a human to provide pair-wise relationships rather than directly assigning class labels [15].

This article is organised as follows: in Section 2, a review of related existing work is given, followed by a description of the proposed Constrained-HIDA model in Section 3. Experimental setup and results are shown in Section 4. Finally, the concluding remarks are given and future work is discussed in Section 5.

2 Literature Review

The emergence of Generative Adversarial Networks (GANs) inspired numerous homogeneous domain adaptation techniques for computer vision. Some of these models aim to extract domain invariant features such as DANN [9, 8] (derived from the original GAN [10]), WDGRL [23] (derived from Wasserstein GAN [1]), DSN [4], etc. Others like CyCADA [14] aim to translate data between domains and are mostly based on image-to-image GAN architectures [32]. It is known, however, that these UDA methods do not scale well to the semi-supervised setting [22]. Methods that specifically aim to use few target labels easily outperform UDA methods [22], motivating the need for specific SSDA algorithms. When target labels are not available at all, many UDA methods fall back on pseudo-labelling [25, 19].

In many cases, even if there are no hard, explicit labels, some background knowledge about the domain is available. This knowledge can be incorporated in the form of instance-level constraints. In constrained clustering, constraints on the pairs of data samples are used to guide the clustering. These constraints can be in the form of must-link or cannot-link, which state that the pair should or should not belong to the same cluster [29]. Zhang et al. propose the deep constrained clustering framework [31] which takes advantage of the benefits of deep learning to learn embedding features and clustering in an end-to-end manner.

Contrastive loss is often used with pair-wise constraints, for example in face recognition [7]. Its simple formula pushes must-link pairs closer in latent space, and cannot-link pairs farther apart. Contrastive learning is therefore a natural choice when learning features for constrained clustering. Hsu et al. used contrastive KL loss on logits of their neural network for constrained clustering [15]. An example of the contrastive loss being used together with clustering in DA is the Contrastive Adaptation Network (CAN) [16].

Constraints found their application in homogeneous DA. Liu et al. [20] pose the problem of unsupervised DA as semi-supervised constrained clustering with target labels missing. The source labels are used to create partition-level constraints. This is especially useful for preserving the structure of domains in multi-source DA. This work, however, does not explore how to use knowledge or preserve the structure in the target domain. Another interesting UDA work

is assigning pairwise pseudo-constraints to samples in the target domain to facilitate the clustering process [17]. In SSDA, soft constraints are used to help tackle the problem of imbalanced classes in medical images [12]. The constraints included, however, are based on labels in the target domain that are already used by the algorithm, with the sole purpose of preserving the structure, and they do not introduce any new knowledge about the target domain.

Heterogeneous domain adaptation is much less present in the literature than homogeneous DA. Most of the heterogeneous DA methods for CV are designed to work with tabular data and focus on adapting between vectorial features extracted from the images of different sizes, such as between SURF and DeCAF [18, 30] and DeCAF and ImageNet features [26], or to adapt from image to text data [24, 5].

The HDA methods applicable to raw-image data of different modalities such as the ones that exist in RS are less frequent. Adversarial Discriminative Domain Adaptation (ADDA) [27] is evaluated on RGB and depth images, but the limitation of the model is that it assumes the same number of channels in the domains. This is also true for Benjdira et al.’s contribution to RS [3], which can be applied to different sensors, but the number of bands must remain the same. Another model for RS by Benjdira et al. [2] can work with a different number of channels, but it is designed for semantic segmentation, and it requires the existence of labelled segmentation masks in the target domain to be used as an intermediate space during the translation process. This approach, therefore, does not extend to classification.

CycleGAN for HDA [28] is a patch classification approach based on image-to-image translation, it is a variant of CycleGAN in which a metric loss, classification loss, and a super-resolution capability are introduced. It is designed to handle RS data of different resolutions, but it may be possible to apply it to domains with different numbers of channels¹ Another work on patch classification of RS data explores the potential of learning domain invariant features in HDA [21]. The paradigm of extracting domain-invariant features is a natural choice for our Constrained-HIDA, as applying the contrastive loss on the constraints is straightforward in the learned common latent feature space.

To the best of our knowledge, there are no other works on using the constraints in HDA, thus making Constrained-HIDA the first such method.

3 Methodology

In this section, the Constrained-HIDA model will be described. Constrained-HIDA extracts deep domain-invariant features from two heterogeneous domains. The learning of a common latent space of invariant features is guided by cross-entropy loss on available (source domain) labels for class discrimination, Wasserstein loss is used to reduce the distance between domains, and contrastive loss on constraints helps to preserve the correct local structure of domains.

¹ It is not clear in the original article [28] if the method was evaluated on the same or different numbers of channels.

Let $X^s = \{x_i^s\}_{i=1}^{n^s}$ be a labelled source dataset of n^s samples from the domain \mathcal{D}_s following the data distribution \mathbb{P}_{x^s} with labels y_i^s , and let $X^t = \{x_j^t\}_{j=1}^{n^t}$ be an unlabelled target dataset of n^t samples from the domain \mathcal{D}_t following the data distribution \mathbb{P}_{x^t} . Constrained-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.

A certain amount of domain knowledge is given in the form of pairwise constraints of two types — must-link and cannot-link. These constraints can be attached to two samples coming from the same domain or from different domains. In this paper, the focus is on the case where there are only inter-domain constraints. The set of constrained samples X^c is usually a small fraction of the whole dataset $X = X^s \cup X^t$. Let $\mathcal{C}^=$ be a set of must-link constraints $C_i^=$, where $C_i^= = (x_{i1}, x_{i2}) \in \mathcal{C}^=$ implies that x_{i1} and x_{i2} should belong to the same cluster/class, and let \mathcal{C}^\neq be a set of cannot-link constraints, where $C_j^\neq = (x_{j1}, x_{j2}) \in \mathcal{C}^\neq$ implies that x_{j1} and x_{j2} should belong to the different cluster/class, $\mathcal{C}^=, \mathcal{C}^\neq \subset X^s \times X^t$, $\mathcal{C}^= \cap \mathcal{C}^\neq = \emptyset$.

Constrained-HIDA’s architecture is presented in Figure 2 and consists of 5 neural network components: 3 feature extractors, a domain critic, and a class discriminator, with the addition of contrastive loss over constraints on extracted features. To work with the data coming from two different spaces, possibly of different input sizes, two different input branches are needed. Therefore, Constrained-HIDA has two separate feature extractors — $FE_s : \mathcal{X}^s \rightarrow \mathbb{R}^{d_1}$ and $FE_t : \mathcal{X}^t \rightarrow \mathbb{R}^{d_1}$ — these have the task of bringing 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 $FE_i : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ is employed to model the similarity of the data domains and to extract domain invariant features — $h^s = FE_i(g^s)$ and $h^t = FE_i(g^t)$.

Wasserstein distance is used to measure the distance between domains. This metric is calculated by solving the optimal transport between two probability distributions μ and ν . Since this is computationally expensive, the domain critic $DC : \mathbb{R}^{d_2} \rightarrow \mathbb{R}$ is trained to approximate it instead [1, 23], accelerating the training process. The loss of this component is defined such that

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

In order to calculate the empirical Wasserstein distance, Eq. (1) needs to be maximised, therefore the domain critic component is trained by solving

$$\max_{\theta_{dc}} (\mathcal{L}_{wd} - \gamma \mathcal{L}_{grad}), \quad (2)$$

where θ_{dc} are the domain critic’s weights and $\gamma \mathcal{L}_{grad}$ is a regularisation term enforcing the Lipschitz constraint. When training our domain critic [23], this

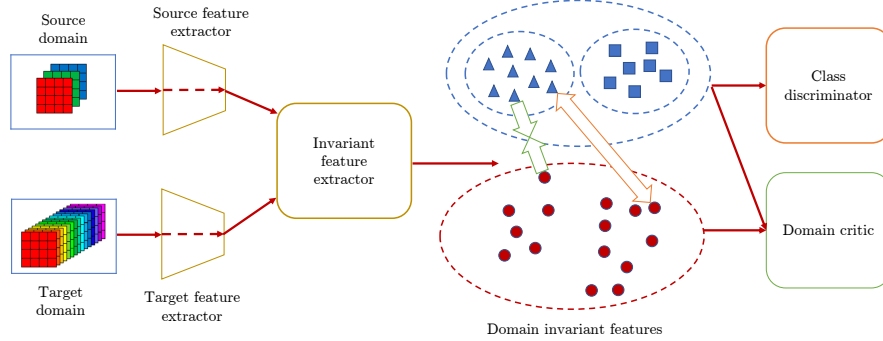


Fig. 2: Overview of the proposed method. Features of the labelled source domain samples are shown in blue, with triangles and squares representing different classes, while features of the unlabelled target domain samples are shown as red circles. Must-link constraints force samples to move towards each other (green arrows), and cannot-link constraints force samples to move apart (orange arrow).

regularisation term amounts to

$$\mathcal{L}_{grad}(\hat{h}) = \left(\left\| \nabla_{\hat{h}} DC(\hat{h}) \right\|_2 - 1 \right)^2, \quad (3)$$

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 [11].

The class discriminator $C : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^c$ (where c is the number of classes) is trained on the extracted features of the labelled source samples (h^s, y^s) (and does not use the unlabelled target data). If labels y^s are one-hot encoded, the cross-entropy classification loss is used, such that

$$\mathcal{L}_c(h^s, y^s) = -\frac{1}{n^s} \sum_{i=1}^{n^s} \sum_{k=1}^c y_{i,k}^s \log C(h_i^s). \quad (4)$$

Contrastive loss is applied to the extracted features of the constrained pairs of samples. Let $I^=$ be an indicator function equal to one when the pair (x_i, x_j) is under must-link constraint, or equal to zero otherwise. Let also I^{\neq} be an indicator function for cannot-link constraints. The contrastive loss is defined such that

$$\mathcal{L}_{con} = \sum_{i,j} I^=(x_i, x_j) \|h_i - h_j\|_2^2 + I^{\neq}(x_i, x_j) \max\left(0, m - \|h_i - h_j\|_2\right)^2 \quad (5)$$

where h_i, h_j are the extracted features of samples x_i, x_j , and m is a threshold that prevents the cannot-link loss from moving towards infinity, i.e. the features of samples under a cannot-link constraint are limited to be a distance of m apart.

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

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

$$\min_{\theta_{fe}, \theta_c} \left\{ \mathcal{L}_c + \lambda_1 \max_{\theta_{wd}} [\mathcal{L}_{wd} - \gamma \mathcal{L}_{grad}] + \lambda_2 \mathcal{L}_{con} \right\}. \quad (6)$$

where λ_1 and λ_2 are the weights of the contrastive loss and Wasserstein loss respectively.

4 Experimental results

4.1 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 3):

- NWPU-RESISC45 [6] (high-resolution aerial RGB images extracted from Google Earth) — dense residential, forest, freeway, industrial area, lake, meadow, rectangular farmland, and river.
- EuroSAT [13] (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 problem to be solved is patch classification, with each patch having a single label. 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.

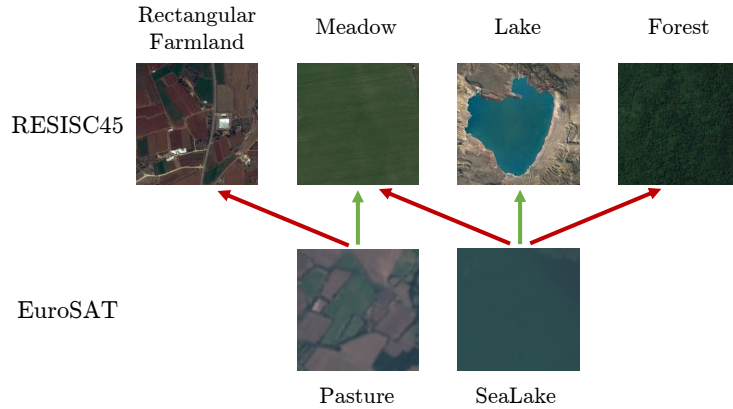


Fig. 4: Examples of issues that can arise during transfer learning between RESISC45 and EuroSAT. Green arrows show which samples should be aligned, and red arrows show which samples tend to be aligned, but should not.

The EuroSAT dataset covers 34 European countries and also consists of data from throughout the year. Both datasets, therefore, have in-domain temporal and geographic variability, making even the in-domain problem of classification very difficult.

Transfer learning brings another level of difficulty, especially with the huge domain shift that exists in the presented datasets. Figure 4 visualises some classes that tend to be misaligned between domains. The Lake class in RESISC45 shows the entire lake with the surrounding area, whereas in EuroSAT only a patch of water is shown, making it more similar to meadow and forest which also present uniform colours. In the following experiments, we show that a huge improvement in performance can be achieved by introducing must-link and cannot-link constraints between such misaligned patches.

One advantage of the proposed Constrained-HIDA approach is that information in all channels can be used. The information provided by non-RGB channels can be discriminative but is often neglected. For example, the multispectral EuroSAT data contain, aside from the visible RGB bands, near-infrared (NIR), short-wave infrared (SWIR) and red edge bands etc.

The datasets are split into the 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.

4.2 Implementation details

Constrained-HIDA is a convolutional architecture. (see Figure 5 for details). The feature extractor for RESISC45 consists of two convolutional layers with 16 and 32 filters respectively. Each convolutional layer is followed by 4×4 max-pooling.

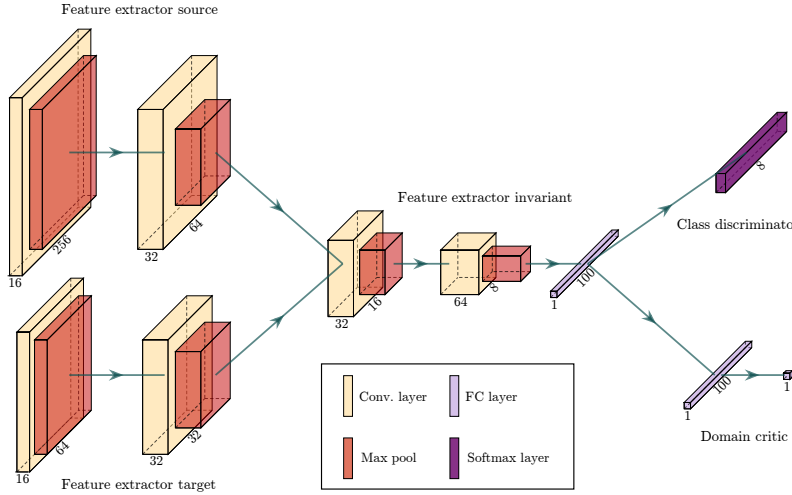


Fig. 5: The architecture of the proposed Constrained-HIDA model, specifically used for the case when the source dataset is RESISC45 and the target dataset is EuroSAT.

The feature extractor for EuroSAT is the same, except that it has 2×2 max-pooling after every convolutional 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) 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_1 is 0.1, and the contrastive loss weight λ_2 is 0.3. The Adam optimiser is used.

The input data is standardised per channel so that each channel has a mean of 0 and a standard deviation of 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 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. In every

batch, there are always 4 pairs of source-target samples with a constraint, either must-link or cannot-link. The model is trained for 40 epochs.

The convolutional architecture used for Constrained-HIDA is not rigorously optimised but was found through initial experiments. The hyper-parameters related to the domain critic, as well as learning rates, optimiser, and loss weights, are taken from the WDGRL method [23]. Data augmentation was chosen based on remote sensing domain experience. Increasing batch size or percentage of constrained pairs per batch did not improve performance further.

The code to reproduce the Constrained-HIDA experiments presented in this article is available online².

4.3 Comparison methods

To the best of our knowledge, there are no other HDA works on using constraints instead of labels in the target domain. Our method is therefore compared with HDA methods for image data in an unsupervised and semi-supervised setting. The first comparison method is CycleGAN for HDA [28], which can be used in both UDA and SSDA, a method tailored for RS and for data with different spatial resolutions. We will denote unsupervised and semi-supervised variants of the method as CycleGAN for U-HDA and CycleGAN for SS-HDA. We further compare with an unsupervised version of our method without using any constraints or any labels in the target domain (denoted U-HIDA), and with a semi-supervised version that uses labels in the target domain, but not constraints nor contrastive loss (denoted SS-HIDA). Semi-supervised methods are evaluated in the situation where 1.25% of labelled target data is available (5 labelled samples per class, 40 in total).

Our Constrained-HIDA is evaluated on a range of different amounts of constraints (40, 80, 160, 320, and 480 constrained pairs), where the ratio of must-link and cannot-link constraints is 1 : 7. Constraints are generated by taking pairs of samples and if their ground truth label is the same, a must-link constraint is created between them, if their ground truth labels differ, a cannot-link constraint is instead added. This is repeated until the correct number and ratio of constraints are found. Each constrained pair has one sample from the source and one from the target domain, they are all therefore inter-domain. No intra-domain constraints were used. Note that in semi-supervised DA comparison methods, the existence of 5 labels per class in the target domain, with 8 classes and 400 samples per class in training sets of each domain, implies the existence of 128,000 inter-domain constraints, and 780 intra-domain constraints in the target domain — a number far greater than what our method is using! For demonstration purposes, we also show the results of our method with all the inter-domain constraints implied by 40 labels in the target domain, (i.e. 16,000 must-link and 112,000 cannot-link), without using any intra-domain constraints, and without directly using any target labels for the training, relying solely on the contrastive loss over constraints in the target domain.

² <https://github.com/mihailoobrenovic/Constrained-HIDA>

	R → E	E → R
CycleGAN for U-HDA	18.48 (8.00)	16.82 (5.74)
U-HIDA	13.61 (11.33)	17.77 (9.37)
Constrained-HIDA 40 constraints	35.52 (7.70)	33.29 (13.59)
Constrained-HIDA 80 constraints	39.09 (10.02)	40.54 (9.43)
Constrained-HIDA 160 constraints	48.59 (7.46)	49.00 (7.17)
Constrained-HIDA 320 constraints	64.68 (3.68)	56.13 (7.12)
Constrained-HIDA 480 constraints	65.27 (2.53)	59.37 (5.48)
Constrained-HIDA all constraints for 40 labels	69.34 (3.60)	63.71 (2.12)
CycleGAN for SS-HDA 40 labels	41.57 (9.20)	47.29 (1.53)
SS-HIDA 40 labels	66.14 (2.92)	62.68 (3.24)

Table 2: Accuracy of the proposed Constrained-HIDA model with different numbers of constraints, UDA methods are shown as lower baselines and SSDA methods are shown as upper baselines. Standard deviations are shown in parentheses.

4.4 Results

The overall accuracy of our and all the comparison methods with RESISC45 as source and EuroSAT as target ($R \rightarrow E$) and vice-versa ($E \rightarrow R$) are shown in Table 2.

For the $R \rightarrow E$ case, the results show that our Constrained-HIDA almost doubles the performance of unsupervised CycleGAN for HDA with as few as 40 constraints, with even higher gains over the unsupervised version of the model (U-HIDA). As more constraints are added, the better Constrained-HIDA performs. With 160 constraints, it already gains 7% over semi-supervised CycleGAN for HDA that uses 40 labels in the target domain. From 320 constraints and on, the results become comparable to the semi-supervised version of our model SS-HIDA.

For the $E \rightarrow R$ case, the findings are similar. Constrained-HIDA with 40 constraints has around 2 times stronger performance than the lower baselines. With 160 constraints it already outperforms semi-supervised CycleGAN for HDA by around 2%, with the gain growing as more constraints are added. When using 480 constraints, the results of Constrained-HIDA become comparable to SS-HIDA.

Constrained-HIDA using all of the inter-domain constraints implied by 40 labels in the target domain (i.e. 120,000 constraints) even outperforms SS-HIDA trained with 40 target labels in both cases — by more than 3% in the $R \rightarrow E$ case, and around 1% in the $E \rightarrow R$ case. This is a very interesting finding, having in mind that the classifier in Constrained-HIDA is trained only on source samples and that only the contrastive loss and Wasserstein loss were affected by target samples, while the classifier of SS-HIDA was trained with all available labelled data including from the target domain. This implies that it might be more important to align the structure of the target domain with the source domain than to use (a small number of) hard target labels.

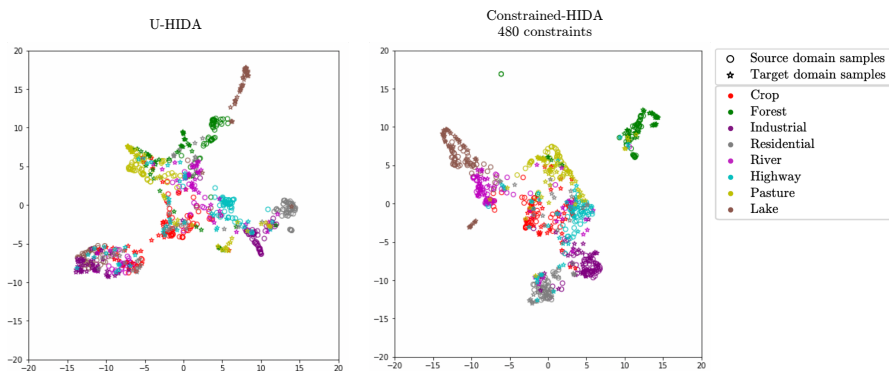


Fig. 6: PaCMAP visualisation of U-HIDA and Constrained-HIDA features in the $R \rightarrow E$ case.

It should be noted, however, that in the case of Constrained-HIDA using 320 and 480 constraints, there are 40 and 60 must-link constraints respectively. This means that there are 40 and 60 target samples, each associated with a source sample that is labelled. One could argue that this indirectly brings information about the labels to the target domain. This information is however still weaker than a label in our experiments. The target labels are used when training the classifier in SSDA comparison methods and directly introduce the information equivalent to a huge number of must-link and cannot-link constraints, both inter-domain and intra-domain. On the other hand, in our experiments, Constrained-HIDA only applies the contrastive loss to inter-domain constraints. Furthermore, the numbers of 320 and 480 constraints still represent only 0.25%, and 0.375% respectively of the total number of 128,000 constraints implied by 40 labels in the target domains.

As shown in Figure 6, Constrained-HIDA learns better discriminative features compared to U-HIDA. In the absence of constraints, U-HIDA tends to flip classes, for example, many target samples of the crop class are matched with the river class, a lot of forest class samples are matched with pasture etc. In contrast, Constrained-HIDA better matches classes and the spread between the domains is reduced such that the overlap is more consistent, explaining the increase in performance observed in Table 2.

Interactive supervision can further reduce the need for constraints. If constraints are manually created with prior knowledge, on samples representing classes that are known to be problematic, fewer constraints can be more effective. By identifying and adding constraints to a certain number of target samples that are misclassified by unsupervised HIDA in the $R \rightarrow E$ case, 8 such constraints are sufficient for Constrained-HIDA to achieve an accuracy of 40.92%, and 80 gives 55.36%, which is 15% better than when using the same

number of randomly chosen constraints, and almost 7% better than when using 160 randomly chosen constraints. This means that number of constraints can be more than halved by carefully choosing them without affecting performance. It should be noted, however, that in these experiments the ratio of must-link and cannot-link constraints is 1 : 1. Still, these initial results show that carefully chosen constraints can provide strong results with very little supervision and that interactive supervision is a very interesting future research direction.

5 Conclusions

This article has proposed a novel approach to heterogeneous image domain adaptation using constraints named Constrained-HIDA. To the best of our knowledge, this is the first such method using constraints instead of labels in a semi-supervised setting. The results show that with a very small number of constraints, Constrained-HIDA strongly outperforms UDA methods, and has comparable results with SSDA methods, even outperforming them when using an equivalent amount of information. This shows that replacing labels with constraints could reduce the need for supervised information in the target domain and could facilitate the job of annotating experts for whom providing constraints might be easier and more natural than providing hard labels.

In the future, Constrained-HIDA could be further improved by introducing pseudo-labels or pseudo-constraints, with e.g. constrained clustering; this could enrich the information about the target domain. The prospect of interactive learning is another interesting direction, allowing the user to put constraints on the examples misclassified by the model, in an iterative manner, could additionally decrease the number of constraints needed. The method could also be evaluated in homogeneous DA, on domains coming from the same input space, but with a huge domain shift.

Acknowledgements 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 and ICube’s internal project funding (RL4MD).

References

1. Arjovsky, M., et al.: Wasserstein generative adversarial networks. In: ICML. pp. 214–223 (2017)
2. 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)
3. Benjdira, B., et al.: Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing* **11**(11), 1369 (2019)

4. Bousmalis, K., et al.: Domain separation networks. In: NIPS. pp. 343–351 (2016)
5. Chen, W.Y., et al.: Transfer neural trees for heterogeneous domain adaptation. In: ECCV. pp. 399–414 (2016)
6. Cheng, G., et al.: Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* **105**(10), 1865–1883 (2017)
7. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189 (2015)
9. Ganin, Y., et al.: Domain-adversarial training of neural networks. *JMLR* **17**(1), 2096–2030 (2016)
10. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
11. Gulrajani, I., et al.: Improved training of Wasserstein GANs. In: NIPS. vol. 30 (2017)
12. Harada, S., Bise, R., Araki, K., Yoshizawa, A., Terada, K., Kurata, M., Nakajima, N., Abe, H., Ushiku, T., Uchida, S.: Cluster-guided semi-supervised domain adaptation for imbalanced medical image classification. arXiv preprint arXiv:2303.01283 (2023)
13. Helber, P., et al.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE JSTARS* **12**(7), 2217–2226 (2019)
14. Hoffman, J., et al.: CyCADA: Cycle-consistent adversarial domain adaptation. In: ICML. pp. 1989–1998 (2018)
15. Hsu, Y.C., Kira, Z.: Neural network-based clustering using pairwise constraints. arXiv preprint arXiv:1511.06321 (2015)
16. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4893–4902 (2019)
17. Li, J., Li, G., Shi, Y., Yu, Y.: Cross-domain adaptive clustering for semi-supervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2505–2514 (2021)
18. Li, J., et al.: Heterogeneous domain adaptation through progressive alignment. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(5), 1381–1391 (2018)
19. Liu, H., Wang, J., Long, M.: Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems* **34**, 22968–22981 (2021)
20. Liu, H., Shao, M., Ding, Z., Fu, Y.: Structure-preserved unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering* **31**(4), 799–812 (2018)
21. Obrenovic, M., Lampert, T., Monde-Kossi, F., Gançarski, P., Ivanović, M.: Sshida: Semi-supervised heterogeneous image domain adaptation. In: MACLEAN: MACHine Learning for EArth ObservatioN Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) (2021)
22. Saito, K., et al.: Semi-supervised domain adaptation via minimax entropy. In: ICCV. pp. 8050–8058 (2019)
23. Shen, J., et al.: Wasserstein distance guided representation learning for domain adaptation. In: AAAI. pp. 4058–4065 (2018)
24. Shu, X., et al.: Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In: ACM Multimedia. pp. 35–44 (2015)

25. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33**, 596–608 (2020)
26. Titouan, V., et al.: CO-Optimal Transport. In: *NeurIPS*. vol. 33, pp. 17559–17570 (2020)
27. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: *CVPR*. pp. 7167–7176 (2017)
28. Voreiter, C., et al.: A Cycle GAN approach for heterogeneous domain adaptation in land use classification. In: *IGARSS*. pp. 1961–1964 (2020)
29. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al.: Constrained k-means clustering with background knowledge. In: *ICML*. vol. 1, pp. 577–584 (2001)
30. Wang, X., et al.: Heterogeneous domain adaptation network based on autoencoder. *J Parallel Distrib Comput* **117**, 281–291 (2018)
31. Zhang, H., Zhan, T., Basu, S., Davidson, I.: A framework for deep constrained clustering. *Data Mining and Knowledge Discovery* **35**, 593–620 (2021)
32. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *ICCV*. pp. 2223–2232 (2017)