Unsupervised Domain Adaptation for Remote Sensing Data Classification Model Transfer

Melanie Böge, Dimitri Bulatov, Edwin Deisling, Gisela Häufel, Kevin Qiu

Fraunhofer IOSB Ettlingen, Germany - (melanie.boege@iosb.fraunhofer.de)

Keywords: Histogram matching, Canonical Correlation, DeepLab, Random Forest

Abstract

In this paper, we explore the application of domain adaptation techniques for semantic land cover segmentation using aerial remote sensing data. We leverage canonical correlation and histogram matching to facilitate the transfer of knowledge from pre-trained classification models to new datasets without the need for additional labeled data. Specifically, we perform Canonical Correlation to align feature distributions between the source and target domains and Histogram Matching to enhance the correspondence of pixel distributions across datasets. The effectiveness of these domain adaptation techniques is assessed through improvements in semantic segmentation performance of the Random Forest and DeepLabV3+ classifiers on German city datasets. Our results indicate a substantial increase in segmentation accuracy when using domain adaptation methods. Furthermore, we examine the role of elevation data, represented by Normalized Digital Surface Models (NDSM), which enhances segmentation performance on unseen datasets. These findings underscore the efficacy of domain adaptation and the value of elevation data in remote sensing classification, particularly in dynamic environments where models encounter new datasets.

1. Introduction

In recent years, the rapid advancement of machine learning has led to the development of powerful classification and semantic segmentation models capable of achieving remarkable performance on various tasks. However, these models are often trained on specific datasets, which can limit their effectiveness when applied to different data sources. This phenomenon, known as domain shift, occurs in the case when the distribution of the training data differs significantly from that of the target data. Consequently, models that perform well in one domain may exhibit degraded performance in another, leading to challenges in real-world applications.

Domain adaptation (DA) has emerged as a crucial area of research aimed at bridging this gap between training and target domains. By leveraging knowledge from the source domain, DA techniques facilitate the transfer of classification models to new environments, enabling them to generalize better to unseen data distributions. This transfer is crucial in numerous fields, such as medical imaging, speech recognition, and remote sensing classification tasks, where labeled data in the target domain may be scarce or expensive to obtain.

As shown in the next section, most DA methods rely on adapting features derived from the target domain images rather than modifying the classification models themselves. We assume the existence of a pre-trained model that is optimized for a source dataset while adapting the target dataset to align with the training source data. This process enhances the model's applicability to a different dataset, allowing it to maintain its effectiveness even when faced with variations in data characteristics. As we continue to confront challenges posed by diverse data conditions in practical applications, it is imperative to explore and develop effective domain adaptation strategies that successfully utilize pre-trained models across varying data sources.

Domain adaptation is significant in high-resolution airborne remote sensing (RS) because the images can vary significantly due to different acquisition conditions, atmospheric effects, or changes in the observed objects. Even when elevation data is available, occlusions and intense variations of appearances within the same class are abundant in urban terrain. At the same time, access to labeled data is costly.

In this paper, we focus on methods to align the feature distributions between the source (S) and target (T) domains for RS applications in urban areas. We seek methodologies that are independent of labeled data and can effectively handle various remote sensing (RS) data modalities. While some advanced techniques demand extensive training data, such resources are often limited in many remote sensing contexts. Deep learning approaches, particularly Generative Adversarial Networks (GANs), require substantial computational power and can be challenging to train. Additionally, there is a risk of overfitting with these powerful methods, especially when working with smaller datasets, which should be avoided. In contrast, traditional methods mitigate these risks and provide the added benefit of producing results that are generally easier to understand and interpret. This clarity is crucial in applications where human decision-making relies on these outcomes.

Two methods, among others, fulfill these requirements: Canonical Correlation Analysis (CCA) can effectively align features across domains by maximizing the correlation between the source and target feature sets. A further simple and effective tool of domain adaptation techniques is histogram matching (HM), which effectively aligns the intensity distributions of source and target images. We employ the conventional Random Forest (RF) classifier and a Deep-learning-based semantic segmentation method (DeepLabV3+) for land cover classification, adjusted to take multi-modal data as input.

2. Related Work

After reading a few survey papers [Tuia et al., 2016, Wang and Deng, 2018, Xu et al., 2022], we could conclude that there are four main categories of methods for DA, whereby the boundaries between them are often fuzzy, so that one can mention a fifth category, denoted hybrid methods.

The first family of DA methods is based on the selection of invariant features, which exhibit a certain robustness regarding their change from the source to the target domain. According to [Tuia et al., 2016], a trade-off must be found here between the domain shift between the features in the subset and their discrimination capability regarding the separation of classes [Bruzzone and Persello, 2009].

Secondly, selection can affect not only features but also the data points. In this case, we speak about instance weighting. In [Cai et al., 2024], the authors cluster target domain samples based on entropy, which helps dynamically produce more accurate pseudo-labels and align clean subdomains with noisy ones. Techniques like the self-adaptive pseudo-label assigner (SPA) adjust class-wise confidence thresholds to generate high-quality pseudo-labels, which are crucial for reducing domain bias [Han et al., 2024]. The data-invariant samples can also be synthetic [Cai et al., 2023]. For example, the network SRDA-Net [Wu et al., 2022] simultaneously performs super-resolution and domain adaptation tasks, addressing resolution discrepancies and enhancing the segmentation accuracy, particularly for small objects.

Thirdly, adjusting the classifier opens the way to semi-supervised classification or active learning. Either way, the parameters of a pre-trained learner are fine-tuned using a few labeled samples of a new dataset. Adding new data references is costly, so they must be carefully selected. In their survey, [Tuia et al., 2016] mentioned a DA technique based on jointly considering the information contained in the source and target domains within a Bayesian framework [Bruzzone and Prieto, 2002]. The work of [Huang et al., 2024] represents a more recent example of this category of methods. The Joint Distribution Adaptive-Alignment Framework (JDAF) combines marginal and conditional distribution alignment to dynamically update and align feature representations, enhancing the model's adaptive performance. The authors optimize the parameters of a backbone network so that the method is deep-learning-based. The method called Class Centroid Alignment aligns class centroids between domains by moving target domain samples toward the source domain, making the data distributions more similar and improving classifier performance [Zhu and Ma, 2016].

The last and most popular category of methods is to perform domain transfer while keeping the classifier and the sample indices fixed. For unsupervised domain adaptation, traditional methods include multidimensional histogram matching, data alignment with PCA [Inamdar et al., 2008], kPCA (kernel Principal Component Analysis, [Nielsen and Canty, 2009]), or CCA (Canonical Correlation Analysis) [Volpi et al., 2015]. Transformation into new, latent space is often implicitly combined with de-noising, opening the way to sparse representations and low-rank reconstructions, which are meant to avoid the influence of outliers and noise in the source domain samples. Although many representation-based domain adaptation methods attempt to find a total transformation matrix for all samples in the source domain and ignore the individual changes in each class, [Shi et al., 2015] attempt to find new representations for samples in different classes in the source domain by multiple linear transformations. Furthermore, manifold alignment methods apply the maximum mean discrepancy method to the nonlinear projections of the feature vectors to adjust the distributions of source and target domains. Three types of manifold learning, according to [Liu et al., 2021], are locally linear embeddings, Laplacian Eigenmaps, and local tangent space alignment. All this is done to maintain the characteristics of the original data structure in the transformed domain for which [Liu et al., 2021] have proposed a manifold regularization framework. Deep-learning-based approaches have been subdivided into generative, adversarial, and self-training [Xu et al., 2022]. Generative methods presuppose (deep-learning-based) transformation from target to source [Wittich and Rottensteiner, 2021] or vice versa. In the first case, the model of S is applied to the modified data of T, and in the second case, it is trained on the modified data of S and then applied to T. The typical losses to be minimized often result from cyclicity (transform from one domain to another and back must match the first). However, other penalty terms, such as "visiting loss", intended to ensure that pixels from the other domain are visited wherever possible, have also been proposed. Again, here, the authors of [Wittich and Rottensteiner, 2021] balance the classification accuracy and generation quality by training the transferred samples with the same classifier as the original ones. In the second case, the transfer affects feature vectors, and new images are not generated. As a frequently cited example, [Ganin and Lempitsky, 2015] introduced a technique called domain-adversarial neural network architecture that consists of a feature extractor, a label predictor, and a domain classifier and minimizes the so-called domain confusion loss within the training process. The domain classifier's goal is to distinguish between features from the source and target domains, while the feature extractor aims to confuse the domain classifier. This adversarial training encourages the model to learn domain-invariant features. Here, advanced concepts like attention and self-attention can be applied. For example, a multi-level attention mechanism, which includes a feature level attention generated by shallow features and an entropy level attention produced by a deep discriminative feature, was presented in [Zheng et al., 2020]. In this regard, the fastest progress is achieved by self-learning algorithms, often based on visual transformer outputs that have been pretrained on massive data sets. However, integrating elevation data is problematic because typical computer vision methods only work with RGB data. To summarize, we will stick to this last category and present two methods for domain adaptation that do not rely on labeled examples.

3. Methodology

In this work, we perform DA using multi-modal RS images of two close-range datasets, described in more detail in Section 4. We consider the raw image channels (red, green, blue, and near-infrared (NIR)), the relative elevation, and the so-called planarity map (PMAP). The relative elevation, denoted as Normalized Digital Surface Model (NDSM) in RS, in the case it is not provided, can be retrieved from the absolute elevation and the ground model using one of the numerous methods previously developed, such as [Bulatov et al., 2012, Piltz et al., 2016]. PMAP assesses how likely a neighborhood of a single pixel can be approximated by a plane. It is computed from the eigenvalues of a structural tensor [Gross and Thoennessen, 2006]. In our implementation, additional robustness has been proposed for a given DSM rather than a purely 3D point cloud. Hereby, pixels with elevation jumps from the structural tensor formation are ignored. To take into account texture information for the superpixel-based Random Forest, we retrieved the responses of the rotationally-invariant MR8 filter bank of Varma and Zisserman [Varma and Zisserman, 2005] from the intensity and NDSM image.

3.1 Domain Adaptation Methods

3.1.1 Canonical Correlation Analysis CCA is important for DA in image analysis. It aligns the feature spaces of source and target domains by maximizing correlations between their feature representations, reducing domain discrepancies, and improving model performance. CCA is a statistical method for exploring the relationships between two multivariate datasets by identifying linear combinations of variables that maximize their correlation. Given two datasets $S \in \mathbb{R}^{n\times p}$ and $T \in \mathbb{R}^{n\times q}$, where n is the number of observations and p and q are the number of variables in each dataset, CCA seeks to find linear transformations $U = SW_s$ and $V = TW_t$ that maximize the correlation between the resulting canonical variables U and V. Hereby, W_s and W_t are weight matrices that define the linear combinations for S and T, respectively.

The canonical correlation can be mathematically expressed as:

$$\rho = \frac{\text{Cov}(U, V)}{\sqrt{\text{Var}(U) \cdot \text{Var}(V)}}, \tag{1}$$

and the goal of CCA is to maximize ρ subject to the constraints:

$$W_s^\top \Sigma_{ss} W_s = I \text{ and } W_t^\top \Sigma_{tt} W_t = I, \tag{2}$$

where Σ_{ss} and Σ_{tt} are the covariance matrices of S and T and I the identity matrix. By $S' = (U - \text{mean}(U) + \text{mean}(V)) \cdot W_t^{-1}$ we compute the adapted data. Note that the CCA can only be applied to random variables of the same size. In this paper, we replicate the shorter dataset to match the larger one, ensuring that no image information is lost, even at the cost of some redundant information.

3.1.2 Histogram Matching HM is a valuable tool in image processing that allows the adjustment of image histograms to achieve desired visual characteristics. This enhances the interpretability and analysis of images in various applications. In a nutshell, it is a technique in image processing used to adjust the histogram of a source image to resemble the histogram of a target image. The primary goal of this method is to modify the source image to enhance its visual quality or to standardize the appearance of images captured under different lighting conditions or environments.

The algorithm for HM involves three steps. First, for each feature of both the source and the target image, we compute the probability density function p of each pixel value r_j among all n pixels and the corresponding cumulative distribution function (CDF)

$$\text{CDF}(r_k) = \sum_{j=0}^{k} p(r_j),\ k \in \{\text{pixel values}\}, \tag{3}$$

represented by a cumulative histogram. Second, a mapping M is created from the intensity values of the source image to those of the target image based on the normalized CDFs, such as $\text{CDF}_{\text{source}}(s) \leq \text{CDF}_{\text{target}}(t),\ M(s) = t$. Finally, this mapping is applied to the source image to generate the matched output image. Therefore, the intensity values in the source image are replaced with the mapped values according to the mapping function.

One interesting theoretical question deals with filter banks, such as [Varma and Zisserman, 2005]. Its filters are primarily based on gradient operators (G), and performing both $M(G(s))$ and $G(M(s))$ has certain advantages and disadvantages. The former strategy presupposes applying HM directly to the output of the filter banks of S and T. This approach could work if the derivatives ∇S and ∇T have similar statistical distributions. However, derivatives often amplify noise and may not have distributions as smooth or comparable as the original features. Applying HM directly on derivatives might yield less reliable results if the derivatives are noisy or irregular. Hence, we opt for the latter strategy, which preserves the statistical relationship between S and T before gradient calculation.

3.2 Classification Models

3.2.1 Random Forest RF, [Breiman, 2001] is an ensemble learning method widely used for classification (and regression) tasks. It operates by constructing multiple decision trees during training and outputting the classification probability p_c as the relative frequency of individual trees.

$$p_c(s) = n_c(s)/n, \tag{4}$$

$n_c(s)$ is the number of trees voting for the class of instance s (pixel or superpixel) to be c, and n is the total number of trees. The class of s corresponds to the most frequent tree.

Random Forest is a quite popular conventional classifier due to its ease of use, minimal parameter tuning, ability to handle missing values and categorical features effectively. This ensemble approach enhances predictive accuracy and controls overfitting by increasing the minimum leaf size parameter, making RF particularly robust in handling large datasets with high dimensionality, including those with multi-modal data.

In our application, the number of trees was set to 20 while the minimum leaf size parameter was set to 4. These are default values for the configurations mentioned in Section 4. Also, we accelerated the computation by training and evaluating the RF on superpixels retrieved using the SLIC algorithm [Achanta et al., 2012], which is quite fast and easy-to-use.

3.2.2 DeepLab Convolutional Neural Networks (CNNs) are a significant advancement in image segmentation. Unlike RF, which relies on handcrafted features, CNNs automatically learn hierarchical features from the data, capturing complex spatial relationships. In computer vision, the input modalities are usually limited to RGB images. However, additional modalities exist in RS, like more spectral bands or elevation information. We therefore follow to [Qiu et al., 2022] and extend a supervised state-of-the-art semantic segmentation model, DeepLabV3+ [Chen et al., 2018], based on a ResNet101 [He et al., 2016] encoder, with a second input branch. While the first branch processes RGB input, the second branch processes a multi-band image consisting of NDVI, PMAP, and NDSM. NDVI is the Normalized Vegetation Index and is calculated using the red and NIR channel. The features f_1 and f_2 of the two input branches, as shown in Eq. (5) and in Fig. 1 below, are fused by convex combination after the first residual block.

$$f = \alpha f_1 + (1 - \alpha) f_2 \quad \text{where } 0 \leq \alpha \leq 1. \tag{5}$$

We train and evaluate the two-branch network, which we will denote as DeepLab for the sake of brevity, with $\alpha = 0$ (only the second branch contributes with features f_2), $\alpha = 0.5$ (both branches contribute equally), and $\alpha = 1$ (the original version

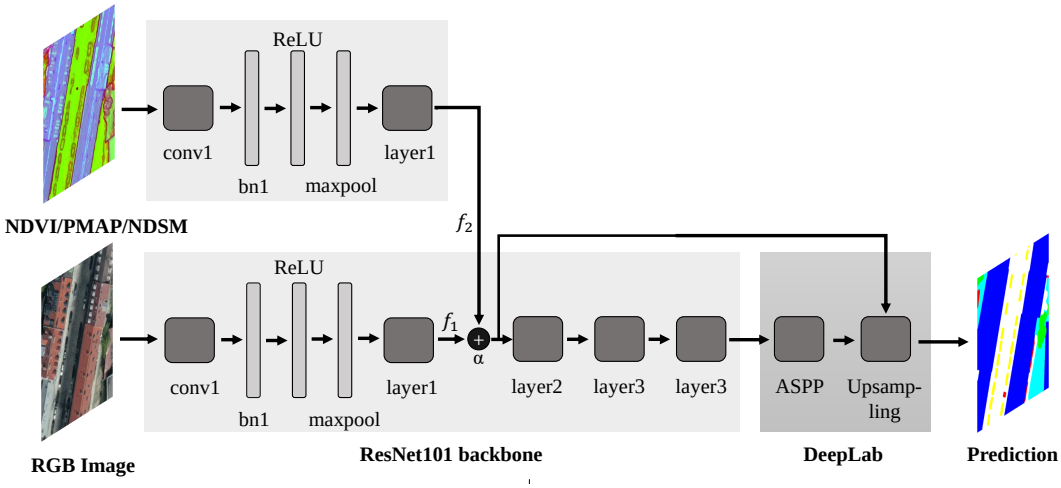


Figure 1. The modified DeepLabV3+ architecture to accept two input images.

of DeepLab, because only the first branch (RGB) contributes while the second branch is disregarded). The DeepLab model is pre-trained on ImageNet [Deng et al., 2009] as well as the ISPRS Potsdam [Rottensteiner et al., 2014] dataset, down-scaled to 10cm to match the resolution of Munich and Moabit. All input data for DeepLab is converted to uint8, with PMAP, NDVI, and NDSM scaled to utilize the entire 0-255 range.

4. Results and Discussion

We evaluate the DA performance by applying to T models trained on S and those trained on T. In the first case, we will assess whether it is worth performing DA at all, and in the second, we will track the so-called performance gap. Please note that to track the performance gap, one cannot use all available data of S but should always keep the validation data. The ratio between training and validation data is 70:30. Due to the limited amount of labelled data, we do not use a separate test set. However, since DA testing is carried out across datasets, this is less of an issue. Domain adaptation is applied to RGB and NIR channels, while 3D data and geometric properties such as planarity remain unaltered.

To track the quantitative results, we use the usual metrics Overall Accuracy (OA), Cohen's Kappa (κ) and the F1-Score (F1). Two latter metrics are important because our datasets contain a few small classes.

4.1 Datasets and domain adaptation

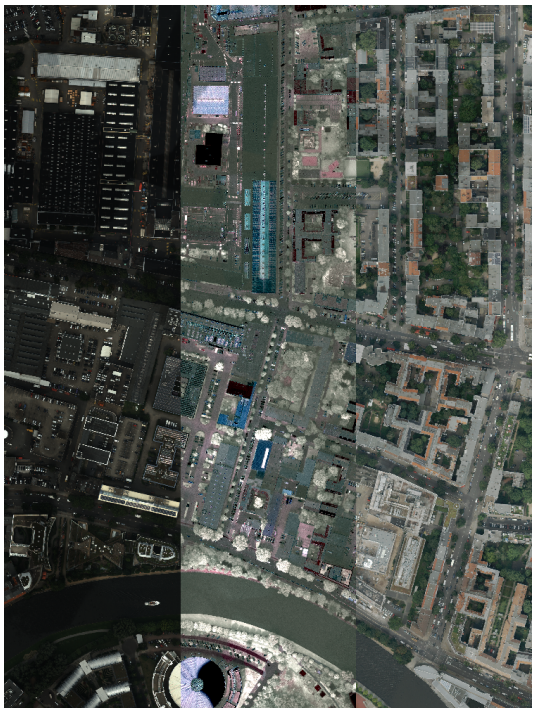
We consider one pair of datasets from two large German cities: Munich's city center and Berlin's Moabit district. Both datasets exhibit a ground sampling distance (GSD) of 0.1m, are available as uint16 data, providing a color depth of 65536, and are results of a photogrammetric reconstruction from a sequence of images taken by two different airborne cameras: an airborne camera DMC II 230 at a GSD of 10cm for Munich and the DLR MACS-HALE camera [Brauchle et al., 2015] for Moabit, respectively. Besides, two different methods – [Bulatov et al., 2012] for Munich and [Piltz et al., 2016] for Moabit – were applied for calculating the NDSM. The acquisition times were also different: The Moabit dataset was captured on a cloudy day in March, so the image is very sallow, with many leafless trees, while Munich, taken during a summer day, is very contrast-heavy, especially between sunlit and conspicuous shadow regions. The Moabit dataset has a more versatile land use because the North-Western fragment resembles a factory, the North Eastern part is residential, and there is a river in the South, which is not even present in the Munich dataset. There are some green areas along the river, but overall, the Moabit dataset contains fewer green areas than the Munich dataset, which are also more difficult to recognize due to lighting conditions. Inspired by the Potsdam dataset [Rottensteiner et al., 2014], we manually labeled 40 patches of 512×512 pixels for each dataset with six classes: building, road, grass, tree, car, and clutter. We split the patches into training and validation set with the aforementioned ratio.

First, we visually assess the domain adaptation results (see Fig. 2). Moabit and Munich are visually quite distinct what can be seen from the first part of each figure. The Moabit dataset appears much darker, with a few very bright objects. By adapting the Munich dataset to Moabit, the Munich image is darkened. While the CCA results flatten the contrast, and the result appears a little blurred, the adaption by HM keeps the coloring and highlighting of building roofs. The adapted results are shown in the middle and right-hand part of each image. The more similar one of these parts is to the first part of the other illustration (here Moabit), the more successful the adaptation has been. Adapting Moabit's dataset to Munich makes the Moabit data brighter, and green areas become visible. Again, results obtained by the CCA approach appear blurred. Further, some flat gray entities, such as the parking lot in the upper left corner and some streets, become more green. The HM approach results in the most visually correct adapted image in both datasets.

4.2 Classification accuracy

The final assessment of the improvement of results through domain adaptation is carried out indirectly by evaluating the classification results. Therefore, the datasets adapted using domain adaptation are classified using Random Forest and DeepLab and compared with ground truth.

4.2.1 Random Forest The results of Random Forest classification are given in Table 1. The best results are always achieved with DA using HM. In all cases, an improvement in the classification accuracy can be achieved. Before adapting the Moabit data according to Munich features (without domain adaptation) and by applying the Munich classification model



(a) Moabit

(b) Munich

Figure 2. Results of the domain adaptation of Moabit and Munich. From left to right: initial RGB status, CCA-based DA and HM-based DA to the respective other data set. In other words, the reader is supposed to compare the left-most column of (a) with the middle and right column of (b) and viceversa.

directly to the original Moabit data, the maximum overall accuracy is 60% even when 3D data is added. With domain adaptation, this value can be increased by +5% with CCA and even increased to 72% with HM. Using only color information, the results can be improved by $+7\%$ in the case of adapting Moabit to Munich and by $+4\%$ the other way around concerning the F1 Score. If DA is performed using CCA, we only improve the F1 Score by incorporating 3D information into the classification process and by adapting Moabit to Munich. Unfortunately, in other cases, the performance is worse than without domain adaptation.

As expected, using 3D information stabilizes the classification strongly since elevation information is an essential parameter for class separation. We obtain the peak accuracy increase in F1 of $+19\%$ for HM.

The impact of filter banks varies from feature set to feature set and from one direction of DA to another. One conspicuous observation is that, in the case of Munich to Moabit, κ and F1 almost always *degrade* up to 3 percent points, meaning that small classes are negatively affected. An improvement effect is barely noticeable, even in the case of fine-tuned data. One possible explanation for this contradiction to many related works is that superpixel-wise features may neutralize the positive effect of a filter bank.

Next, we discuss the performance gap. Regarding F1 Scores, it measures around 15% and 20% of difference for HM in case of considering respectively not considering the 3D data. Regarding overall accuracy, Moabit had a better outcome than Munich, reaching 4 percent points versus 11. This means either the Moabit data's adaptability to the Munich data works better, or the extracted classification model on Moabit data is less suitable for the classification process. Both match the visual impression that many details are lost in the very dark Moabit data set.

4.2.2 DeepLab

The results of DeepLab are shown in Table 2.

DA by HM yields better results than CCA. At $\alpha = 1$ (RGB only), HM improves the F1 score by about +9% when adapting the Munich dataset and +29% when adapting the Moabit dataset over the non-adapted versions. The gaps to the fine-tuned models now are about 24pp. and 7pp. respectively. There is a significant difference between the two datasets. We will discuss this in the qualitative analysis. Interestingly, the results for $\alpha = 0$ and $\alpha = 0.5$ also improve by about $+1\%$ to $+4\%$ over the unadapted versions in both datasets. Since NDSM and PMAP are unadapted, the changes must be traced back to the adapted red and NIR channels used to calculate the NDVI.

Unfortunately, the results are worse with CCA as the domain adaptation method than without any adaptation. In the $\alpha = 1$ case, it performs about -23% worse in both directions compared to no adaptation. The other α values also perform worse. The performance gap between the results with or without domain adaptation is less significant for $\alpha = 0$ and $\alpha = 0.5$ than for $\alpha = 1$. This is not surprising since the datasets have totally different RGB looks, whereas the NDVI and NDSM represent an index and a physical attribute and, therefore, vary less between datasets. Overall configurations, the second input branch aids the segmentation task when combined with the RGB data ($\alpha = 0.5$).

The best results are yielded by models trained and evaluated on the same dataset, also called fine-tuned and shown in the right half of the table. The $\alpha = 0.5$ configuration is mostly the best, reaching an OA of about 86% on Munich and 92% on Moabit, though the other configurations are not far behind. This means that even in the fine-tuned case, NDSM and NDVI information mainly benefits classification. This confirms the finding of [Qiu et al., 2022].

4.3 Qualitative Reults

In Figure 3, the predictions on the Moabit dataset are shown. The classes can be worked out using only RGB information, and DeepLab is far better than RF. However, even with DeepLab, misclassifications between buildings and grass occur. This deficiency will be remedied as soon as elevation information is included, so the higher classes will always be well separated from the lower classes. In RF, HM adaptation dramatically improves the prediction by reducing false predictions of

DA	RF param		Moabit to Munich			Munich to Moabit			Munich on Munich			Moabit on Moabit		
mode	3D	MR8	OA	κ	F1	OA	κ	F1	OA	κ	F1	OA	κ	F1
no DA	n	n	39.62	5.93	17.72	45.46	27.50	29.86	69.79	53.44	46.07	69.50	57.85	53.75
	n	y	42.22	5.00	16.10	45.26	26.61	26.76	69.47	52.25	46.04	68.62	57.02	56.79
	y	n	53.68	36.67	33.22	79.70	71.67	57.02	85.98	78.41	65.63	84.50	78.91	72.99
	y	y	60.44	42.92	33.11	77.74	68.85	51.88	85.93	78.22	64.95	84.74	79.19	72.75
CCA	n	n	36.85	3.57	14.73	37.41	8.62	16.93						
	n	y	38.60	4.88	15.06	33.59	2.17	14.61						
	y	n	64.36	49.58	40.82	61.41	43.40	30.68						
	y	y	66.00	50.17	36.83	59.53	40.90	29.66						
HM	n	n	43.11	14.73	24.65	46.98	24.81	33.62						
	n	y	44.51	15.37	27.14	46.45	22.78	33.88						
	y	n	74.17	62.50	52.93	81.95	74.52	58.15						
	y	y	72.57	60.08	49.54	80.56	72.51	54.88						

Table 1. Random Forest classification results. In the first column, we record the mode for domain adaptation. The RF classifier uses mean and standard deviation (STD) of the raw channels and, if there is a y (=yes) in the second or third column, 3D information and filter banks MR8, are included, respectively.

DA		Moabit to Munich			Munich to Moabit			Munich on Munich			Moabit on Moabit		
mode	α	OA	κ	F1	OA	κ	F1	OA	κ	F1	OA	κ	F1
no DA	1	64.78	52.66	51.27	55.10	30.10	39.46	83.95	76.35	72.10	89.67	84.80	77.68
	0.5	89.06	84.03	73.36	75.81	66.22	66.33	85.51	78.67	70.59	92.06	88.27	81.45
	0	87.26	81.14	68.23	82.26	74.06	63.92	83.60	75.67	67.52	90.73	86.28	76.08
CCA	1	48.82	21.70	28.44	41.57	7.76	16.43						
	0.5	77.62	65.94	50.60	61.27	46.25	34.45						
	0	72.63	58.51	42.29	57.58	42.14	41.11						
HM	1	83.50	75.79	70.99	58.16	40.95	48.22						
	0.5	91.09	86.88	77.10	80.42	72.31	70.52						
	0	87.93	82.11	68.90	83.19	75.37	67.28						

Table 2. DeepLab classification results. With α = 1 only RGB information will be used for classification; $\alpha = 0$ only uses information of the second branch, namely NDVI, NDSM, and PMAP; $\alpha = 0.5$: features proceedings from both branches are equally weighted. In case of domain adaptation (DA), color information only (RGB and NDVI) is adapted.

low vegetation on the street and trees on the buildings. With CCA, the prediction worsens compared to the unadapted case. With DeepLab at $\alpha = 0.5$, the unadapted prediction is already quite good, thanks to that second input branch. HM adaptation improves the predictions by removing false positive clutter along the building outline and cars. Again, CCA worsens the prediction. Compared to RF, most cars could be retrieved with DeepLab.

On the Munich dataset in Figure 4, DeepLab without DA yields incorrect predictions in the 2D case, with the building class overriding most other classes. Domain adaptation using HM improves the segmentations; the road class is often classified as low vegetation, and cars are overlooked. With RGB and 3D data ($\alpha = 0.5$), higher and lower classes are better separated when using HM and in the unadapted case (not shown). Still, shadowy areas are often classified as low vegetation or clutter.

Overall, adapting Munich data according to Moabit features delivers worse classification results than the other way around, especially with DeepLab. One reason is the higher intra-class variability of Munich data. From the right Tables 1 and 2, we can see that Moabit has better classification results despite appearing more versatile. Adjusting the features according to Munich allows the classifier to retain its essential properties, with fewer opportunities to overfit the classifier. Contrarily, in Munich alone, the shadows are sufficiently misleading for all classifiers, as, e.g., Fig. 4 shows.

The results demonstrate that DeepLab is a more powerful classification tool than RF and provides better classification performance. DA has a meaningful impact on both classification methods, and the application of HM can significantly increase the accuracy of both classification models. Furthermore, the second input branch, which includes elevation information, strongly aids in semantic segmentation across datasets and within a dataset in all cases.

5. Conclusion and Future Work

We demonstrated the effectiveness of DA using CCA and HM to improve image classification by a peak value of almost 20%. Our approach reduced the differences between the source and target domains and laid a strong foundation for classification. Utilizing Random Forest and DeepLab for adapted image data highlighted their respective strengths. Random Forest proved to be robust and efficient, particularly in processing data with lower complexity, while DeepLab enhanced semantic segmentation through deeper feature extraction.

In DeepLab, over all configurations, be it fine-tuned, adapted using any method, or no adaptation, the second input branch, consisting of NDVI index and elevation information, is greatly helpful in the segmentation task. This second branch is less susceptible to domain shifts, so using it alone achieves good results even without any adaptation. Similarly, including 3D data in Random Forest also stabilizes the predictions on a different dataset.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-17-2025 | © Author(s) 2025. CC BY 4.0 License.

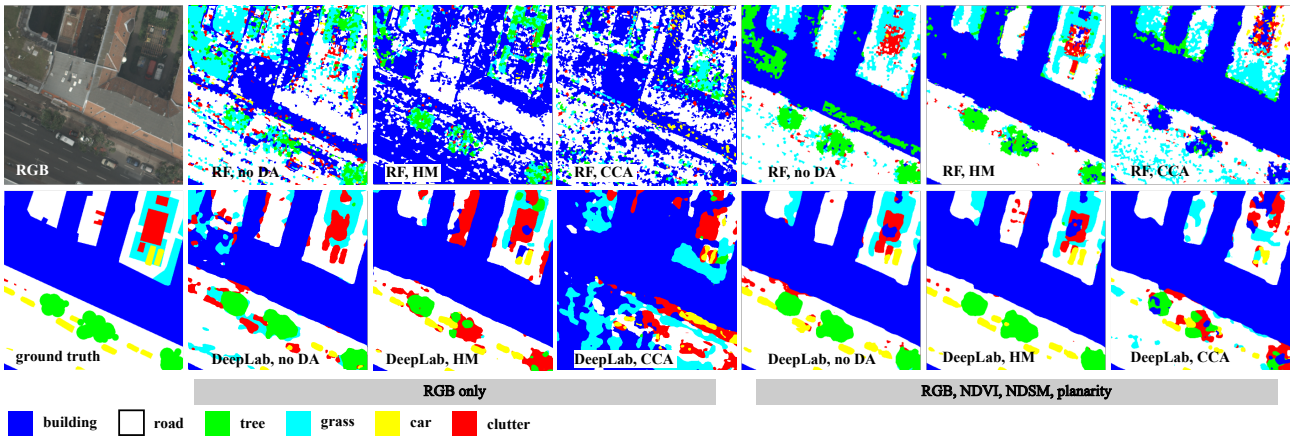


Figure 3. Prediction results of RF and DeepLab trained on Munich, applied to Moabit. The RGB was image was brightend for better visibility. The colors for the classes are provided in the legend below.

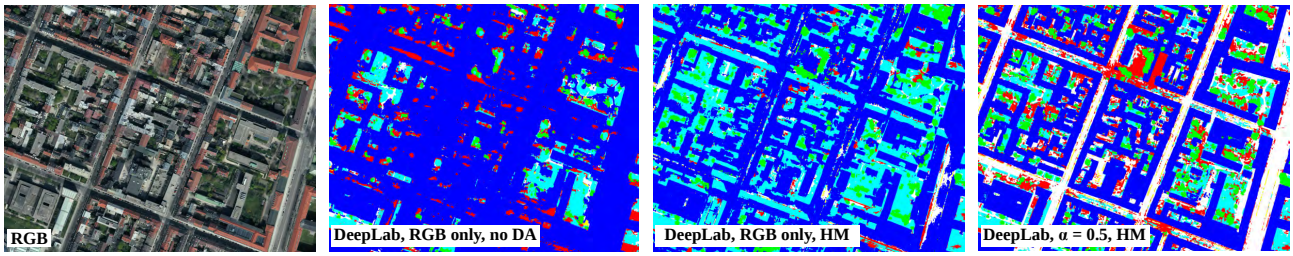


Figure 4. Prediction results of DeepLab trained on Moabit, applied to Munich, with the classes as in Figure 3

The findings suggest combining DA with advanced classifiers increases accuracy and model robustness across diverse datasets.

Future work should refine and apply these techniques to datasets from varying regions and climates. Evaluating methods on complex datasets with high variability and noise will be essential to assess their generalizability. Exploring transfer learning and adversarial training could improve performance by enabling more sophisticated mappings between domains. This way, not only color differences, but differences in texture and noise can be transferred as well.

As the direction-dependent data evaluation has shown, heavily shadowed areas enormously influence adaptability. Future work should consider these semantics and, if necessary, adjust or even preprocess shaded areas separately. Foundation models promise a future, in which dataset-specific fine-tuning becomes unnecessary: by training on massive corpora, they gain powerful generalization abilities. Future work should explore whether domain adaptation still adds value—especially for highly unusual datasets, like the Moabit dataset in this work.

Acknowledgment

We thank the German Aerospace Center (DLR) for acquiring, pre-processing, and providing us the Moabit dataset. For a more appealing formulation of the motivation section, AI generative techniques were used, but thoroughly revised by the authors.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2274–2282.

Brauchle, J., Hein, D., Berger, R., 2015. Detailed and highly accurate 3D models of high mountain areas by the MACS-Himalaya aerial camera platform. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XL-7/W3, 1129–1136.

Breiman, L., 2001. Random forests. *Machine learning*, 45, 5–32.

Bruzzone, L., Persello, C., 2009. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Transactions on Geoscience and Remote Sensing*, 47(9), 3180–3191.

Bruzzone, L., Prieto, D. F., 2002. A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images. *Pattern Recognition Letters*, 23(9), 1063–1071.

Bulatov, D., Wernerus, P., Gross, H., 2012. On applications of sequential multi-view dense reconstruction from aerial images. *ICPRAM (2)*, 275–280.

Cai, Y., Shang, Y., Yin, J., 2024. Multidan: Unsupervised, multistage, multisource and multitarget domain adaptation for semantic segmentation of remote sensing images. *Proceedings of the 32nd ACM International Conference on Multimedia*, 1168–1177.

Cai, Y., Yang, Y., Shang, Y., Shen, Z., Yin, J., 2023. DASR-SNet: Multitask domain adaptation for super-resolution-aided semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 248–255.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, 1180–1189.

Gross, H., Thoennessen, U., 2006. Extraction of lines from laser point clouds. *Symposium of ISPRS Commission III: Photogrammetric Computer Vision PCV06. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36number Part 3, 86–91.

Han, J., Yang, W., Wang, Y., Chen, L., Luo, Z., 2024. Remote sensing teacher: Cross-domain detection transformer with learnable frequency-enhanced feature alignment in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, H., Li, B., Zhang, Y., Chen, T., Wang, B., 2024. Joint distribution adaptive-alignment for cross-domain segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

Inamdar, S., Bovolo, F., Bruzzone, L., Chaudhuri, S., 2008. Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4), 1243–1252.

Liu, X., Zhan, Z., Yuan, J., 2021. Domain adaptation algorithm based on manifold regularization. *IEEE International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, IEEE, 30–33.

Nielsen, A. A., Canty, M. J., 2009. Kernel principal component and maximum autocorrelation factor analyses for change detection. *Image and Signal Processing for Remote Sensing XV*, 7477, SPIE, 266–271.

Piltz, B., Bayer, S., Poznanska, A.-M., 2016. Volume based DTM generation from very high resolution photogrammetric DSMs. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41, 83–90.

Qiu, K., Budde, L. E., Bulatov, D., Iwaszczuk, D., 2022. Exploring fusion techniques in U-Net and DeepLab V3 architectures for multi-modal land cover classification. *Earth Resources and Environmental Remote Sensing/GIS Applications XIII*, 12268, SPIE, 190–200.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 256-271.

Shi, Q., Du, B., Zhang, L., 2015. Domain adaptation for remote sensing image classification: A low-rank reconstruction and instance weighting label propagation inspired algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10), 5677–5689.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 41–57.

Varma, M., Zisserman, A., 2005. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2), 61–81.

Volpi, M., Camps-Valls, G., Tuia, D., 2015. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107, 50–63.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135–153.

Wittich, D., Rottensteiner, F., 2021. Appearance based deep domain adaptation for the classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 82–102.

Wu, J., Tang, Z., Xu, C., Liu, E., Gao, L., Yan, W., 2022. Super-resolution domain adaptation networks for semantic segmentation via pixel and output level aligning. *Frontiers in Earth Science*, 10.

Xu, M., Wu, M., Chen, K., Zhang, C., Guo, J., 2022. The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data. *Remote Sensing*, 14(17), 4380.

Zheng, J., Fu, H., Li, W., Wu, W., Zhao, Y., Dong, R., Yu, L., 2020. Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 154–177.

Zhu, L., Ma, L., 2016. Class centroid alignment based domain adaptation for classification of remote sensing images. *Pattern Recognition Letters*, 83, 124–132.