Geospatial Correspondences for Multimodal Registration

Diego Marcos
University of Zurich
diego.marcos@geo.uzh.ch

Raffay Hamid
DigitalGlobe Inc.
raffay@cc.gatech.edu

Devis Tuia
University of Zurich
devis.tuia@geo.uzh.ch

Abstract

The growing availability of very high resolution (<1 m/pixel) satellite and aerial images has opened up unprecedented opportunities to monitor and analyze the evolution of land-cover and land-use across the world. To do so, images of the same geographical areas acquired at different times and, potentially, with different sensors must be efficiently parsed to update maps and detect land-cover changes. However, a naïve transfer of ground truth labels from one location in the source image to the corresponding location in the target image is generally not feasible, as these images are often only loosely registered (with up to ± 50m of non-uniform errors). Furthermore, land-cover changes in an area over time must be taken into account for an accurate ground truth transfer. To tackle these challenges, we propose a mid-level sensor-invariant representation that encodes image regions in terms of the spatial distribution of their spectral neighbors. We incorporate this representation in a Markov Random Field to simultaneously account for nonlinear mis-registrations and enforce locality priors to find matches between multi-sensor images. We show how our approach can be used to assist in several multimodal land-cover update and change detection problems.

1. Introduction

In recent years, there has been a tremendous increase in the amount and resolution of commercially available satellite imagery [14]. This growth has opened numerous avenues to monitor and analyze the land-cover and land-use around the world, resulting in many novel applications including precision agriculture [45], population density estimation [23], and location based services [32].

A key challenge common to these applications is the efficient generation of land-cover maps, *i.e.* segmenting remotely sensed images into semantic classes such as forests, roads, buildings *etc*. This problem is exacerbated by the need to frequently update these maps by accounting for the constant natural and man-made changes on the Earth's surface. The growing number of available air and space-borne sensors, together with their short revisit time makes the automatic updating of such maps with remote sensing data an important and challenging research direction [28].

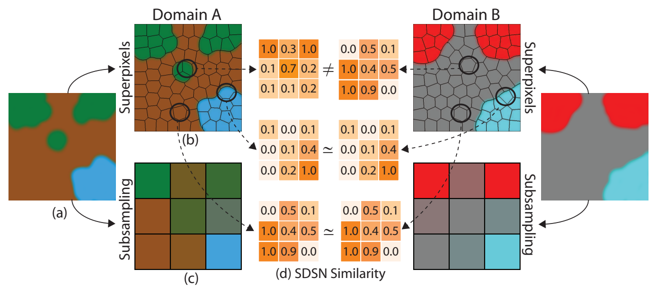


Figure 1. Illustration of Spatial Distribution of Spectral Neighbors (SDSN). (a): Remotely sensed image $I_{\mathcal{A}}$ in domain A. (b): Superpixels computed for $I_{\mathcal{A}}$. (c): Downsampled version of $I_{\mathcal{A}}$. (d): The SDSN features are computed as similarities of a superpixel in (b) with every value in (c). The same procedure is applied to the image in domain $\mathcal{B}$. Superpixels belonging to the same land-cover class tend have similar SDSN values across domains.

Traditional mapping approaches cannot be directly applied to solve this problem, as the appearance-consistency assumption they make does not generally hold true for multi-sensor multi-temporal (heterogeneous) images. This is due to the large variation in acquisition conditions *e.g.* frequency bands, resolution, acquisition times and geometry. It is therefore common to view multi-sensor land-cover update from the perspective of domain adaptation [2] where correspondences between the *source* and *target* domains are defined using a *shared* feature-space.

In this work, we propose a novel mid-level representation that assists in performing domain adaptation by extracting a domain-invariant feature for every image region (a super-pixel, Figure 1b) in each image in terms of the *spatial distribution* of its *spectral neighbors* (SDSN). Our representation is particularly geared towards image series acquired over the same geographical area at different times and using different sensors. In order to obtain domain invariance we exploit the fact that satellite images are loosely geo-registered, usually up to a non-uniform registration error of ±50 m. We can therefore divide the images into a relatively coarse set of patches (Figure 1c) in order to obtain an approximately registered shared coordinate system.

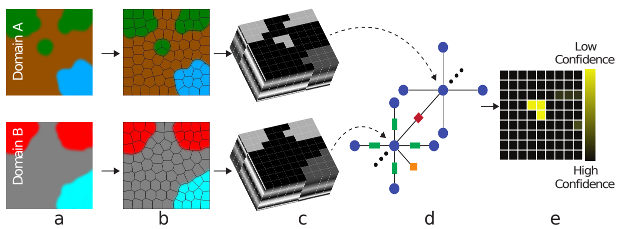


Figure 2. (a): Remotely sensed images. (b): Superpixel segmentation of the images. (c): Stack of domain invariant features computed for each superpixel maintaining the neighborhoods. (d): Graphical model used to match superpixels from both domains. (e): Contribution of each match to MRF cost function is used as confidence map of the matching, enabling the detection of areas with higher probability of having undergone a land cover change.

By encoding image regions in terms of their *spectral distances* from the mean value of each patch in their respective domains (Figure 1d), SDSN is able to provide a simple yet effective way to map information across different satellite sensors. This sensor invariance allows to use these features for multimodal image registration. We incorporate our SDSN representation in a Markov Random Field (Figure 2) where intra-domain edges are used to encourage smoothness and favor matches over short distances, while inter-domain edges encourage matching superpixels with similar domain-invariant features (Figure 2d). We show how this can be used for domain adaptation using two different strategies: direct transfer of land-cover ground truth (GT) (§ 4.1) and finding super-pixel pairs for unsupervised manifold alignment (§ 4.2). We also show how this approach can be used effectively for detecting land-cover changes in an unsupervised manner (§ 4.3).

2. Related Work

The problem of land-cover segmentation in multi-sensor and multi-temporal scenarios can be formulated as an instance of the more general problem of *domain adaptation*, which addresses the transfer of available domain-specific knowledge to a different but related domain. Both semi-supervised [10, 43] and unsupervised [20, 21] approaches have been proposed to perform domain adaptation.

For semi-supervised approaches, techniques involving co-training [38], label propagation [44], variants of expectation maximization [9] and SVM [16] have been successfully proposed. More recently, approaches involving co-regularization [26] and data rotation [31] have also been put forth. These approaches still require some labeled examples in the target domain, which prevents their application to problems where such labels are not available.

Unsupervised domain adaptation is generally considered a harder problem since we do not have any labeled correspondence between the domains. In this regard, approaches relying on source-target partial distribution similarity [19], clustering [35, 24, 7], structural correspondence learning [6], domain divergence minimization [5], manifold alignment [42] and deep learning [17] have been proposed. However, these methods require a certain level of correlation between the distributions of both domains. In this work, we put forth a strategy that is able to deal with distributions from different domains without requiring them to be correlated by exploiting the fact that for our setting the images are loosely spatially geo-registered.

A third approach, not always explicitly referred to as domain adaptation, consists of using engineered domain invariant features. It has been successfully applied in both remotely sensed [39, 22] and natural images [27]. Some of these features (*e.g.* SIFT [27] or shape descriptors [22]) achieve domain-invariance at the cost of discarding relevant information, *e.g.* color in R-G-B imagery, while focusing only on geometrical information. In contrast, we propose to fully take into account spectral information while also offering the possibility to include task-specific appearance descriptors in the process. SDSN is related to the local self-similarity (LSS) [33] and global self-similarity (GSS) [11] features. However, unlike these approaches, SDSN uses approximate geographical correspondences to allow for cross-domain comparisons, hence offering both the expressiveness of GSS and the simplicity of LSS.

In remote sensing, domain adaptation has been traditionally used for land-cover map update tasks [8, 3]. Most of these pipelines assume a perfect pixel-to-pixel registration between multi-temporal images, which is a serious limiting factor for high resolution and multi-sensor data. An object-based variant resides in the semantic tie points strategy proposed in [30] and used in [29] for remote sensing domain adaptation. An MRF-based approach [27] significantly relaxing the co-registration constraint is presented in [39], where registration and change detection are simultaneously performed. They use several correlation similarity measures that imply using the same number of spectral bands in both domains. Our approach extends this work to the multi-sensor setting where feature spaces are usually composed by different types and number of spectral channels, making it more general.

3. Proposed Method

We use super-pixels as our basic computational unit since they reduce the size of the problem while offering a meaningful spatial support. In this work we use the SLIC segmentation method presented in [1]. Given an image $I_{\mathcal{D}}$ of size $m \times n$ in domain $\mathcal{D}$, and a SLIC segment size parameter s, we build a super-pixel image $H_{\mathcal{D}}$ of size roughly $(m/s \times n/s)$.

We formulate our problem as an object matching problem by using a Markov Random Field (MRF), similar to [39, 27]. An important feature of this model is that the contribution to the MRF energy associated to each matched pair can be used as an estimate of matching confidence. Every super-pixel $H^j_{\mathcal{B}} \in H_{\mathcal{B}}$ in domain $\mathcal{B}$ (target) is matched to super-pixel $H^i_{\mathcal{A}} \in H_{\mathcal{A}}$ in the domain $\mathcal{A}$ (source) with a certain confidence relative to rest of the matches.

3.1. Spatial Distribution of Spectral Neighbors

Our main hypothesis is that objects that are spectrally similar in one domain tend to be spectrally similar in other domains, except when they have undergone a land cover change. For instance, a patch of vegetation in an RGB image is likely to have a similar color to other areas of vegetation in that image. At the same time, a patch of vegetation in a near infrared (NIR) image is likely to look very similar to other vegetated areas in the same NIR image. This within-image similarity is independent to how similar or dissimilar a particular patch of vegetation might look across the two images. We use this observation to encode each super-pixel $H^i_{\mathcal{D}}$ in domain $\mathcal{D}$ in terms of its similarity to other regions of the image (see Figure 1).

To do so, we start by computing a downscaled version $J_{\mathcal{D}}$ of the original image $I_{\mathcal{D}}$ as the average spectral signature of every non-overlaping $d \times d$ patch in $I_{\mathcal{D}}$. Here $J_{\mathcal{D}}$ is of size $(m/d \times n/d)$ and contains $Q = (mn)/d^2$ elements. We then compute the SDSN feature $\mathbf{f}^i_{\text{SDSN}}$ for $H^i_{\mathcal{D}}$ as:

$$\mathbf{f}^i_{\text{SDSN}} = [f^{i1}_{\text{SDSN}} \cdots f^{iq}_{\text{SDSN}} \cdots f^{iQ}_{\text{SDSN}}], \quad (1)$$

where

$$f^{iq}_{\text{SDSN}} = e^{-\sigma \| S(J^q_{\mathcal{D}}) - S(H^i_{\mathcal{D}}) \|^2}. \quad (2)$$

Here, $S(J^q_{\mathcal{D}})$ and $S(H^i_{\mathcal{D}})$ are the spectrum (*e.g.* RGB color) associated to $J^q_{\mathcal{D}}$ and the mean spectrum of $H^i_{\mathcal{D}}$ respectively. Note that $\mathbf{f}^i_{\text{SDSN}}$ has Q dimensions. Each element f^{iq}_{SDSN} of $\mathbf{f}^i_{\text{SDSN}}$ encodes the similarity of the spectrum of $H^i_{\mathcal{D}}$ to the average spectrum of a particular patch of the image, $J^q_{\mathcal{D}}$. Downscaling allows for robustness against registration noise between the images in the different domains. For example, a 15 pixel shift in the original image becomes a sub-pixel shift of 0.15 pixels using downscaling factor of 100.

3.2. Matching Formulation

We build on previous matching approaches relying on MRF such as [34, 27, 15, 39]. We define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where every edge $\epsilon_{ij} \in \mathcal{E}$ connects two nodes $i, j \in \mathcal{V}$ with a weight $c(\epsilon_{ij})$. Every node i corresponds to a super-pixel $H^i_{\mathcal{B}}$ in the target domain. Since we expect the misregistration shifts to be consistent within a region of the image, possibly with shift larger than the superpixel size, we consider mid-range connections for node i, $\mathcal{N}_i$, beyond first order neighborhoods. In our experiments (Section 4) we use a 25×25 super-pixel neighborhood. We make use of the SLIC grid initialization to define the neighborhood systems efficiently. We set weights $c(\epsilon_{ij})$ inversely proportional to the geographical distance between node i and its neighbors and we normalize them such that $\sum_{\epsilon_{ij} \in \mathcal{N}_i} c(\epsilon_{ij}) = 1$. Each node is defined by its geographical coordinates $\mathbf{p}_i = (x, y)$ and is assigned to a matching vector $\mathbf{w}_i = (u, v)$ towards a super-pixel $M_i = H^k_{\mathcal{A}}$ in the source domain, defined by its coordinates $\mathbf{q}_k = \mathbf{p}_i + \mathbf{w}_i$. M is a look-up table storing the currently selected matches. The matching process is formulated as an energy minimization over the graph $\mathcal{G}$ as:

$$E(M) = \sum_i \Theta_{\text{data}}(H^i_{\mathcal{B}}, H^k_{\mathcal{A}}) + \lambda_{\text{small}} \sum_i \Phi_{\text{small}}(\mathbf{w}_i) + \lambda_{\text{smooth}} \sum_i \Phi_{\text{smooth}}(\mathcal{N}_i). \quad (3)$$

The data term Θ_{data} measures the dissimilarity between $H^i_{\mathcal{B}}$ and its match $H^k_{\mathcal{A}}$, defined by:

$$\Theta_{\text{data}} = \sum_{\mathbf{f} \in F} \alpha_{\mathbf{f}} \Theta_{\mathbf{f}} \quad (4)$$

where $\alpha_{\mathbf{f}}$ is the weight given to each dissimilarity measure $\Theta_{\mathbf{f}}$, computed using the feature $\mathbf{f}$, *e.g.* SDSN, SIFT, color, *etc*. Here F defines the set of all features considered.

The dissimilarity between a pair of superpixels H^k_A and H^j_B in feature $\mathbf{f} \in F$ is computed as:

$$\Theta_{\mathbf{f}}(H^k_{\mathcal{A}}, H^i_{\mathcal{B}}) = -log\big(\mathbf{f}(H^k_{\mathcal{A}})^{\top} \cdot \mathbf{f}(H^i_{\mathcal{B}})\big) \quad (5)$$

We normalize each feature to have unit ℓ_2-norm. To further spread the samples over the unit ball, we center every vector to zero mean. Note that the matrix version of this formulation can use optimized BLAS Level-3 [18] and therefore can be computed efficiently by optimally using all the resources of modern computing architecture.

In Equation (3), the term Φ_{small} penalizes big matching displacements and depends only on the matching vector:

$$\Phi_{\text{small}}(i) = \|\mathbf{w}_i\|_2 \quad (6)$$

Similarly, Φ_{smooth} penalizes matching vectors deviating too much from the average matching vector in a neighborhood:

$$\Phi_{\text{smooth}}(\mathcal{N}_i) = \|\mathbf{w}_i - \sum_{j \in \mathcal{N}_i} c(\epsilon_{ij}) \mathbf{w}_j\|_2 \quad (7)$$

where all $j \in \mathcal{N}_i$ are the neighbors of i and each ϵ_{ij} the corresponding edge, with $c(\epsilon_{ij})$ being the edge weight.

The confidence of the match of node i in the target domain $\mathcal{B}$ is then defined as $-E(M_i)$.

3.3. Optimization

Since satellite images are loosely pre-aligned, the optimal solution does not have large $\mathbf{w}$. Therefore, we limit the search for a match for $i \in \mathcal{B}$ to a window of size $w \times w$ around the initial match. In practice, we initialize the system on the geographically nearest super-pixel in $\mathcal{A}$. Note that we can see the matching problem as a classification problem with w^2 classes, corresponding to every possible match for each super-pixel in $\mathcal{B}$[15]. To find a set of matches M that minimize Equation (3), we employ the Iterated Conditional Modes (ICM) algorithm [4]. Thanks to the grid structure of the graph we can use Fast Fourier Transform (FFT) to compute the energy in the form of a convolution, which significantly improves the efficiency of the algorithm. The fact that the initialization is never very far from the solution [37], the use of super-pixels and the FFT means that, for image pairs used in this work, the presented method typically converges in less than 10 seconds using ICM on a standard personal computer.

4. Experiments and Results

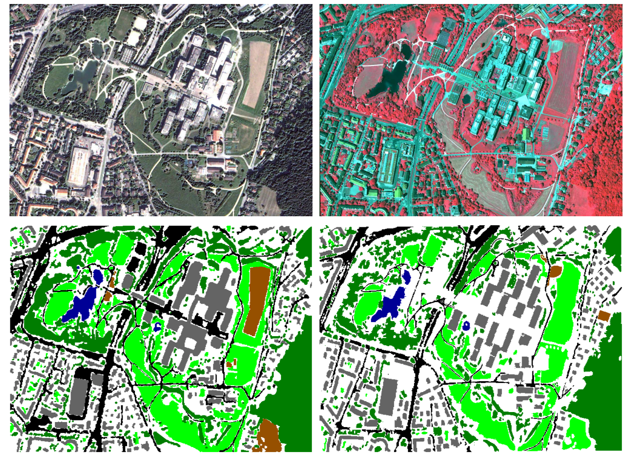
We apply our proposed representation to three different problems within the context of multimodal registration. In all the experiments the SDSN feature is compared to a multi-scale SIFT feature over the average color channel with patch sizes of 9, 17 and 33 pixels [40] and a feature consisting of the common spectral bands, thereafter referred to as "color". In all the experiments using a set of two features, the values of $\alpha_{\mathbf{f}}$ have been set to 0.5.

4.1. Ground Truth Transfer

We aim to transfer the available ground truth (GT) from the source image to the target image, while simultaneously avoiding the regions that have likely undergone some land cover change. This transferred GT is then used to train a kNN classifier in the target domain to generate an updated land cover map. The choice of kNN classifier is due to its simplicity and distribution independence. We use a hand labeled GT of the target domain to validate the map obtained.

4.1.1 Dataset and Setup

The source domain consists of five QuickBird [12] satellite images of Zurich Switzerland taken in August 2002. They have four channels: near infrared, red, green and blue (NIR-R-G-B), and a resolution of about 0.62 cm/pixel. These image are a subset of the Zurich Summer dataset presented in [41]. The target domain is a corresponding set of five NIR-R-G aerial images of the same area, with nearly the same footprint, captured during the campaign of summer 2013 and provided by the Swiss Federal Office of Topography [36]. We refer to this dataset as NIR-R-G Orthophoto data. The resolution of the target images is 25 cm/pixel. To test our approach in the case where source and target only share the R and G bands, we discard the NIR band of the QuickBird images and use exclusively the R-G-B bands throughout the experiments. Figure 3 shows an example image pair and the corresponding GT maps. In this dataset, the geo-registration error of the image pairs ranges from 5 m to 15 m. Each image in the source domain has a quite dense GT land-cover map consisting of between four and six classes among Roads, Buildings, Trees, Grass, Bare Soil, Water, Railways and Swimming Pools.



(a) (b)

Figure 3. One of the 5 image pairs from the Zurich dataset. (a): Quickbird [12] image, the source domain, and the corresponding GT. (b): False color Representation of the 3 band NIR-R-G Orthophoto data [36] (2013) and its GT: Roads, buildings, trees, grass, water and bare soil. Best viewed in color.

We treat each of the five image pairs as an independent GT transfer problem. For each image pair, we first re-scale them to the size of the smaller image in the pair. Note that this step is not required but it results in obtaining a similar number of superpixels, which helps getting a good matching. The images are then segmented with SLIC [1], with the superpixel size of 10 pixels and the regularization parameter values set to 10. The SDSN features are computed using a downsampling factor d of 20 and a σ of 0.5. For the MRF matching we used $\lambda_{\text{smooth}} = 0.05$ and $\lambda_{\text{small}} = 0.05$.

After matching, 90% most confident matches are used for transferring the GT to the target image. We then used all the transferred GT to train a pixel-wise kNN classifier with $k = 5$. The classified land cover map is compared to the hand labeled target GT for validation. We report results from QuickBird [12] to Orthophoto [36] and vice versa.

4.1.2 Results

The classification results are shown in in Table 1. Using SIFT or color features alone produces results that are significantly worse than the results obtained using the proposed SDSN features. Using SDSN in conjunction with SIFT produces the best results on average. This is because SDSN and SIFT encode very different properties of the super-pixels, *i.e.* the former encodes spectral information in terms of global interactions across the whole image, while the latter encodes the local geometry. These two forms of information complement each other in describing land-cover changes, resulting in better GT transfer. We also compare our results with those obtained using a multi-modal mutual information-based registration method [25] (Table 2).

#	Transfer Dir.	Color	SIFT	Color +SIFT	SDSN	SDSN +SIFT	SDSN +Color
1	$\mathcal{A} \to \mathcal{B}$	0.652	0.551	0.579	0.694	**0.716**	0.678
	$\mathcal{B} \to \mathcal{A}$	0.281	0.438	0.371	0.548	**0.572**	0.414
2	$\mathcal{A} \to \mathcal{B}$	0.665	0.629	0.678	0.684	**0.690**	0.681
	$\mathcal{B} \to \mathcal{A}$	0.396	0.510	0.475	0.536	**0.541**	0.466
3	$\mathcal{A} \to \mathcal{B}$	0.711	0.714	0.705	0.731	**0.749**	0.733
	$\mathcal{B} \to \mathcal{A}$	0.538	0.582	0.561	0.561	**0.584**	0.563
4	$\mathcal{A} \to \mathcal{B}$	0.585	0.644	0.551	**0.730**	0.694	0.597
	$\mathcal{B} \to \mathcal{A}$	0.453	0.548	0.484	0.498	**0.557**	0.480
5	$\mathcal{A} \to \mathcal{B}$	0.559	0.766	0.756	0.782	**0.790**	0.786
	$\mathcal{B} \to \mathcal{A}$	0.599	0.690	0.705	**0.723**	0.711	0.720
Mean		0.544	0.607	0.586	0.649	**0.660**	0.612

Table 1. Average classification accuracy for ground truth transfer. The first column correspond to image pairs. $\mathcal{A}$ and $\mathcal{B}$ correspond to QuickBird [12] Orthophoto [36] domains.

Method	SDSN+SIFT	Affine[25]	Non-rigid[25]
AA	66.0%	63.7%	64.0%

Table 2. Numerical comparison with [25], Average Accuracy.

4.1.3 Parameter Sensitivity and Circular Validation

We now focus on the sensitivity of our method to the choice of parameter values used when transferring the GT from source to target $\mathbf{A} \to \mathbf{B}$. In the left column of Figure 4 we see the result of applying this concept to the values of λ_{small} and λ_{smooth}, the SDSN's σ, and the down-scaling factor d. It can be observed that most image-pairs are not very sensitive to variations in the tested parameters showing the robustness of our framework to its various parameters.

We also study how a circular validation strategy [7] can help with estimating a good set of parameters for our framework. In the case of GT transfer, this can be done by transferring the GT from source to target $\mathbf{A} \to \mathbf{B}$, and then from target back to source $\mathbf{A} \to \mathbf{B} \to \mathbf{A}$, where it is compared to the original GT for evaluation. This setting corresponds to the right column of Figure 4. It can be observed that the optimal values obtained during validation $\mathbf{A} \to \mathbf{B} \to \mathbf{A}$ are similar to the optimal values required for our original problem $\mathbf{A} \to \mathbf{B}$. This result shows that we can employ this circular validation strategy [7] in practice to select the optimal parameter values required by our framework. Note that

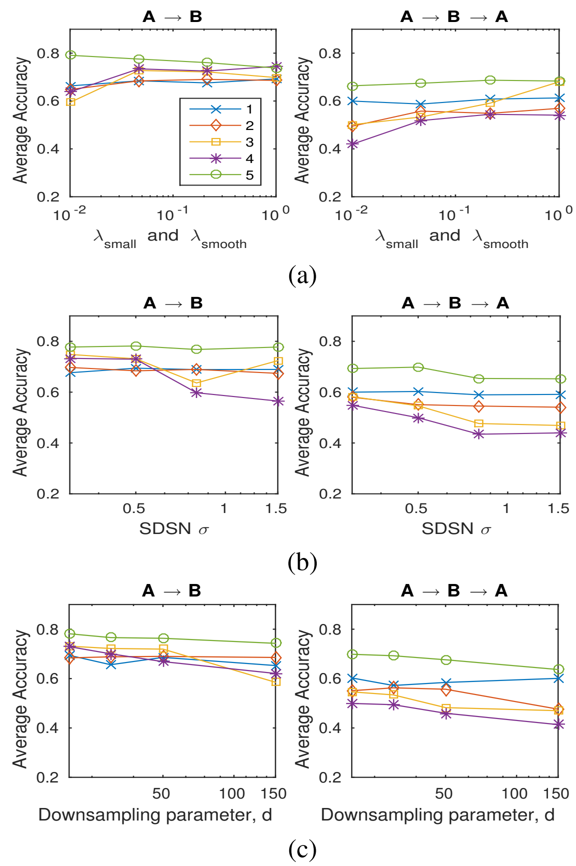


Figure 4. Classification accuracy in the five Zurich image pairs for different values of (a): λ_{small} and λ_{smooth} (both take the same value in this experiment), (b): σ value for the SDSN feature, (c): the downscaling parameter d for the SDSN feature. On the left, we transfer the GT from source to target. On the right is from source to target and back to source.

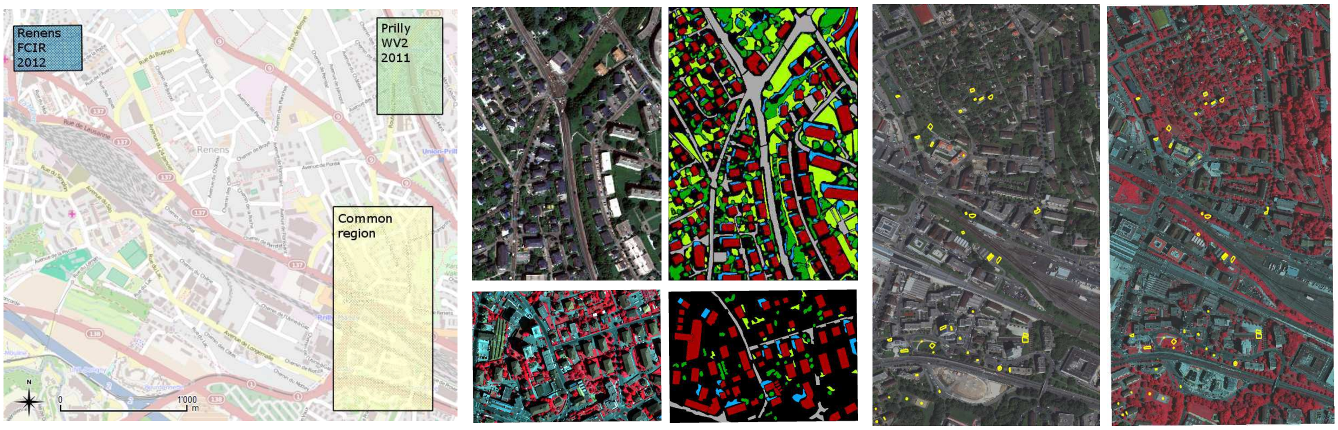
the downsampling factor d determines the number of operations required to compute the SDSN feature and affects the computation time in a quadratic manner (see Table 3). However, the results in Figure 4 suggest that this time can be greatly reduced at a small cost in accuracy.

d	2	5	10	20	50	100
time (s)	4.71	0.75	0.255	0.073	0.04	0.04

Table 3. Time (images 700 $\times$ 1000 pixels on a single CPU) to compute SDSN features wrt. downsampling, d.

4.1.4 Sensitivity to Perturbations in the Input

We use the image shown in Figure 3a in order to explore how sensitive our method is to three types of perturbations: 1) amount of change, 2) displacement and 3) rotation between the images. To do so, we match the image in Figure 3a with a perturbed version of itself. The land-cover changes were added by substituting vegetation superpixels



(a) (b) (c)

Figure 5. (a): Layout of the used images in the area of Lausanne, Switzerland. (b): Representation of the 2 images used (left) and the respective GTs (right). From top to bottom: Prilly 2011 (WorldView 2 [13]) and Renens 2012 (NIR-R-G Orthophoto). Color legend: buildings, roads, grass, trees and shadows. (c): Hand labeled super-pixel pairs (in yellow). Figure best viewed in color.

with bare soil ones. Performance was measured as the percentage of changed superpixels recalled, assuming that the amount of change had been correctly estimated. When considering displacement and rotation, the amount of change is fixed at 20%. Results in Table 4 show high robustness even for the most extreme amounts of change and displacement, which are the main sources of perturbation to be expected in remotely sensed images.

Change amount (%)	6	12	30	42
Recall (%)	96	97	96	84
Displacement (m)	10	20	30	50
Recall (%)	92	86	85	82
Angle (°)	1	3	6	10
Recall (%)	89	78	62	57

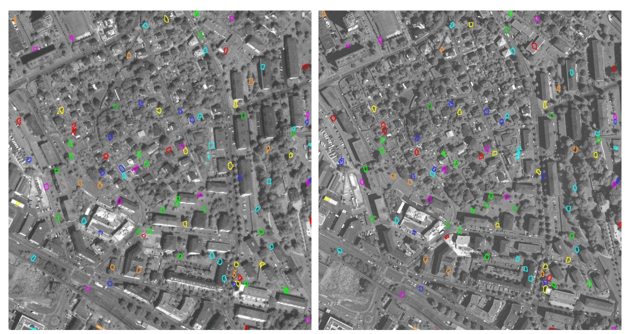
Table 4. Sensitivity to changes, displacements and rotations in the target ($\mathcal{B}$). Domains $\mathcal{A}$ and $\mathcal{B}$ consist of the RBG and NIR-R-G bands of Figure 3a respectively.

4.2. Unsupervised Manifold Alignment

We now explore the problem where the source and target images have a partial overlap, but there is no GT available in the overlap area. To perform domain adaptation, we cannot simply register the labeled super-pixels directly. However, we can project both domains in a common latent space where both domains are similarly distributed. We use the same settings presented in [29], where a hand-labeled set of super-pixel pairs in the overlapping area are used to perform manifold alignment between the two domains. Instead of manual selection, we use the proposed method to automatically find a set of these super-pixel pairs.

We use the manifold alignment algorithm in [43], where the local geometrical structure of the domain is preserved while enforcing weak class consistency using the matched super-pixel pairs. Once the domains are aligned, one can then directly train and test in the aligned domain.

To find a set of confident super-pixel matches that are also representative of the different land covers in the image, we partition the super-pixel spectra in the target domain using k-means clustering and select the most confident matches in each cluster.



(a) (b)

Figure 6. Automatically generated super-pixel pair map using the proposed SDSN features. (a): Grayscale version of the WorldView 2 [13] image used as source. (b): Grayscale version of the NIR-R-G Orthophoto [36] image used as target. Matched segment pairs are within the neighborhood of each other and are marked with same color in order to imply correspondence. Best viewed in color.

4.2.1 Dataset and Setup

For these experiments, we use the dataset presented in [29]. The images cover an area of Lausanne, Switzerland. The source domain is a WorldView 2 [13] image, with 8 spectral bands in the visible and infrared region, taken in 2011 and the target domain is a 25 cm/pixel resolution NIR-R-G Orthophoto [36]. Figure 5b shows the areas of interest, with their corresponding GT, and Figure 5c the area common to both domains. The land-cover classification includes 5 classes, as shown in Figure 5b.

The parameters for SLIC were set to segment size of 20 pixels and regularization parameter of 10. The σ value for the SDSN was set to 0.5 and the downsampling factor for the low resolution image was set to 100. For the MRF matching, we used $\lambda_{\text{smooth}} = 10^{-2}$ and $\lambda_{\text{small}} = 10^{-2}$. We then partitioned the superpixels in the target domain into 26 clusters and randomly took 10 confident matches from each of the 20 most populated clusters.

For manifold alignment, we used the same settings as in [29]. After projecting the data onto the latent feature space, we tested the performance of the alignment by classifying in the test image using only the labeled pixels from the source image. We report results using kNN with $k = 5$ training with 400 labeled pixels per class. We also tried other classifiers, such as Random Forest and SVM (not shown in this paper), obtaining comparable trends. For each type of feature (color, SIFT, SDSN and SDSN+SIFT) we generated 5 instances of the super-pixel pair set. For each instance, as well as for the hand labeled super-pixel pairs, we computed 10 realizations of the manifold alignment and classification training set.

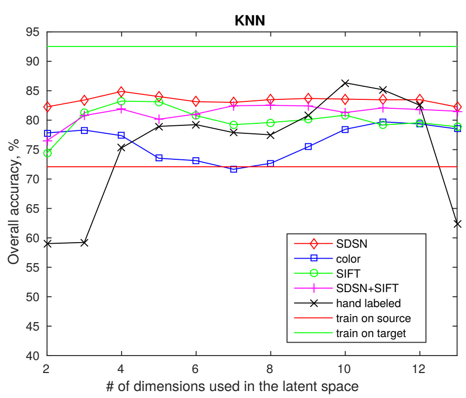


Figure 7. Classification results (as overall accuracy) on the manifold alignment experiment using kNN.

4.2.2 Results

An example of an automatically generated super-pixel pair set with SDNS feature matching is shown in Figure 6. Notice how the high-confidence super-pixels pairs are distributed across the whole image evenly.

In Figure 7 we see the average classification results. We see how the automatically generated super-pixel pair sets using SDSN, SIFT or SDSN+SIFT perform substantially better than the hand labeled map in the vast majority of cases. Moreover, in this experiment, using SDSN+SIFT does not seem to have an advantage over SDSN alone. This is possibly due to the presence of higher rise buildings, compared to the Zurich dataset, and the acquisition angle difference between domains. This has a big impact on the local geometry, to which SIFT is highly sensitive.

4.3. Change Detection

The main assumption underlying the SDSN feature is that spectral neighborhood relations are domain invariant and can be used to match with high confidence the areas that remain unchanged between domains. Considering the other end of the confidence spectrum, we can build a map of low confidence areas that can be used for change detection.

4.3.1 Datasets and Setup

To test our framework on change detection, we use an image-pair from the dataset in § 4.1.1, shown in Figure 8a and 8b, with change GT in green. We also apply our method to an RGB image-pair from Google Earth of an agricultural area near Melbourne, Australia taken on 10/17/2014 and 01/03/2015, Figure 10a-b. We chose this location as it represents a case where the three similarity measures, using SDSN, SIFT and color, provide different notions of change. Our pre-processing for this experiment is the same as in § 4.1. The MRF parameters λ_{smooth} and λ_{small} are set to 0.2. We set them higher in this case compared to the GT transfer setting because here we want to penalize non-smooth high-confidence matches more. The matching is done in both directions, *i.e.*, $\mathbf{A} \rightarrow \mathbf{B}$ and $\mathbf{B} \rightarrow \mathbf{A}$. The two resulting confidence maps are then combined by taking, for each location, the minimum value among the two maps.

4.3.2 Results

Figures 8c-f show as heat-maps the low confidence matches, in yellow. We calculated the confidence using 4 different invariant features within our MRF framework: the common bands (color), SIFT, SDSN and SDSN+SIFT. In the case of SIFT, Figure 8e, the algorithm detects the illumination changes in the forest (lower right corner of the image) as the dominant changes. Using the common bands R and B, Figure 8c, highlights mostly illumination changes on the roads and rooftops, with only changes #3, 6, 12 and 14 being clearly detected. SDSN, Figure 8d, shows a better correlation with the labeled changes, while detecting also changes in building's shadows. This undesired effect is reduced by using SDSN+SIFT. SDSN clearly detects changes #2, 3, 4, 5, 6, 8, 9, 10, 11, 12 and 14 while maintaining a false positive ratio comparable to or even lower than the color feature. We present the ROC curves for change detection in Figure 9. It can be observed that results incorporating the SDSN feature are significantly better than those

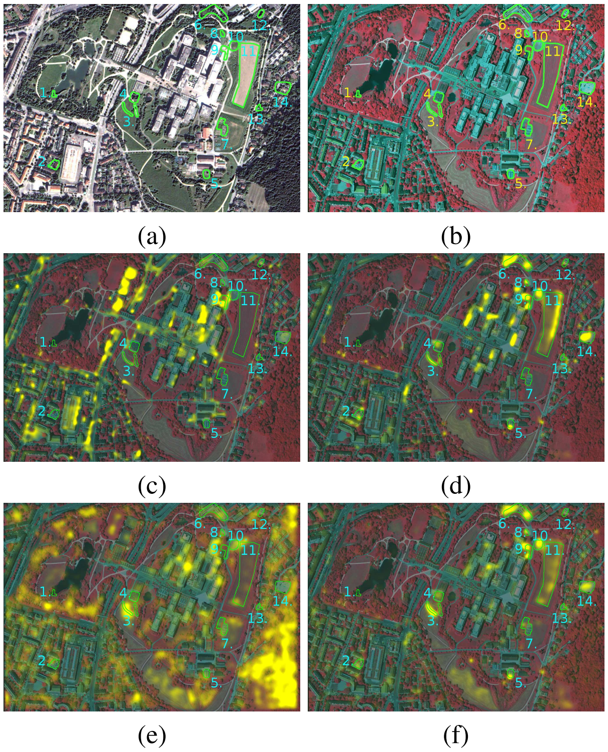


Figure 8. (a) Image in domain $\mathcal{A}$ and (b) $\mathcal{B}$. Change GT marked in green. (c-f) The low confidence areas, those contributing the most to the MRF energy, are shown in yellow. The matching has been performed with the same parameters using the following features: (a) common spectral bands, (b) SIFT , (c) SDSN and (d) SDSN+SIFT. Best viwed in color.

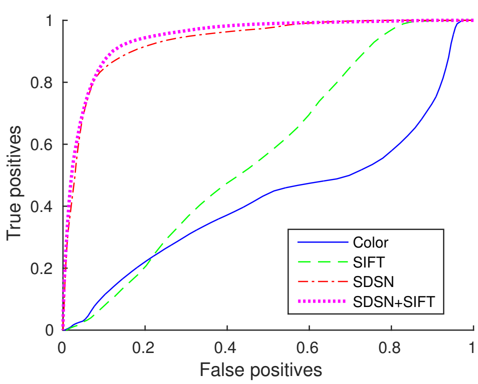


Figure 9. Correctly detected change pixels versus false positives for all different values of the detection threshold.

obtained using either color or SIFT features. Once more, using SDSN+SIFT results in a slightly better performance than SDSN alone.

Figure 10c-e show the non confident areas on the Google Earth image pair using the common bands, SIFT and SDSN features respectively. We can clearly see how each map corresponds to a different notion of change. Using the common bands, in this case all R, G and B, highlights the areas in which the color change is stronger, such as the dark green vegetation turning into bright bare soil. SIFT is correlated with geometrical changes in the image, such as the cloud shadow in the center top of Figure 10a. On the other hand, SDSN highlights the the areas that undergo an uncommon transformation compared to the predominant set of transformations. Given that most areas have either changed from vegetation to bare-soil or *vice versa*, these uncommon transformations include vegetation that continues to be vegetation (*e.g.* the field near top left corner) and bare soil that stays bare soil (*e.g.* roads). While the former transformation is an anomaly we would like to detect, the latter is an artifact due to the spectral similarity between roads and bare-soil classes.

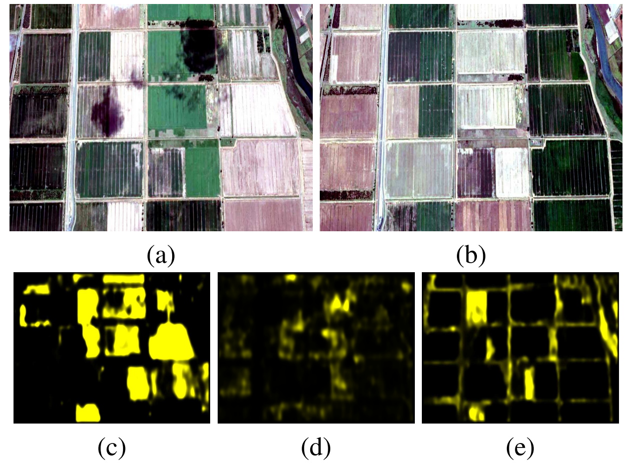


Figure 10. (a, b): Google Earth images taken in 10/2014 and 01/2015 respectively. (c-e): Low matching confidence areas are shown in yellow, (c) using color, (d) SIFT, (e) SDSN.

5. Conclusions

We proposed the spatial distributions of spectral neighbors (SDSN) as a cross-domain feature for multi-sensor, multi-temporal image pairs of sub-meter resolution. We showed that SDSN can help to match the super-pixels between two images from overlapping areas while distinguishing between the spectral changes that are the artifacts of different acquisition conditions from those due to real land-cover changes. We showed the usefulness of SDSN for land-cover map update and for change detection.

We incorporate the SDSN representation into a Markov Random Field to account for nonlinear misregistrations and to enforce a locality prior in order to find matches between multi-sensor, multi-temporal images. Furthermore, we compare SDSN with other features commonly used in remote sensing image registration. Our results demonstrate that SDSN performs significantly better than the alternatives considered, maximizing domain invariance and resulting in better classification and change detection.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012. 2, 4

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 1

[3] C. Benedek and T. Szirányi. Change detection in optical aerial images by a multilayer conditional mixed Markov model. *IEEE Transactions on Geoscience and Remote Sensing*, 47(10):3416–3430, 2009. 2

[4] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302, 1986. 4

[5] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008. 2

[6] J. Blitzer, M. Dredze, F. Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007. 2

[7] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010. 2, 5

[8] L. Bruzzone and S. B. Serpico. An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):858–867, 1997. 2

[9] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Transferring naive bayes classifiers for text classification. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 540. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007. 2

[10] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006. 2

[11] T. Deselaers and V. Ferrari. Global and efficient self-similarity for object classification and detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1633–1640. IEEE, 2010. 2

[12] DigitalGlobe. QuickBird datasheet. http://global.digitalglobe.com/sites/default/files/QuickBird-DS-QB-Prod.pdf/, 2000. 4, 5

[13] DigitalGlobe. WorldView 2 datasheet. http://global.digitalglobe.com/sites/default/files/DG_WorldView2_DS_PROD.pdf/, 2000. 6

[14] J. Doe. Commercial satellite imaging market - global industry analysis, size, share, growth, trends, and forecast, 2013 - 2019. *Transparency Market Research*, 2014. 1

[15] P. S. Dragomir Anguelov, H.-C. Pang, D. Koller, and J. D. Sebastian Thrun. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, volume 17, page 33. MIT Press, 2005. 3, 4

[16] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 289–296. ACM, 2009. 2

[17] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2

[18] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012. 3

[19] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of The 30th International Conference on Machine Learning*, pages 222–230, 2013. 2

[20] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012. 2

[21] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006. IEEE, 2011. 2

[22] L. Gueguen and R. Hamid. Large-scale damage detection using satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1321–1328, 2015. 2

[23] J. Harvey. Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing*, 23(10):2071–2095, 2002. 1

[24] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002. 2

[25] D.-J. Kroon and C. H. Slump. MRI modality transformation in demon registration. In *Proc. ISBI*, pages 963–966, 2009. 5

[26] A. Kumar, A. Saha, and H. Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in neural information processing systems*, pages 478–486, 2010. 2

[27] C. Liu, J. Yuen, and A. Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 2, 3

[28] D. Lu, G. Li, and E. Moran. Current situation and needs of change detection techniques. *International Journal of Image and Data Fusion*, 5(1):13–38, 2014. 1

[29] D. Marcos-Gonzalez, G. Camps-Valls, and D. Tuia. Weakly supervised alignment of multisensor images. In *Geoscience and Remote Sensing Symposium (IGARSS)*, July 2015. 2, 6, 7

[30] J. Montoya-Zegarra, C. Leistner, K. Schindler, et al. Semantic tie points. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 377–384. IEEE, 2013. 2

[31] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 2

[32] J. Schiller and A. Voisard. *Location-based services*. Elsevier, 2004. 1

[33] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 2

[34] A. Shekhovtsov, I. Kovtun, and V. Hlaváč. Efficient mrf deformation model for non-rigid image matching. *Computer Vision and Image Understanding*, 112(1):91–99, 2008. 3

[35] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 2

[36] Swisstopo. Swiss Federal Office of Topography SWISSIMAGE FCIR. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/images/ortho/swissimage/SWISSIMAGE_FCIR.html. 4, 5, 6, 7

[37] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov Random Fields. In *Computer Vision–ECCV 2006*, pages 16–29. Springer, 2006. 4

[38] G. Tur. Co-adaptation: Adaptive co-training for semi-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3721–3724. IEEE, 2009. 2

[39] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios. Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69, 2015. 2, 3

[40] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 4

[41] M. Volpi and V. Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2015. 4

[42] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI*, volume 2, page 3, 2009. 2

[43] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1541, 2011. 2, 6

[44] D. Xing, W. Dai, G.-R. Xue, and Y. Yu. Bridged refinement for transfer learning. In *Knowledge Discovery in Databases: PKDD 2007*, pages 324–335. Springer, 2007. 2

[45] C. Yang, J. H. Everitt, Q. Du, B. Luo, and J. Chanussot. Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture. *Proceedings of the IEEE*, 101(3):582–592, 2013. 1