A Sensor-Driven Hierarchical Method for Domain Adaptation in Classification of Remote Sensing Images

Claudia Paris, *Member, IEEE*, Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—This paper presents a Sensor-Driven Hierarchical Domain Adaptation (DA) method that aims at transferring the knowledge from a source domain (RS image where reference data are available) to a different but related target domain (RS image where no labeled reference data are available) for solving a classification problem. Due to the different acquisition conditions, a difference in the source and target distributions of the features representing the same class is generally expected. To solve this problem, the proposed method takes advantage from the availability of multisensor data to hierarchically detect features subspaces where for some classes data manifolds are partially (or completely) aligned. These feature subspaces are associated with invariant physical properties of classes measured by the sensors in the scene, i.e., measures having almost the same behaviour in both domains. The detection of these invariant feature subspaces allows us to infer labels of the target samples that result more aligned to the source data for the considered subset of classes. Then, the labeled target samples are analyzed in the full feature space to classify the remaining target samples of the same classes. Finally, for those classes for which none of the sensors can measure invariant features we perform the adaptation via a standard Active Learning (AL) technique. Experimental results obtained on two real multisensor datasets confirm the effectiveness of the proposed method.**

***Index Terms*—Transfer learning, Domain Adaptation (DA), multi-sensor data acquisition, invariant features, data fusion, classification, Remote Sensing (RS).**

I. Introduction

THE POSSIBILITY of generating accurate land-cover maps by applying supervised classification approaches to RS data has been extensively analyzed in the literature. The main drawback of these methods is the need of reference data for training the classification algorithm, which requires expensive and labour intensive field data collection. Therefore, in a real application scenario it is not reasonable to assume to have ground reference data available each time that a new RS data is acquired. To mitigate the need of labeled samples, the new RS image can be classified by exploiting the ground reference data associated with an image acquired by the same sensor in a region with comparable properties (i.e., same set of land-cover classes). However, when transferring the knowledge among pairs of RS images, even tough they are similar to each other, it is necessary to face many problems. The different acquisition conditions of the two data (i.e., illumination, atmosphere, look/view angles, sensor parameters) affect the radiometry of the scene. Moreover, the phenological state of the vegetation or the differences in the soil moisture can lead to crucial variations in the spectral response of the same land-cover classes (e.g., bare soil, crops). From the statistical view point, all these factors result in a shift of the probability distribution of the classes between the images. Hence, the direct application of the classifier trained on the source domain (RS data where reference data are available) to the target domain (RS data where no reference data are initially available) results in a low classification accuracy of the obtained land-cover map.

In machine learning and pattern recognition literature, the issue mentioned above is addressed by using DA methods in the framework of transfer learning. The main idea is to transfer the knowledge learned on the source domain to a target domain by modelling the differences among the areas [1], [2]. Several methods have been presented in the framework of change detection, where the source and the target domains are acquired in the same geographical area but at different times. Few approaches have been developed to normalize the images if they have been taken under identical acquisition condition [3]–[7], while others adopt image processing for matching as more as possible the statistical distributions of the images [7], [8]. More sophisticated techniques aim at adapting the classification model estimated on the source domain to the target domain [9]–[11]. In [9], [10] the main idea is to use in an unsupervised way the samples of the target image to tune the classifier in order to update the land-cover map generated on the source domain. In [11] the authors first apply an unsupervised change-detection method to the source and the target domains, then they exploit the unchanged pixels associated with the groundtruth samples of the source domain to generate the reference data of the target domain. However, all these methods can address the case of classification of time-series images acquired on the same area.

In the case of images acquired on different regions, it is not possible to use the change information or the temporal correlation between areas for addressing the DA issue. To solve this problem, a common DA strategy consists in weighting the samples of the source domain in order to use them in the classification of the target domain [12]–[15]. In [12], [13] the source samples weighting is combined with feature transformations in order to construct a feature representation that minimizes the domain difference, while in [14], [15] the re-weighting technique is jointly used with an active learning (AL) method to evaluate the similarity between source and target domains. Note that AL methods aim at iteratively expanding the original training set by selecting the most informative unlabeled samples of the target

domain. Typically, an interactive process with a supervisor is required to manually label the selected samples, thus strongly improving the classification accuracy [16]. Several approaches address the DA problem by gradually adapting the training distribution to the target domain rather than performing a samples weighting [17], [18]. In [17] the authors present a deep learning model to gradually replace source domain samples with target domain samples, while in [18] a unified architecture which jointly address feature learning, domain adaptation and classifier learning is proposed.

Many DA problems have been also addressed by semisupervised learning methods [19]–[27] when the reference data are not sufficient to represent the real distributions of the land-cover classes. Indeed, often ground data are acquired over contiguous sites easily to access, thus resulting in a unrepresentative pool of samples that affect the training of the classifier. SSL methods aim to solve this problem by taking advantage from the unlabeled samples of the image to better model the distributions of the classes and thus train the classifier. In [19] the authors present an iterative algorithm which exploits a weighting strategy based on a time-dependent criterion to include in the training set the unlabeled samples of the image. At each iteration a Support Vector Machine (SVM) classifier is trained with the enlarged set of labeled samples, thus gradually searching the optimal classification function. In particular, they observe that the most informative unlabeled samples are the ones close to the margin boundaries of the SVM. Recently, graph-based methods brought a great contribution in solving semisupervised classification problems due to their solid mathematical background [28]–[33]. Typically, both the labeled and unlabeled samples are considered as nodes of the graph, while the weights between the nodes represent the similarity among pairs of samples. This condition allows one to drive the labeling process in a natural way among samples of the same classes under the assumption of consistency (i.e., nearby points should belong to the same class) [28]. In [34], the graph represents the structure of the land-cover classes to highlight possible changes between the domains. Thus, the method does not require to have the same set of land-cover classes between source and target domains. First, the number of classes of the target domain is detected by means of a clustering algorithm. Second, a sub-graph matching algorithm is proposed to detect the common classes and to identify possible changes among pairs of land-cover classes. To match the classes of the different domains, the data are projected into a higher dimensional kernel induced feature space which allows a linear class separation.

In the computer vision community, several methods addressed unsupervised DA problems by learning new feature representations that are domain-invariant [35], [36]. This can be done by projecting the source and target domains to low-dimensional subspaces and then aligning the source subspace to the target one [37], or by directly learning a projection of the data where the distance between the domains is minimized [38]. In [39], the authors propose a transfer subspace learning approach by minimizing the Bregman divergence between the domains in lower dimensional spaces. A significant effort has been devoted to detect an intermediate feature subspace using the information conveyed by both the domains [40]–[43]. In [40], the authors address a DA problem in which the data from the source and the target domain are represented by heterogeneous features having different dimensions. Both the source and target data are projected onto an augmented common feature subspace [44], where it is possible to transfer the knowledge across the domains. The feature augmentation method has been extended to detect a manifold of intermediate domains [41]–[43]. In [41] the authors present a kernelized manifold-based approach which extends and improves the method reported in [42]. The main idea is to exploit a geodesic flow kernel to model the domain shift by integrating an infinite number of subspaces that characterize changes in geometric and statistical properties between the domains. In [45] a new framework based on the optimal transportation problem is presented to transport the source samples to the target distribution. According to the proposed regularization schemes, the class structure of the source domain is encoded to guarantee that samples belonging to the same class must undergo similar transformation. In [46] the authors propose a method which aims to select a subset of features that are characterized by both invariant spatial behaviour and discrimination ability among the set of land-cover classes. The feature selection is performed considering a novel criterion function based on a standard measure of distance between the classes and a novel metric that evaluates the stationary behaviour of features between the domains. Due to this feature-selection phase, the generalization capability of the classification method is strongly improved with respect to the standard techniques. In [47] a nonlinear deformation based on vector quantization and graph matching is presented to adapt the source domain to the target domain. The data manifolds of the images are locally deformed to facilitate the statistical alignment. Therefore, by maximizing the similarity of the graphs representing the two domains it is possible to transfer the knowledge from the source to the target domain in an unsupervised way. In [48]–[51] the authors address the adaptation from a local perspective, trying to adapt locally the samples of the images while preserving the geometrical structure of the entire distribution. In [51] the authors propose an alignment method which works directly on the manifolds of the images, thus addressing the cases of having multiangular, multitemporal, and multisource image classification problems. In greater details, the method aims to pull close samples of the same classes while preserving the geometry of each manifold along the transformation. However, to detect possible spaces where the manifolds are aligned the presented method requires a set of labeled data from the target image. Indeed, the alignment transformation is defined by using both labeled and unlabeled samples. Similarly, in [49] the authors focus on the feature extraction phase to statistically align the distributions of the source and the target domains either in a semisupervised or in an unsupervised way. In particular, they present a Transfer Component Analysis, which allows the preservation of the local geometry (data manifold) while minimizing the distance between the domains, thus improving the classification accuracy of the target domain regardless of the classifier.

From this brief analysis of the literature, it turned out that

accurate DA results can be achieved by defining a feature representation that minimizes the difference between the distributions in the source and target domains. However, the effectiveness of these data-driven methods strongly depends on both the discrepancy between the source and target distributions and some properties of the considered set of land-cover classes. In [52], [53] the authors pointed out that, unless the training samples are prohibitively large, there are no guarantees of success for unsupervised DA problems. Thus, DA problems become intractable when the similarity between source and target distributions is not sufficient. For this reason, differently from the literature, in the proposed approach the detection of the optimal features space relies on a physical-based approach that exploit the capability of the sensors to measure some almost invariant physical properties of the classes, i.e., having almost the same behaviour in the domains. When a sensor is able to measure physical properties, associated with subset of features, characterized by an invariant (similar) behaviour on one (or more) class(es) across the domains, it is possible to transfer the knowledge from the source to the target domain in an unsupervised yet reliable way. Although these feature subspaces guarantee a better alignment of the data with respect to the original space, for some specific classes they are not sufficient to model the entire distribution of the target domain. Thus, the main idea of the proposed method is to exploit hierarchically different invariant feature subspaces to infer the labels of the target samples of some classes almost aligned to the source data. The labeled target samples are then analyzed in the full feature space to properly represent the Probability Density Function (PDF) of the target domain for the considered subset of classes. Finally, the subset of classes for which none of the sensors available can measure invariant properties are adapted by using an AL procedure which takes advantage from the adaptation performed in the previous step. This step can be introduced to further increase the level of detail of the land-cover map already obtained in an unsupervised way for the invariant classes. Note that the proposed method cannot be applied to any domain adaptation problem and a preliminary analysis is required to design the hierarchical decomposition. However, the only assumption of the method is that in the considered DA problem there are sensors capable to measure biophysical parameters that have similar values the domains. Experiments conducted on two real multisensor datasets confirm the effectiveness of the proposed method.

The rest of the paper is organized as follow. Section II introduces the problem formulation and the notation used in the paper. Section III presents the proposed DA method, while Section IV describes the considered dataset. Section V illustrates and discusses the experimental results. Finally, Section VI draws the conclusion of this paper.

II. Problem Formulation

In this section we formalize the sensor-driven DA problem and define the notation used in the paper. Let us assume to have N sensors Ψ_n, with $n = [1, N]$, which can acquire data on two different geographical areas. The feature vector extracted by the nth sensor Ψ_n is defined as $\mathbf{x}^{\Psi_n} = (x_1^{\Psi_n}, x_2^{\Psi_n}, .., x_{z_n}^{\Psi_n})$, with $\mathbf{x}^{\Psi_n} \in \mathbb{R}^{z_n}$. Accordingly, the vector $\mathbf{x} \in \mathbb{R}^{\mathrm{d}}$ of features extracted by all the available N sensors is as follows:

$$\begin{aligned}\mathbf{x} &= (\mathbf{x}^{\Psi_1} \cup \mathbf{x}^{\Psi_2} \cup .. \cup \mathbf{x}^{\Psi_{\mathrm{N}}}) \\ &= (x_1^{\Psi_1}, .., x_{z_1}^{\Psi_1}, x_1^{\Psi_2}, .., x_{z_2}^{\Psi_2}, x_1^{\Psi_{\mathrm{N}}}, .., x_{z_{\mathrm{N}}}^{\Psi_{\mathrm{N}}})\end{aligned} \tag{1}$$

Let $\mathcal{D}_s$ be the source domain and $\mathcal{D}_t$ the target domain that we assume share the same set of M land-cover classes $\Omega = \{\omega_m\}_{m=1}^{M}$. Let $T_s = \{\mathbf{x}_{s,j}, y_{s,j}\}_j$ be the training set available on the source domain and $\{\mathbf{x}_{t,j}\}_j$ be the set of unlabeled samples of the target domain, where $\mathbf{x}_{s,j}, \mathbf{x}_{t,j} \in \mathbb{R}^{\mathrm{d}}$ and $y_{s,j} \in \Omega$. From the statistical view point, each class $\omega_m \in \Omega$ is characterized by the prior probability $P(\omega_m)$ and by its relation with the data through the class conditional probability $P(\mathbf{x}|\omega_m)$. Accordingly, the distribution of the source domain can be written as:

$$p_s(\mathbf{x}) = \sum_{\omega_m \in \Omega} P_s(\omega_m) P_s(\mathbf{x}|\omega_m) \tag{2}$$

and the distribution of the target domain:

$$p_t(\mathbf{x}) = \sum_{\omega_m \in \Omega} P_t(\omega_m) P_t(\mathbf{x}|\omega_m) \tag{3}$$

Due to the different scenes represented by $\mathcal{D}_s$ and $\mathcal{D}_t$ (e.g., different acquisition condition, different ground condition, different geographical locations), usually a shift in the probability distributions of the classes is observed (i.e., $p_s(\mathbf{x}) \neq p_t(\mathbf{x})$). The main goal of the DA method is to overcome this shift, in order to transfer knowledge from $\mathcal{D}_s$ to $\mathcal{D}_t$ in a reliable way. Typically, this is done by adapting the classifier model trained on $\mathcal{D}_s$ to $\mathcal{D}_t$ [54], by searching feature spaces where the data are more aligned from the global view point [55] or by matching local deformations [56], [57]. Differently from all the methods present in the literature, in the proposed approach the detection of these features subspaces relies on the physical meaning of the properties measured by the available sensors. When one (or more) sensor(s) takes a measure (or a subset of measures) that is not affected by the specific domain considered (e.g., the height value of a tree), in the associated feature subspace the probability distributions result almost statistically aligned. This is exploited by the proposed method to infer knowledge to $\mathcal{D}_t$ as described in the next section.

III. Proposed Sensor-driven Hierarchical DA Method

The acquisition of multisensor data results in the collection of complementary measurements on the scene. Due to the capability of each sensor of measuring different physical properties of the land-cover classes in many cases, it is possible to identify a set of properties which are expected to be almost invariant between $\mathcal{D}_s$ and $\mathcal{D}_t$ for one or more specific classes. A physical measure can be considered invariant for a class when the class presents almost the same distribution in the feature space defined by the measure in $\mathcal{D}_s$ and $\mathcal{D}_t$. This condition allows us to bridge the gap between source and target distributions. Fig. 1 shows the block-scheme of the proposed

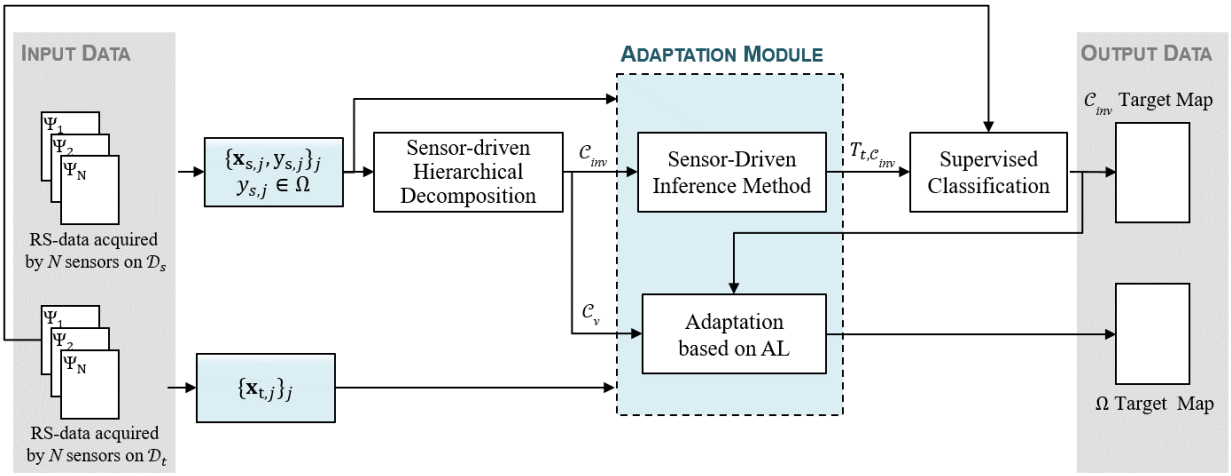


Fig. 1. Block scheme of the proposed sensor-driven hierarchical domain adaptation method.

DA approach. The main assumptions of the proposed method are: i) the availability of multisensor data acquired on $\mathcal{D}_s$ and $\mathcal{D}_t$, ii) the same set of land-cover classes is shared between the domains. Note that even though the assumption of having the same set of land-cover classes between the domains can be critical in some application, it has been extensively employed in the literature [9], [22], [58]–[62] and it is reasonable in many real cases.

A. Sensor-driven Hierarchical Decomposition

The first step, *Sensor-driven Hierarchical Decomposition*, seeks to decompose the DA problem to identify a subset of classes for which invariant measurements can be obtained. If for a subset of classes some features have almost the same behaviour in both the domains, for those classes it is possible to infer knowledge from $\mathcal{D}_s$ to $\mathcal{D}_t$ in an unsupervised but reliable way. To generate the hierarchical decomposition of the set of land-cover classes, first the invariant measures are detected. To this end, it is necessary to analyze the properties of both the considered classification problem and the sensors available. Note that the detection of the invariant feature subspaces can be performed by considering more sophisticated combination of features acquired by different sensors or by combining different features. The use of data-driven methods can identify features potentially effective but with results implicitly less reliable/robust given the fact that the DA problem is ill-posed [52], [53]. The exploitation of a physic-driven step can mitigate the possible drawback on the reliability of the results of data-driven methods. Thus, in the considered implementation, we focused the attention on the physic of both the signals recorded by the sensors and their interaction with the set of classes to detect the measures carachterized by invariant bahaviour. Many papers in the literature studied the interaction between RS sensors and the physical properties of the scene [63]–[65]. For example, in [63] the authors define a set of physical indices for urban scene to automatically classify high-resolution imagery in an unsupervised way, while in [64] similar indices are employed to extract the ground truth data directly from the high resolution image.

Procedure 1: Sensor-driven Hierarchical Decomposition

begin
inputs: The set of land-cover classes $\Omega = \{\omega_m\}_{m=1}^{M}$ and the set of sensors Ψ_n, $n = [1, N]$
for $\omega_m \in \Omega$ **do**
  Analyze the set of measures $\mathbf{x} = \{\mathbf{x}^{\Psi_1} \cup .. \cup \mathbf{x}^{\Psi_N}\}$ to detect the sensor Ψ_n (or the set of sensors) which provides an invariant feature subspace $\mathbf{x}^{\Psi_n}$
  if there is at least one invariant measure for ω_m **then**
    Insert ω_m into the set of classes $\mathcal{C}_{inv}$
  else
    Insert ω_m into the set of classes $\mathcal{C}_v$
  end if
end for
Define the number of hierarchical levels equal to the number of detected invariant feature subspaces. According to a top-down approach, starting from the whole set of classes Ω:
for each level of the hierarchy **do**
  select one of the detected invariant feature subspaces
  partition the meta-classes present in the hierarchical level into two (or more) disjoint meta-classes by keeping in the same meta-class the classes having similar behaviour in the considered invariant feature subspace.
end for
Partition the remaining meta-classes into $\omega_m \in \mathcal{C}_v$.
outputs: The hierarchical tree structure $\mathcal{C} = \{c_k\}_{k=1}^{N_k}$, $\omega_m \in \mathcal{C}_{inv}$ and $\omega_m \in \mathcal{C}_v$ and the related invariant feature subspaces identified at each level.

In light of this, the user expertise together with results published in the literature, should be employed to detect feature subspaces where a subset of classes is expected to be characterized by an invariant behaviour from the physical view point. Procedure 1 describes the steps defined to

generate the hierarchical tree structure. First, we perform an invariance analysis which takes into account the considered classification problem and the set of sensors available. The goal of this analysis is to evaluate, for each land-cover class, the physical measures collected by the sensors in order to identify if one (or more) measure(s) is (are) invariant for that class. For instance, the land-cover class "Tree" is expected to have almost always high height values regardless of the geographical location. Thus, the height measured by a LiDAR sensor can be considered a physical invariant measure for the "Tree" class. At the end of this phase, we identify the invariant feature subspaces and associate each class ω_m either to: i) a subset of classes $\mathcal{C}_{inv}$ for which invariant features exist, or ii) a subset of classes $\mathcal{C}_v$ for which none of the sensors available can measure invariant properties. In the second phase of the proposed procedure, we define the hierarchical tree structure by iteratively partitioning the multiclass problem into meta-class problems. Note that, the hierarchical structure allows us to decompose the DA problem, thus facilitating the identification of the most discriminant feature subspace (among the invariant ones) where each land-cover class can be accurately classified. To this end, a top-down approach is considered. At each level of the hierarchy a specific sensor, among the ones able to measure invariant properties, drives the definition of the meta-classes. Note that the order in which the sensors are used to define the hierarchy is not relevant, since the leaf-nodes (land-cover classes) are discriminated due to the joint use of the detected invariant features. Let Ψ_n be the sensor that drives the definition of the meta-classes at the first level of the hierarchy. The whole set of classes Ω is partitioned into two (or more) disjoint meta-classes so that the classes assigned to one meta-class are more similar to each other than to the classes associated to the other meta-classes in the considered invariant feature subspace $\mathbf{x}^{\Psi_n}$. Note that, if the considered sensor is able to measure invariant features only for one land-cover class, at that level of the hierarchy there will be a leaf node representing that land-cover class vs a meta-class including all the remaining classes. The process is repeated until all the detected invariant measures have been employed. Accordingly, all the $\omega_m \in \mathcal{C}_{inv}$ are identified as leaf-nodes, while the $\omega_m \in \mathcal{C}_v$ are assigned to the remaining meta-classes. Let us consider a hierarchical tree structure made up of the set of classes $\mathcal{C} = \{c_k\}_{k=1}^{N_k}$, where c_k can be a meta- or a data- class (i.e., ω_m). For all the hierarchy levels starting from two, each class c_k is connected to a unique parent-class and a set of child-classes $F(c_k) = \{c_{k_1}, c_{k_2}, .., c_{k_{f_k}}\}$, where f_k is the number of classes included in the meta-class c_k (see Fig. 2). At each level of the tree the PDFs of $\mathcal{D}_s$ and $\mathcal{D}_t$ for a feature vector $\mathbf{x}$ can be written as follows:

$$p_s(\mathbf{x}) = \sum_{c_k \in \mathcal{C}_v} P_s(c_k) P_s(\mathbf{x}|c_k) + \sum_{c_k \in \mathcal{C}_{inv}} P_s(c_k) P_s(\mathbf{x}|c_k) \quad (4)$$

and

$$p_t(\mathbf{x}) = \sum_{c_k \in \mathcal{C}_v} P_t(c_k) P_t(\mathbf{x}|c_k) + \sum_{c_k \in \mathcal{C}_{inv}} P_t(c_k) P_t(\mathbf{x}|c_k) \quad (5)$$

Note that the sets $\mathcal{C}_v$ and $\mathcal{C}_{inv}$ are different for different levels of the tree (for simplifying the notation we do not include explicitly the dependence). First, we adapt the set of classes $\mathcal{C}_{inv}$ by means of a sensor-driven inference method. Then, we address the adaptation of classes characterized by variant behaviour, $\mathcal{C}_v$, by using a standard AL technique.

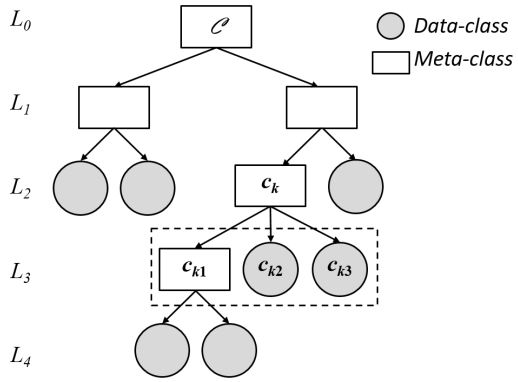


Fig. 2. Example of 5 levels hierarchical tree structure of 7 land-cover classes. The generic class c_k is represented connected to its child-classes $F(c_k) = \{c_{k1}, c_{k2}, c_{k3}\}$. In the considered example f_k is equal to three.

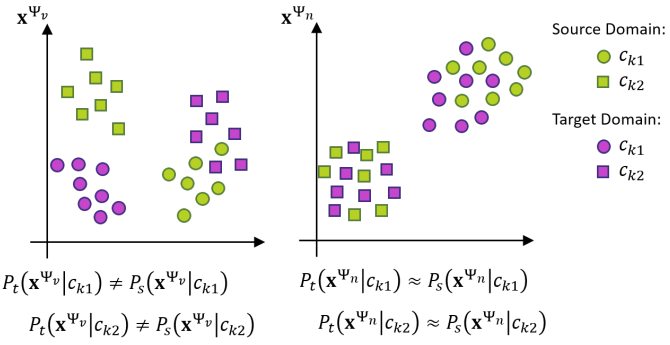


Fig. 3. Example of classes that result statistically aligned in the feature space $\mathbf{x}^{\Psi_n}$, which is invariant from the physical view point, while they are not aligned in a generic feature subspace $\mathbf{x}^{\Psi_v}$. Note that, the definition of invariant behaviour between the domains means that c_{k_1} and c_{k_2} present almost the same behaviour in $\mathbf{x}^{\Psi_n}$.

B. Sensor-driven Inference Method

The second step, *Sensor-driven Inference Method*, seeks to infer the labels of the $\mathcal{C}_{inv}$ classes from $\mathcal{D}_s$ to $\mathcal{D}_t$, thus generating a training set adapted to $\mathcal{D}_t$ for those classes. Note that no labeled samples are required from the target domain. Let us focus the attention on a generic meta-class c_k that includes 2 child-classes $\{c_{k_1}, c_{k_2}\}$. Without loosing generality, let us assume that from the previous step it turned out that one sensor Ψ_n, among the N available ones, provides an invariant features subspace $\mathbf{x}^{\Psi_n}$ where the classes $\{c_{k_1}, c_{k_2}\}$ can be discriminated. The conditional PDFs in that feature subspace $\mathbf{x}^{\Psi_n}$ result almost aligned (the probability distributions are similar), i.e.,:

$$\begin{aligned} P_t(\mathbf{x}^{\Psi_n}|c_{k_1}) \approx & P_s(\mathbf{x}^{\Psi_n}|c_{k_1}), \text{ and} \\ & P_t(\mathbf{x}^{\Psi_n}|c_{k_2}) \approx P_s(\mathbf{x}^{\Psi_n}|c_{k_2}) \end{aligned} \quad (6)$$

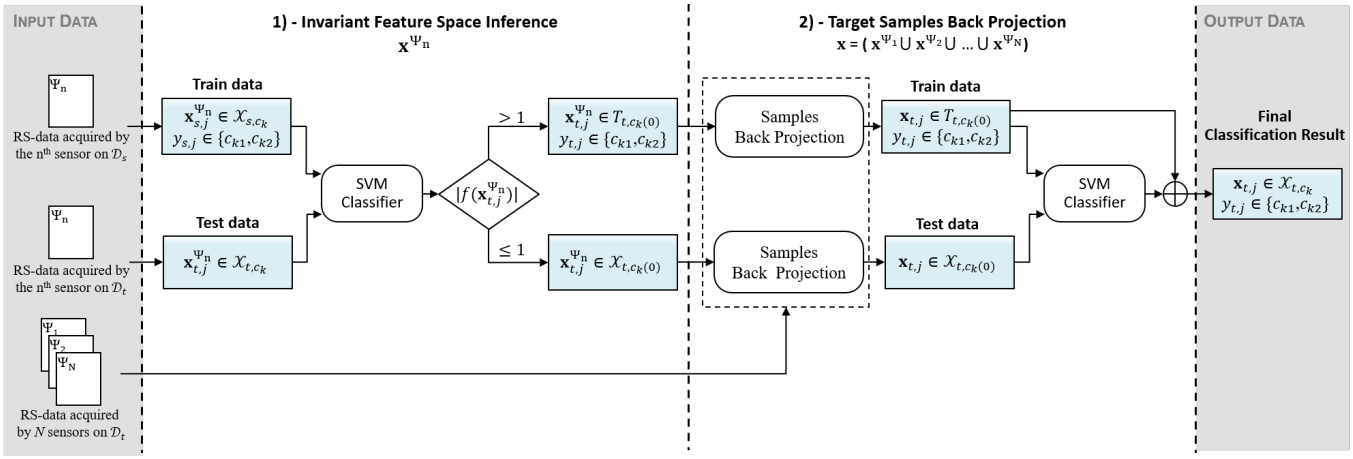


Fig. 4. Block scheme of the sensor-driven inference step. In the example, for the considered meta-class c_k the sensor Ψ_n provides an invariant features subspace $\mathbf{x}^{\Psi_n}$ where the classes $F(c_k) = \{c_{k1}, c_{k2}\}$ present almost the same behaviour and can be discriminated.

whereas they may be not aligned for all other feature subspaces:

$$P_t(\mathbf{x}^{\Psi_v}|c_{k_1}) \neq P_s(\mathbf{x}^{\Psi_v}|c_{k_1}), \text{ and}$$
$$P_t(\mathbf{x}^{\Psi_v}|c_{k_2}) \neq P_s(\mathbf{x}^{\Psi_v}|c_{k_2}) \quad (7)$$
$$\forall v \in [1, N], \text{ with } v \neq n$$

Fig. 3 shows the qualitative example of the considered classes statistically aligned in the invariant feature space $\mathbf{x}^{\Psi_n}$, while they are not aligned in a generic feature subspace $\mathbf{x}^{\Psi_v}$. The alignment condition of the data allows us to transfer the knowledge from $\mathcal{D}_s$ to $\mathcal{D}_t$ in an unsupervised but reliable way. In greater detail, this is accomplished by means of two main steps. First, we infer class labels of the target samples that result more aligned to the source samples (i.e., that have the highest probability of being correctly labeled) in the invariant feature subspaces. Then, we analyze the labeled target samples in $\mathbf{x} = (\mathbf{x}^{\Psi_1} \cup \mathbf{x}^{\Psi_2} \cup .. \cup \mathbf{x}^{\Psi_N})$ to accurately model the target distribution of the considered classes in the entire feature space, thus labeling the remaining samples of that classes (see Fig. 4).

1) Invariant Feature Space Inference: the proposed inference strategy exploits a classifier trained on $\mathcal{D}_s$ in the feature subspace $\mathbf{x}^{\Psi_n}$ to predict the labels of $\mathcal{D}_t$ for the invariant classes c_{k_1} and c_{k_2}. Due to the hierarchical tree structure, we focus the attention only on the unlabeled samples of $\mathcal{D}_t$ belonging to the meta-class c_k of the considered level of the tree. While at the first level of the hierarchy the entire set of unlabeled samples is considered, for all the levels starting from two we inherit the samples of the identified meta-class c_k. Let $\mathcal{X}_{s,c_k}$ and $\mathcal{X}_{t,c_k}$ be the source and the target samples belonging to the class c_k, respectively. To transfer the knowledge between the domains, in our study we consider the SVM classifier, extensively employed for the classification of RS [66]–[69]. However, since the data alignment is due to the invariance of the feature space, any other classification technique can be employed. Let $T_{s,c_k} = \{(\mathbf{x}_{s,j}^{\Psi_n}, y_{s,j}) | \ \mathbf{x}_{s,j}^{\Psi_n} \in \mathcal{X}_{s,c_k}\}$ be the source training set for the considered class c_k represented in the detected invariant feature space $\mathbf{x}^{\Psi_n}$, with $\mathbf{x}_{s,j}^{\Psi_n} \in \mathbb{R}^{z_n}$ and $y_{s,j} \in \{c_{k_1}, c_{k_2}\}$. The SVM classifier trained in $\mathbf{x}^{\Psi_n}$ with the labeled samples of the source domain T_{s,c_k} is applied to the unlabeled target samples $\mathbf{x}_{t,j}^{\Psi_n} \in \mathcal{X}_{t,c_k}$. The labels of the target samples are predicted according to the sign of the discriminant function $f(\mathrm{x}^{\Psi_n})$ associated to the SVM hyperplane, i.e., sign$[f(\mathrm{x}^{\Psi_n})]$. Although in the considered feature subspace $\mathbf{x}^{\Psi_n}$ the data are more aligned that in the original space, this subspace is not expected to be sufficient to properly model the entire target PDFs of c_{k_1} and c_{k_2}. For this reason, we generate an initial training set for $\mathcal{D}_t$ by considering only the samples that fall outside the SVM margin (i.e., $|f(\mathrm{x})| > 1$) because they are the ones having the highest probability of being correctly classified (see Fig. 5). Let $T_{t,c_k(0)} = \{(\mathbf{x}_{t,j}^{\Psi_n}, y_{t,j}) | \ \mathbf{x}_{t,j}^{\Psi_n} \in \mathcal{X}_{t,c_k}, \ |f(\mathbf{x}_{t,j}^{\Psi_n})| > 1\}$ be the initial training set of $\mathcal{D}_t$ generated by transferring the knowledge in $\mathbf{x}^{\Psi_n}$ for the invariant classes c_{k_1} and c_{k_2}. Let $\mathcal{X}_{t,c_k(0)} = \{\mathbf{x}_{t,j}^{\Psi_n} | \ \mathbf{x}_{t,j}^{\Psi_n} \in \mathcal{X}_{t,c_k}, \ |f(\mathbf{x}_{t,j}^{\Psi_n})| \leq 1\}$ be the remaining set of unlabeled target samples.

2) Target Samples Back Projection: the obtained initial target training set $T_{t,c_k(0)}$ is analyzed in the entire feature space to represent the PDFs of the considered classes on all the available features (i.e., $P_t(\mathbf{x}|c_{k_1})$ and $P_t(\mathbf{x}|c_{k_2})$). This condition allows us to use all the available information to ensure an accurate classification of the remaining unlabeled samples. Then, the classifier is trained on the initial training set $T_{t,c_k(0)}$ in the entire feature space to classify the remaining unlabeled samples $\mathcal{X}_{t,c_k(0)}$, thus obtaining the final target training set for the considered class c_k, i.e., $\{(\mathbf{x}_{t,j}, y_{t,j}) | \mathbf{x}_{t,j} \in \mathcal{X}_{t,c_k}\}$. Note that at each level of the hierarchy, the adaptation is performed by classifying the target samples in the invariant feature subspace using the source reference data. Thus, we are adapting the reference data available in the $\mathcal{D}_s$ to $\mathcal{D}_t$ via the invariant feature subspaces. When the adaptation of c_{k_1} and c_{k_2} is complete, we can focus the attention on their child classes, i.e., $F(c_{k_1})$ and $F(c_{k_2})$. Due to the hierarchical decomposition of the DA problem, the adaptation of $F(c_{k_1})$ and $F(c_{k_2})$ is address separately since they belong to different meta-classes. If $F(c_{k_1})$ and $F(c_{k_2})$ have been identified as invariant classes $\mathcal{C}_{inv}$, the adaptation is performed by using the sensor-driven inference method, otherwise the adaptation

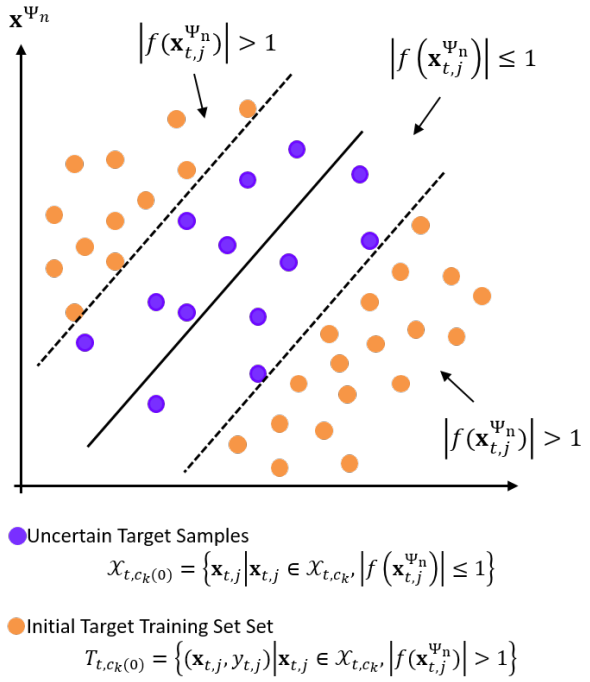


Fig. 5. Classification result obtained on the unlabeled target samples $\mathcal{X}_{t,c_k}$ by using the labeled samples of the source domain T_{s,c_k} in the detected invariant feature space $\mathbf{x}^{\Psi_n}$. According to the decision function $|f(\mathbf{x}_{t,j}^{\Psi_n})|$, the set of unlabeled target samples is divided into $T_{t,c_k(0)}$ and $\mathcal{X}_{t,c_k(0)}$.

is based on AL method (see Section III-D).

C. Supervised Classification of the Target Domain

By applying the label inference method to all the classes $\mathcal{C}_{inv}$, in the previous step we generate an initial training set for the target domain $T_{t,\mathcal{C}_{inv}} = \{(\mathbf{x}_{t,j}, y_{t,j})\}_j$ with $\mathbf{x}_{t,j} \in \mathbb{R}^d$ and $y_{t,j} \in \mathcal{C}_{inv}$. By using the considered training set, we classify $\mathcal{D}_t$ in the entire feature space, thus generating a classification map representing all the invariant data-classes $\omega_m \in \mathcal{C}_{inv}$ and the target samples belonging to the variant classes $\mathcal{X}_{t,\mathcal{C}_v}$. In the considered implementation we employed the SVM classifier by considering the One Against All (OAA) multiclass strategy [19]. Note that, this classifier is at the state of the art of the RS classification data because of its high generalization capability [66], [67], high classification accuracy when compared with other classifiers and effectiveness in handling ill-posed problems (i.e., low ratio between the number of training samples and the number of features) [68], [69]. By inferring the knowledge on the classes $\mathcal{C}_{inv}$, we simplify the adaptation problem by: i) reducing the number of classes that should be adapted in the next step, and ii) introducing constraints in the adaptation of the remaining classes $\mathcal{C}_v$ that increase the reliability of adaptation based on AL method. The amount of simplification depends on the number of classes $\mathcal{C}_{inv}$ on which we can detect invariant feature subspaces between $\mathcal{D}_s$ and $\mathcal{D}_t$.

D. Adaptation based on Active Learning

The last step of the proposed method, *Adaptation based on AL*, aims to complete the land-cover map of $\mathcal{D}_t$ generated at the previous step, by integrating the classes $\omega_m \in \mathcal{C}_v$. This is done by obtaining the labels of few informative samples for generating a representative training set of $\mathcal{D}_t$. To minimize the number of training samples of $\mathcal{D}_t$ we exlpoit an AL methods. AL methods are employed in the supervised classification of RS data to optimize the definition of the training set by selecting the most informative samples. The initial training set is iteratively expanded by means of an interactive procedure which involves a supervisor (i.e., a human expert) who correctly assigns the labels to the selected uncertain (i.e., informative) samples [70]–[72].

The AL strategy employed in the proposed method focuses the attention only on the $\mathcal{C}_v$ classes by taking advantage from: i) the hierarchical decomposition of the DA problem, ii) the adaptation performed in the previous step on the classes $\mathcal{C}_{inv}$. Due to the hierarchical decomposition of the classes, we are in the condition of identifying the target samples belonging to the variant classes $\mathcal{X}_{t,\mathcal{C}_v}$. This condition allows us to refine the target training set used to classify the invariant classes $\mathcal{C}_{inv}$ by integrating only unlabeled target samples belonging to the classes $\omega_m \in \mathcal{C}_v$.

It is worth mentioning that the proposed method can use any kind of AL strategy. Here we consider the marginal sampling by closest support vector (MScSV) [73]. This very simple tecnique selects the sample, lying within the margin, closest to the hyperplane of the SVM classifier. In the multiclass strategy, the MScSV considers the most uncertain sample for each binary SVM according to the one-against-all (OAA) architecture [74]. Please note that more sophisticated techniques can be employed. However, in the considered operational scenario we selected a simple AL strategy to avoid the tuning of any parameter. At the end of this step, we generate the land-cover map of $\mathcal{D}_t$ by means of the training set obtained with the inference method integrated to the AL approach.

IV. Datasets Description and Design of Experiments

A. Datasets Description

The proposed DA method was evaluated experimentally on two different multisensor datasets:

1) Forest Dataset: in this dataset (see Fig. 6), we considered two spatially disjoint forest areas located in the Southern Italian Alps, Trentino region. The first study area is located in Val di Sella (1090 Ha), whereas the second one is located in Padergnone (175 Ha), hereafter referred as *Vds* and *Pad*, respectively. The sensors available were a hyperspectral (HSI) sensor, a LiDAR system and a color camera. The HSI presents a spectral range between 402.9 nm and 989.1 nm and a spatial resolution of 1 m. In *Vds*, the acquisition was performed on 16th July 2008 with a spectral resolution of 4.6 nm (126 spectral bands), whereas in *Pad* the data were taken on 4th September 2007 with a spectral resolution of 9.2 nm (63 spectral bands). The *Vds* image was subsampled by applying a Gaussian model with a Full Width at Half Maximum (FWHM) equal to the band spacings (i.e., 9.2 nm) by matching the corresponding wavelength ranges. LiDAR data were acquired jointly with the HSI with an average point density of 5

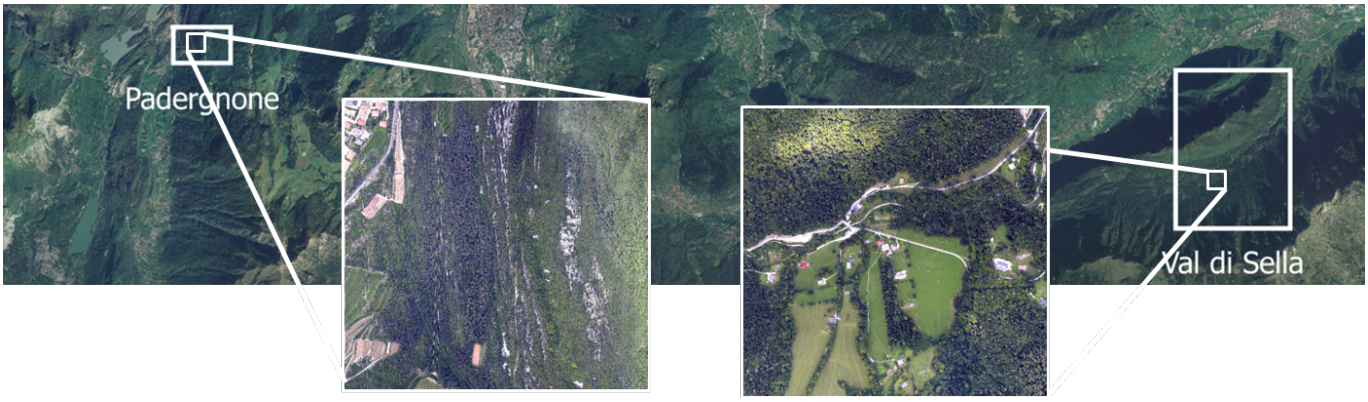


Fig. 6. Forest dataset: color composition of the orthophoto acquired on a portion of the Trentino region. The study areas are highlighted in the white rectangles overlapped on the optical image. A small portion of the high-resolution optical images of the dataset is represented for both the study areas.

pts/m^2. The Digital Terrain Model (DTM) was produced and subtracted from the LiDAR data to obtain the relative height of the targets with respect to the terrain. The obtained LiDAR point cloud was rasterized to generate the Canopy Height Model (CHM) image. The high resolution optical image has 3 spectral bands acquired in the visible range (RGB) with a spatial resolution of 0.2 m. The spatial resolution of the optical image was degraded to 1 m to be coherent with the HSI and the CHM, using the nearest neighbour re-sampling. The multi-sensor data were manually coregistered by guaranteeing a data shift up to 1 m. To this end, we used as reference targets the buildings present in the scene. The considered domains share nine classes: 3 Conifers species (i.e., Norway Spruce, European Larch, Scots Pine), 2 Broadleaves species (i.e., European Beech, Hop Hornbeam), Buildings, Roads, Soil and Grass.

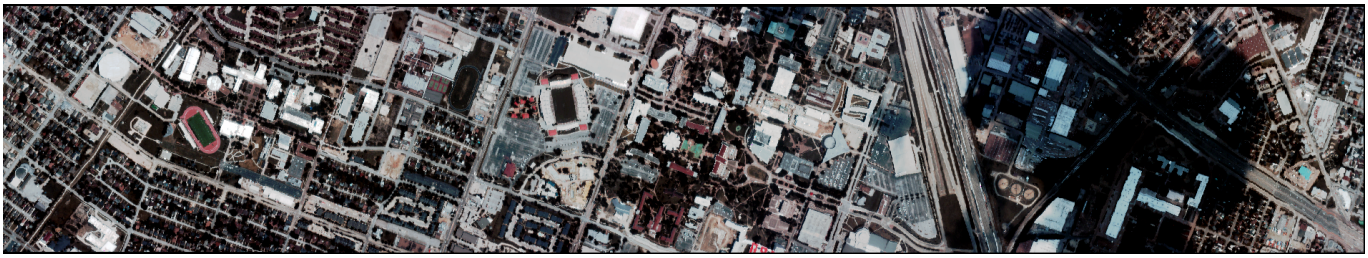
2) Urban Dataset: in this dataset (see Fig. 7) we considered as first domain a multisensor data acquired over the University of Houston campus and the neighboring urban area [75]. The data consist of an HSI image acquired at the spatial resolution of 2.5 m, having 144 spectral bands in the spectral range between 380 nm and 1050 nm and a LiDAR point cloud acquired over the same scene. By processing the LiDAR data a CHM was generated at the same spatial resolution of the HSI data. The LiDAR data were acquired on 22nd June 2012, while the HSI data were acquired on 23rd June 2012. The second domain is a urban area located in Pellizzano, Trentino region. The multisensor data available are an HSI image having 1 m spatial resolution and, 65 spectral bands in the range between 400 nm and 990 nm and LiDAR data acquired with an average point density $>$ 10 pts/m^2. By processing the LiDAR data a CHM of 1 m spatial resolution was generated to represent the elevation value of the object with respect to the ground. The HSI data were acquired on 23rd June 2013, while the LiDAR data were acquired in September 2012. To perform the adaptation, the spectral channels of the two HSI images were compared, thus extracting from the 144 spectral bands of the Houston (*Hst*) dataset the corresponding 65 of the Pellizzano (*Plz*) dataset. Starting from the ground reference data available on the *Hst* dataset, we took six thematic classes which are present also in the *Plz* dataset for our classification tasks: Buildings, Roads, Trees, Soil, Water and Grass. Please note that in this dataset we considered RS data acquired by different sensors and in completely different environmental conditions.

B. Experimental Setup

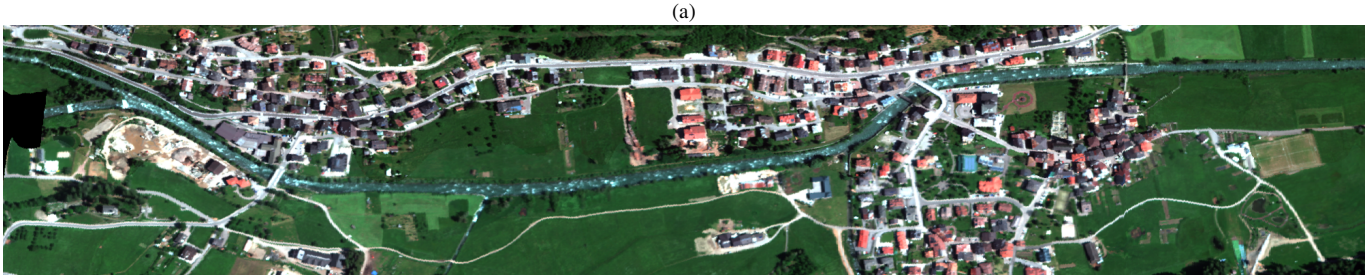
To assess the performance of the proposed approach, all the areas were considered as $\mathcal{D}_s$ and $\mathcal{D}_t$. Tab. I reports the number of samples per area divided in Training (TR) and Test (TS) and pool sets for the Forest Dataset. When an area is considered as $\mathcal{D}_s$, the training set is exploited to infer the knowledge on the unlabeled samples of $\mathcal{D}_t$. The test set of the other area (i.e., $\mathcal{D}_t$) is used for the accuracy assessment and the pool set is the unlabeled set of samples used by the AL technique. Tab. II reports the number of samples per area divided in Training (TR) and Test (TS) for the Urban Dataset. In this case, no AL was performed since for all the considered classes invariant feature were measured, thus we did not generate a pool set of samples. Four random datasets per area have been generated and the average results obtained on the test set of the $\mathcal{D}_t$ for four trials are reported.

In all the experiments carried out, a radial basis function (RBF) kernel was adopted in the SVM classifier. In the experiments with AL methods, at the first iteration the model selection phase was performed using a grid strategy on the validation set of $\mathcal{D}_s$, thus tuning the RBF kernel width and the SVM regularization parameter. The same grid strategy was used to tune the model parameters for the standard supervised SVM. The proposed method was compared to the state-of-the-art DA method Transfer Component Analysis (TCA) [76], the semisupervised LapSVM method [29] and the Geodesic Flow Kernel (GFK) [41]. We applied a grid strategy to tune the parameters of the considered baselines, thus selecting the ones that resulted in the highest classification accuracy on the test set of $\mathcal{D}_t$. The samples of $\mathcal{D}_s$ were used for training the LapSVM, the TCA and the GFK to perform the classification of $\mathcal{D}_t$. The classification results were evaluated in terms of Overall Accuracy (OA), Producer Accuracy (PA) and User Accuracy (UA).

To quantitatively evaluate the dissimilarity between $\mathcal{D}_s$ and $\mathcal{D}_t$ in the invariant feature subspaces $\mathbf{x}^{\Psi_n}$ we calculate per class



(a)



(b)

Fig. 7. Urban dataset: (a) false color composition of the HSI of the Houston data, (b) false color composition of the HSI of the Pellizzano data.

TABLE I
NUMBER OF AVAILABLE LABELED SAMPLES OF THE LAND-COVER CLASSES IN THE SOURCE AND THE TARGET DOMAINS FOR THE FOREST DATASET.

Class Name	**Number of samples**					
	Vds			Pad		
	TR	**TS**	**Pool**	**TR**	**TS**	**Pool**
Norway Spruce (ω_1)	478	238	239	43	21	21
Silver Fir (ω_2)	378	189	189	105	52	53
European Larch (ω_3)	373	186	187	388	194	194
European Beech (ω_4)	512	255	256	815	407	408
Hop Hornbeam (ω_5)	83	41	41	145	72	72
Grass (ω_6)	178	89	89	172	86	86
Building (ω_7)	143	71	71	137	68	68
Roads (ω_8)	148	73	74	137	68	68
Soil (ω_9)	190	94	95	200	99	100

TABLE II
NUMBER OF AVAILABLE LABELED SAMPLES OF THE LAND-COVER CLASSES IN THE SOURCE AND THE TARGET DOMAINS FOR THE URBAN DATASET.

Class Name	**Number of samples**			
	Hst		Plz	
	TR	**TS**	**TR**	**TS**
Building (ω_1)	194	193	990	990
Soil (ω_2)	93	93	430	430
Grass (ω_3)	194	194	474	474
Roads (ω_4)	192	192	496	496
Water (ω_5)	91	91	831	831
Trees (ω_6)	94	94	410	410

the Jeffrey-Matusita (JM) distance. Let us focus on the class c_{k_1}, the JM distance between the class conditional probability $P_s(\mathbf{x}^{\Psi_n}|c_{k_1})$ and $P_t(\mathbf{x}^{\Psi_n}|c_{k_1})$ can be computed as follows:

$$JM_{k_1} = \sqrt{2(1 - e^{-\mathbf{B}_{k_1}})} \tag{8}$$

where $\mathbf{B}_{k_1}$ is the Bhattacharyya distance, i.e.,:

$$\mathbf{B}_{k_1} = -ln\left\{ \int_x \sqrt{P_s(\mathbf{x}^{\Psi_n}|c_{k_1}) P_t(\mathbf{x}^{\Psi_n}|c_{k_1})} \right\} \tag{9}$$

Adopting the assumption that the classes follow multivariate Gaussian distributions, the Bhattacharyya distance $\mathbf{B}_{k_1}$ can be rewritten as follows:

$$\mathbf{B}_{k_1} = \frac{1}{8}(\boldsymbol{\mu}^{\boldsymbol{s}}_{\boldsymbol{k_1}} - \boldsymbol{\mu}^{\boldsymbol{t}}_{\boldsymbol{k_1}})^{\mathrm{T}} \left(\frac{\Sigma^s_{k_1} + \Sigma^t_{k_1}}{2} \right)^{-1} (\boldsymbol{\mu}^{\boldsymbol{s}}_{\boldsymbol{k_1}} - \boldsymbol{\mu}^{\boldsymbol{t}}_{\boldsymbol{k_1}}) + \\ + \frac{1}{2} ln \left(\frac{1}{2} \frac{|\Sigma^s_{k_1} + \Sigma^t_{k_1}|}{\sqrt{|\Sigma^s_{k_1}||\Sigma^t_{k_1}|}} \right) \tag{10}$$

where $\boldsymbol{\mu}^{\boldsymbol{s}}_{\boldsymbol{k_1}}$ and $\boldsymbol{\Sigma}^s_{k_1}$ are the mean and the covariance matrix of the class c_{k_1} in $\mathcal{D}_s$, while $\boldsymbol{\mu}^{\boldsymbol{t}}_{\boldsymbol{k_1}}$ and $\boldsymbol{\Sigma}^t_{k_1}$ are the mean and the covariance matrix of the class c_{k_1} in $\mathcal{D}_t$.

V. EXPERIMENTAL RESULTS

In this section, the experimental results obtained on both the datasets are presented. In particular, the hierarchical tree structures and the results obtained for both $\mathcal{C}_v$ and $\mathcal{C}_{inv}$ classes are reported.

A. Results on the Forest Dataset

1) Sensor-driven Hierarchical Decomposition: Fig. 8 shows the hierarchical tree structure defined for the considered DA problem. The structure is derived according to the procedure described in Procedure 1. First, we identify the physical measures that can be considered almost invariant across the domains for each land-cover class. The HSI sensor provides the Normalized Difference Vegetation Index (NDVI), the LiDAR sensor records the elevation values (H), while the high resolution optical images allows the identification of textural features, i.e., homogeneity textural feature derived by processing the Green band and mean and variance of Red, Green and Blue spectral channels. At the first level of the hierarchy, we focus the attention on the LiDAR sensor. The sensor is able to measure the relative height value (H) of the objects present in the scene with respect to the ground. In the related feature subspace, we can partition the whole set of classes Ω into two meta-classes: the first one ("Elevated") aggregates classes characterized by high height values (i.e., Trees and Building), while the second one ("Flat") includes classes having almost zero H values (i.e., Soil, Grass and Roads). At the second level of the hierarchy, the NDVI measure provided by the HSI sensor is considered. The "Elevated" meta-class can be partitioned into two meta-classes and one leaf-node class: i) the meta-class characterized by high positive NDVI values (i.e., European Beech and Hop Hornbeam) denoted as "Broadleaves"; ii) the meta-class characterized by medium positive NDVI values (i.e., Scots Pine, Norway Spruce and European Larch) denoted as "Conifers"; and iii) the leaf-node class having negative NDVI values (i.e., Building). The "Flat" meta-class can be partitioned into two meta-classes, one having negative NDVI values (i.e., Roads) and the second one having non-negative NDVI values (i.e., Soil and Grass). Note that, the land-cover classes are automatically identified as leaf nodes by the defined procedure due to the joint use of the invariant features. Thus, the Building class is the only one having high H values and negative NDVI, while the Roads class is the only one having low H values and negative NDVI. At the last level of the hierarchy, the textural measures provided by the high resolution optical image are used to partition the "Land" meta-class into Soil and Grass. The hierarchical tree structure is finally completed by partitioning the remaining meta-classes "Conifers" and "Broadleaves" into the $\omega_m \in \mathcal{C}_v$ classes for which none of the sensors measures invariant features.

2) Sensor-driven Inference Method: in the considered scenario, the detected $\mathcal{C}_{inv}$ classes are Conifers, Broadleaves, Soil, Roads, Grass and Building. A representation of the DA problem is shown in Fig. 9. Due to the different acquisition conditions of the source and the target images, even though the two domains appears similar, the feature subspaces represented in Fig. 9c and Fig. 9d point out a strong shift in the distributions. In contrast, by relying on the physical properties of the classes, the feature space defined by the NDVI and the H results in a good statistical alignment of the distributions (see Fig. 9a and Fig. 9b). These results are confirmed by the quantitative evaluation of Tab. III, where the class-wise JM distance between *Pad* and *Vds* is presented considering the invariant feature space $\mathbf{x}^{\Psi_n}$, the first 3 components of the principal component analysis (PCA) and the whole feature space $\mathbf{x}$. The obtained result confirms the selected invariant feature spaces for the considered forest DA problem. Tab. IVa and Tab. IVb show the classification results obtained when *Pad* and *Vds* are $\mathcal{D}_s$, respectively. The proposed sensor-driven DA method SVM^{inf} is compared with the SVM classifier trained on $\mathcal{D}_s$ and the DA TCA and GFK methods and the semisupervised LapSVM method. By directly applying to $\mathcal{D}_t$ the standard SVM trained on $\mathcal{D}_s$, the shift in the sample distributions strongly affects the classification accuracies of Grass, Roads, Building and Soil. This effect is encountered in both the DA problems, thus generating very low PA and UA. In particular, the Soil class is completely misclassified due to the strong shift of the class distribution (see Fig. 9c and Fig. 9d). By focusing the attention on Conifers and Broadleaves, the labeled samples of *Pad* properly represent the unlabeled samples of *Vds*. In contrast, by considering *Vds* as $\mathcal{D}_s$, the sample distributions of the same classes are not effective in representing the *Pad* distributions.

The TCA, the LapSVM and the GFK methods sharply increase the OA with respect to the use of the supervised SVM classifier. However, the methods are not able to handle the adaptation of Soil, Roads and Grass classes as proven by the obtained PA and UA on $\mathcal{D}_t$. In contrast, the proposed SVM^{inf} method accurately classifies $\mathcal{D}_t$. The obtained PA and UA are all higher than 80% for both the adaptation problems, thus

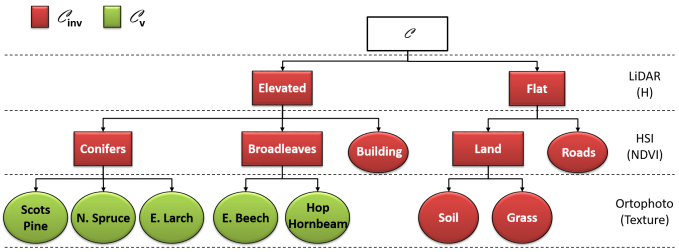


Fig. 8. Hierarchical tree structure derived at the end of the invariance analysis for the considered Forest Dataset. The invariant feature subspaces used to derive the meta-classes are reported for each level of the hierarchy.

TABLE III
JM DISTANCE BETWEEN *Pad* AND *Vds* COMPUTED PER CLASS CONSIDERING: (I) THE INVARIANT FEATURE SPACE $\mathbf{x}^{\Psi_n}$, (II) THE FIRST 3 COMPONENTS OF THE PRINCIPAL COMPONENT ANALYSIS (PCA) AND, (III) THE WHOLE FEATURE SPACE $\mathbf{x}$.

$\mathcal{C}_{inv}$	JM distance		
	$\mathbf{x}^{\Psi_n}$	PCA	$\mathbf{x}$
Conifers	0.259	1.024	1.414
Broadleaves	0.427	1.409	1.414
Grass	0.859	1.414	1.414
Building	0.838	1.273	1.414
Roads	0.480	1.379	1.414
Soil	0.558	1.381	1.414

TABLE IV
AVERAGE CLASSIFICATION RESULTS (OVER FOUR TRIALS): (a) *Pad* IS THE SOURCE DOMAIN; (b) *Vds* IS THE SOURCE DOMAIN. OA%, PA% AND UA% OBTAINED BY APPLYING: 1) THE SVM CLASSIFIER TRAINED ON THE SOURCE DOMAIN; 2) THE DA TCA METHOD; 3) THE SEMISUPERVISED LAPSVM; 4) THE DA GFK METHOD; 5) THE PROPOSED SENSOR-DRIVEN DA METHOD SVM^{inf}.

Classes	**Standard**								**Proposed**	
	SVM		**TCA**		**LapSVM**		**GFK**		**SVM^{inf}**	
	PA%	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**
Conifers	81.25	93.49	90.56	95.99	79.85	91.03	93.56	80.91	90.77	95.65
Broadleaves	85.47	73.49	91.05	75.12	88.01	68.42	71.28	84.46	93.24	83.57
Grass	93.25	45.91	43.54	58.94	51.69	43.19	41.73	87.92	100	100
Building	21.12	100	13.03	36.63	37.68	81.68	87.64	27.46	93.66	97.44
Roads	76.63	46.03	65.75	42.57	79.45	95.87	60.74	72.60	92.47	99.26
Soil	0	0	96.01	96.27	56.12	45.97	20.09	11.70	100	96.41
OA %	72.87		81.61		75.43		73.29		93.05	

(a)

Classes	**Standard**								**Proposed**	
	SVM		**TCA**		**LapSVM**		**GFK**		**SVM^{inf}**	
	PA%	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**
Conifers	74.19	37.26	100	65.7	100	48.71	75.89	85.53	97.07	84.44
Broadleaves	46.72	76.59	81.09	99.15	62.44	99.66	91.44	84.23	90.14	98.26
Grass	42.44	98.64	30.23	100	2.62	100	70.91	45.35	99.13	97.71
Building	58.08	40.10	50.74	27.88	33.46	24.4	33.51	45.96	99.26	98.90
Roads	21.69	15.16	38.24	35.62	41.54	33.83	21.80	41.91	100	80.24
Soil	0	0	2.27	4.48	23.48	48.44	0	0	82.07	100
OA %	47.77		69.31		59.78		68.14		93.04	

(b)

sharply increasing the classification accuracies of the critical classes. Note that similar OA results for both the DA problems (i.e., 93.05% and 93.04%) are achieved, regardless of the accuracies obtained by directly applying the classifier (i.e., 72.87% and 47.77%).

3) Adaptation based on AL: Although the adaptation performed in the previous step achieves an accurate classification map of $\mathcal{D}_t$ for the invariant classes without any labeling cost, the AL technique is integrated in the inference method for addressing the adaptation of the remaining forest species (i.e., $\mathcal{C}_v$ classes). Tab. Va and Tab. Vb show the classification accuracies obtained on $\mathcal{D}_t$ for each class by applying the supervised SVM classifier trained on $\mathcal{D}_s$, the standard AL method for DA, SVM$_{AL}$, and the proposed inference method SVM$^{inf}_{AL}$ integrated with the MScSV AL method. For both the experiments 45 samples were added by means of the AL methods. Note that the MScSV at each iteration selects a sample for each binary classifier of the OAA multiclass strategy. While the standard MScSV AL selects one sample per class (i.e., at each iteration 9 samples are added) the SVM$^{inf}_{AL}$ selects samples only for the $\mathcal{C}_v$ classes, (i.e., at each iteration 5 samples are added).

Let us focus the attention on the results obtained when *Pad* is $\mathcal{D}_s$ (see Tab. Va). The PM improves the OA accuracy of approximatively a 13% with respect to the standard SVM$_{AL}$ and of a 32% with respect to the SVM. Moreover, the PA and the UA of the forest species sharply increased compared to the ones obtained with the SVM$_{AL}$. Therefore, due to the accurate adaptation performed in the previous step for the classes Grass, Building, Roads and Soil, by requiring the labels of few samples for the remaining classes we obtained an accurate land-cover map of the $\mathcal{D}_t$. Tab. Vb shows similar results when *Vds* is the source domain. The PM improves the OA accuracy of 8% with respect to the SVM$_{AL}$ and of 40% with respect to the SVM. Fig. 10a and Fig. 10b depict the average (on four trials) classification accuracies obtained on $\mathcal{D}_T$ versus the number of new labeled samples with the proposed method SVM$^{inf}_{AL}$ and the standard AL method SVM$_{AL}$ considering *Pad* and *Vds* as $\mathcal{D}_s$, respectively. The results obtained confirm that the proposed approach allows a significant reduction in the number of target labeled samples, and thus in the labelling cost, required for obtaining a given classification accuracy. As an example, when *Pad* is the source domain, to reach an OA of 70% our method requires 15 samples whereas the SVM$_{AL}$ needs 80 samples (see Fig. 10a). To reach the same OA when *Vds* is the source domain, SVM$^{inf}_{AL}$ requires 10 samples, whereas the SVM$_{AL}$ requires 45 samples (see Fig. 10b). Furthermore, even tough the AL exploited in the PM selects samples only from the variant set of classes $\mathcal{C}_v$, by adding a high number of labeled samples from the $\mathcal{D}_t$

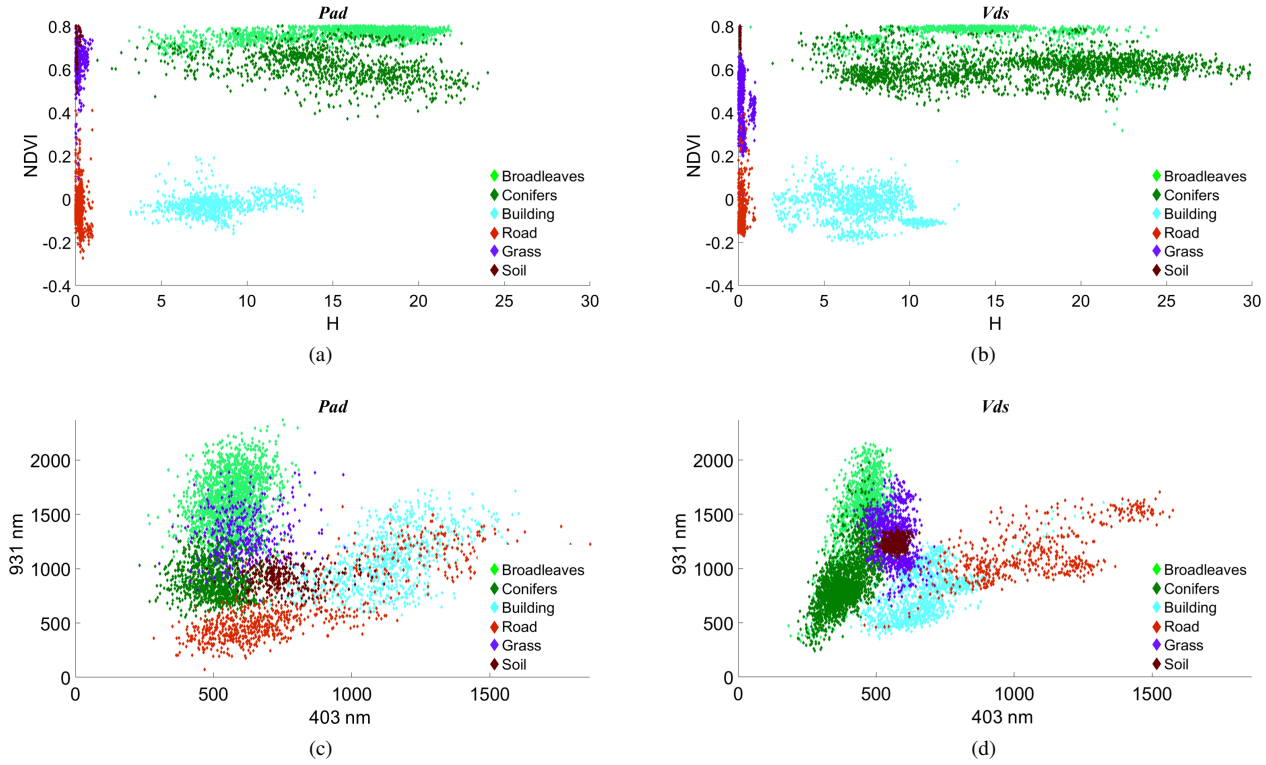


Fig. 9. Distributions of the labeled samples of *Pad* (left) and *Vds* (right) represented in: (a-b) the invariant feature subspace defined by the NDVI (HSI scanner) and the H (LiDAR sensor); and (c-d) the feature subspace defined by the spectral channels at the wavelength of 403 nm and 931 nm, respectively.

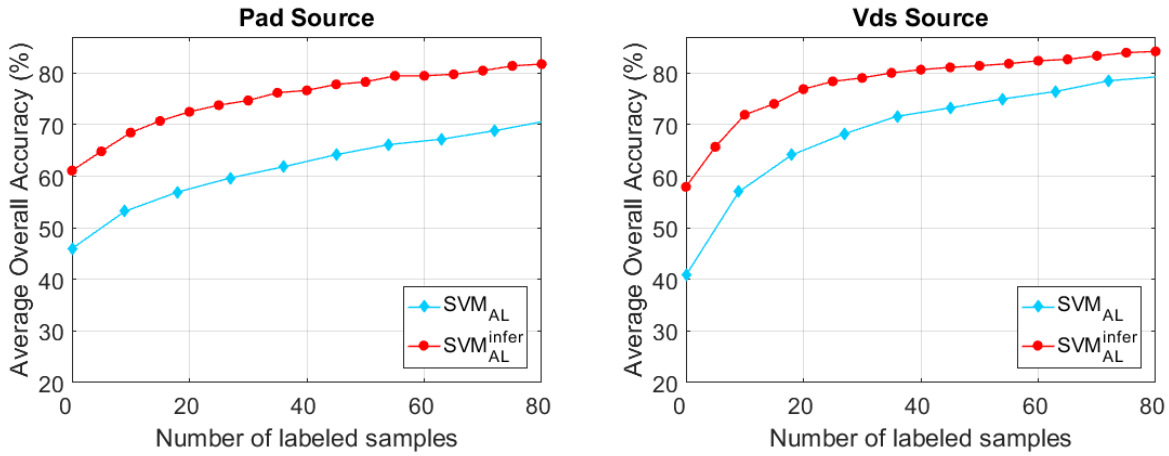


Fig. 10. Average (over four trials) Overall accuracy (%) vs the number of target labeled samples annotated by AL with the proposed method (SVM_{AL}^{infer}) and the standard DA AL method (SVM_{AL}) considering: (a) *Pad* as source domain; (b) *Vds* as source domain.

the standard AL method still achieves comparable OA. Thus, the results obtained confirm the effectiveness of the inference method in adaptation the invariant set of classes $\mathcal{C}_{inv}$.

B. Results on the Urban Dataset

1) Sensor-driven Hierarchical Decomposition: Fig. 11 shows the hierarchical tree structure defined for the urban DA problem according to the procedure described in Procedure 1. The HSI sensor provides the NDVI and the Normalized Difference Water Index (NDWI), while from the LiDAR data we extract the elevation value with respect to the ground (H). Moreover, we derive the homogeneity textural feature from the Green spectral band of the HSI image. At the first level of the hierarchy, the LiDAR sensor drives the definition of the meta-classes, thus partitioning the whole set of classes Ω into classes characterized by high height values (i.e., Trees and Building), denoted as "Elevated", and classes having low H values (i.e., Soil, Water, Grass and Roads), denoted as "Flat". At the second level of the hierarchy, the HSI sensor is used to partition the meta-classes of the previous level into the feature subspace defined by the NDVI. Thus, the "Elevated" meta-class is partitioned into the two leaf nodes Building and Trees

TABLE V
AVERAGE CLASSIFICATION RESULTS (OVER FOUR TRIALS) OBTAINED BY ADDING 45 SAMPLES WHEN THE SOURCE DOMAIN IS: (a) *Pad* , (b) *Vds*. OA%, PA% AND UA% OBTAINED BY APPLYING: 1) THE SUPERVISED SVM CLASSIFIER TRAINED ON THE SOURCE DOMAIN; 2) THE MSCSV STANDARD DA AL METHOD; 3) THE PROPOSED ADAPTATION METHOD SVM_{AL}^{inf} INTEGRATED WITH THE AL STEP FOR ADAPTING THE $\mathcal{C}_v$ CLASSES.

	Classes	Standard				Proposed	
		SVM		SVM_{AL}		SVM_{AL}^{inf}	
		PA%	UA%	PA%	UA%	PA%	UA%
Forest	N. Spruce	30.46	77.45	54.26	75.86	65.11	75.41
	E. Larch	11.72	41.31	36.48	74.25	46.21	79.41
	S. Pine	76.61	39.39	79.03	54.49	80.24	60.06
	E. Beech	61.57	77.53	80.10	78.41	85	82.73
	Hornbeam	55.49	16.49	54.88	27.69	73.78	37.46
No Forest	Grass	88.20	44.35	100	47.91	100	100
	Building	21.13	100	84.51	93.75	97.54	97.19
	Roads	73.63	39.45	95.55	67.72	92.47	99.63
	Soil	0	0	0	0	100	96.41
	OA %	45.97		64.18		77.81	

(a)

	Classes	Standard				Proposed	
		SVM		SVM_{AL}		SVM_{AL}^{inf}	
		PA%	UA%	PA%	UA%	PA%	UA%
Forest	N. Spruce	55.95	9.89	59.52	17.48	63.10	23.45
	E. Larch	22.56	5.59	15.90	11.48	26.15	20.24
	S. Pine	35.89	50.95	51.48	62.99	49.60	70.02
	E. Beech	52.74	78.69	88.36	88.30	94.08	90.22
	Hornbeam	2.08	35.29	43.40	89.29	75.35	91.18
No Forest	Grass	61.63	98.15	96.22	99.70	99.42	97.71
	Building	65.81	37.37	63.24	69.35	99.26	100
	Roads	41.91	21.27	77.21	60.17	100	80.24
	Soil	0	0	90.66	99.45	82.07	100
	OA %	40.80		73.26		81.12	

(b)

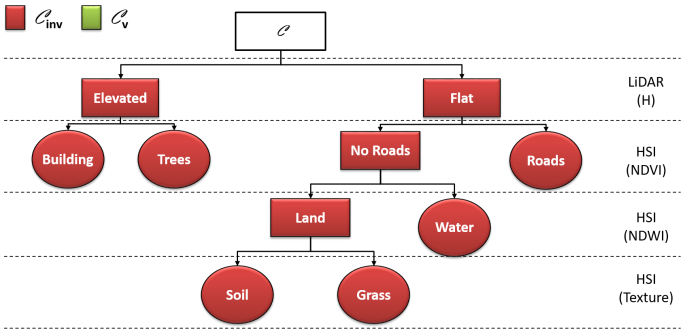


Fig. 11. Hierarchical tree structure derived at the end of the invariance analysis for the Urban Dataset. The invariant feature subspaces used to derive the meta-classes are reported for each level of the hierarchy.

having negative and positive NDVI, respectively. Similarly, the "Flat" is splitted into the Roads leaf node class and "No Roads" (i.e., Soil, Grass and Water) metaclass due to their negative and non-negative NDVI behaviour, respectively. At the third level of the hierarchy, in the feature space defined by the NDWI, we discriminate the Water leaf node class since it is characterized by extremely positive NDWI values compared to the meta-class "Land" (i.e., Soil and Grass). Finally, due to the different textural behaviour the "Land" meta-class is partitioned into Soil and Grass classes.

2) Sensor-driven Inference Method: Tab. VIIa and Tab. VIIb show the classification results obtained when *Plz* and *Hst* are $\mathcal{D}_s$, respectively. The proposed method SVM^{inf} is compared with the SVM classifier trained on $\mathcal{D}_s$, the TCA AND GFK DA method and the semisupervised LapSVM method. Similarly to the Forest Dataset, there is a strong shift in the distributions in the feature subspaces as presented in Fig. 12c and Fig. 12d. Moreover, Tab. VI, where the JM distances between *Hst* and *Plz* are presented, confirms the selected invariant feature spaces for the urban DA problem. The direct use of the standard SVM trained on $\mathcal{D}_s$ and applied to $\mathcal{D}_t$ confirms the shift of the PDFs between the two domains, thus resulting in very low OA (i.e., 57.92% and 59.98%

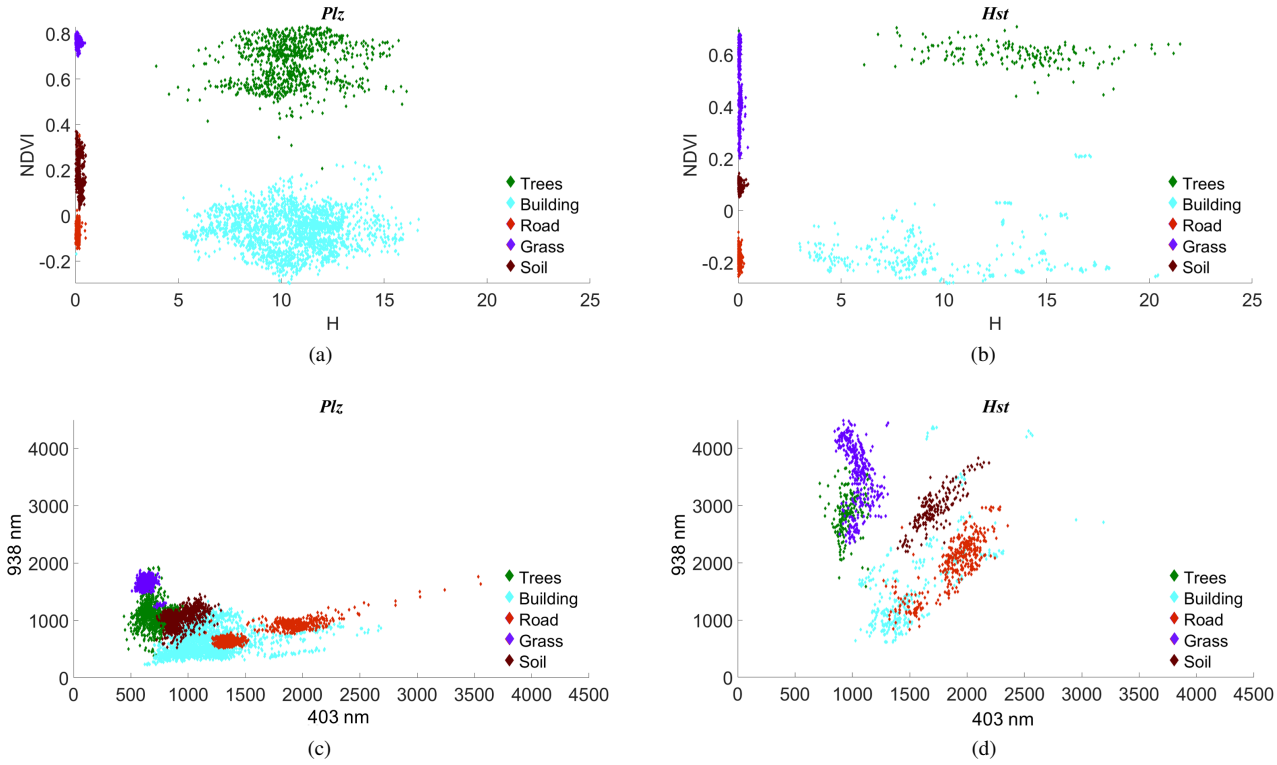


Fig. 12. Distributions of the labeled samples of *Plz* (left) and *Hst* (right) represented in: (a-b) the invariant feature subspace defined by the NDVI (HSI scanner) and the H (LiDAR sensor); and (c-d) the feature subspace defined by spectral channels at the wavelength of 403 nm and 938 nm, respectively.

when *Plz* and *Hst* are $\mathcal{D}_s$, respectively). According to the results obtained for the Forest dataset, the TCA, the GFK and the LapSVM methods improve the classification accuracies of those classes that do not present severe changes in the class distributions. However, for those classes having sharp differences in the PDFs (i.e., Soil and Grass) the effectiveness of the literature approaches decreases. In contrast, the proposed SVM^{inf} is able to achieve a high classification accuracies for all the classes due to the similar behaviour of the distributions in the invariant feature subspaces. Fig. 12a and Fig. 12b show the invariant feature space defined by the NDVI and the H where one can observe that the class distributions are statistically aligned. This is confirmed by the classification results obtained by the proposed SVM^{inf} method, which always improve the results obtained by the literature methods. In particular, OA of 94.22% and 98.26% are achieved when *Plz* and *Hst* are $\mathcal{D}_s$, respectively. Moreover, all the UA and PA accuracies are higher than 84% regardless of the area considered as $\mathcal{D}_s$.

TABLE VI
JM DISTANCE BETWEEN *Plz* AND *Hst* COMPUTED PER CLASS CONSIDERING: (I) THE INVARIANT FEATURE SPACE $\mathbf{x}^{\Psi_n}$, (II) THE FIRST 3 COMPONENTS OF THE PRINCIPAL COMPONENT ANALYSIS (PCA) AND, (III) THE WHOLE FEATURE SPACE $\mathbf{x}$.

$\mathcal{C}_{inv}$	JM distance		
	$\mathbf{x}^{\Psi_n}$	PCA	$\mathbf{x}$
Building	0.063	0.688	1.414
Soil	0.354	0.893	1.414
Grass	0.299	1.402	1.414
Roads	0.612	1.145	1.414
Water	0.410	0.842	1.414
Trees	0.483	1.414	1.414

VI. Discussion and Conclusion

In this paper a sensor-driven hierarchical DA method based on invariant features for the classification of RS images has been presented. The proposed method transfers the knowledge between $\mathcal{D}_s$ and $\mathcal{D}_t$ under the assumptions that: i) they share the same set of land-cover classes, and ii) the same set of multisensor RS data is available for both the domains. In particular, the method takes advantage from a multisensor scenario to mitigate the shift in the PDFs of the classes due to the different acquisition condition. By exploiting the peculiarity of each sensor of measuring different physical properties of the scene, it is possible to detect hierarchically feature subspaces where subsets of classes result statistically aligned. Instead of using data-driven feature extraction methods, the detection of invariant features is driven by the physical properties of the land-cover classes. To this end, a preliminary invariance analysis of the classification problem is required in order to detect the invariant features subspaces for the considered set of land-cover classes.

The proposed approach is based on four main steps. In the first step we perform a sensor-driven hierarchical decompo-

TABLE VII
AVERAGE CLASSIFICATION RESULTS (OVER FOUR TRIALS) WHEN: (a) *Plz* IS THE SOURCE DOMAIN; (b) *Hst* IS THE SOURCE DOMAIN. OA%, PA% AND UA% OBTAINED BY APPLYING: 1) THE SUPERVISED SVM CLASSIFIER TRAINED ON THE SOURCE DOMAIN; 2) THE DA TCA METHOD; 3) THE SEMISUPERVISED LAPSVM; THE DA GFK METHOD; 5) THE PROPOSED SENSOR-DRIVEN DA METHOD SVM^{inf}.

Classes	**Standard**								**Proposed**	
	SVM		**TCA**		**LAPSVM**		**GFK**		**SVM^{inf}**	
	PA%	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**
Building	62.64	79.64	88.20	94.29	85.81	70.80	100	77.67	98.87	100
Soil	43.27	16.21	57.26	35.32	47.58	69.96	41.29	100	99.46	70.34
Grass	85.82	82.12	57.22	98.23	90.98	98.88	98.77	82.47	84.14	98.78
Roads	0	0	81.90	72.97	56.38	71.81	74.42	25	95.70	96.20
Water	87.08	98.75	94.78	86.47	100	64.54	60.87	100	94.78	100
Trees	98.57	52.65	97.73	95.03	97.73	99.14	89.12	97.73	97.72	100
OA %	57.92		77.84		78.80		73.71		94.22	

(a)

Classes	**Standard**								**Proposed**	
	SVM		**TCA**		**LAPSVM**		**GFK**		**SVM^{inf}**	
	PA%	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**	**PA%**	**UA%**
Building	97.07	41.17	97.70	97.70	98.31	98.01	84.98	91.31	99.97	99.84
Soil	19.59	87.76	2.67	70.77	85.17	83.38	70.31	59.88	90.34	100
Grass	88.81	98.07	100	63.20	100	82.72	94.38	100	100	100
Roads	0	0	97.13	57.85	12.65	98.82	64.30	68.35	96.01	91.94
Water	72.59	92.41	79.24	99.96	99.94	66.88	95.22	97.08	100	97.67
Trees	24.93	100	93.48	100	78.17	100	100	79.09	99.63	100
OA %	59.98		81.97		83.37		85.53		98.26	

(b)

sition by focusing the attention on the physical properties of the scene. Thus, the adaptation problem can be decomposed by dividing the classes into two groups: i) a subset of classes for which invariant features are measured, ii) a subset of classes for which none of the sensors available can measure invariant properties. In the second step, we bridge the gap between the distributions according to a sensor-driven DA strategy, which infers labels from $\mathcal{D}_s$ to $\mathcal{D}_t$ for the invariant classes. By taking advantage from the statistical alignment of the domain distributions in the detected features subspaces, we generate an initial training set for $\mathcal{D}_t$ selecting the target samples more aligned to the source data. This training set is analyzed in the full feature space to properly model the PDF of the target domain for the considered subset of classes. Indeed, even though the invariant feature subspaces present a reliable distribution alignment between the domains, they may be not sufficient to accurately solve the complex classification problem of the target domain. The obtained training set is employed to perform in the third step a supervised classification of $\mathcal{D}_t$ for the set of invariant classes. Finally, the proposed method addresses the adaptation of the remaining land-cover classes characterized by variant features behaviour using an AL technique.

Experimental results obtained on two real multisensor dataset show that: i) by means of the invariant feature subspace we correctly transfer the knowledge of $\mathcal{D}_s$ to the more aligned unlabeled samples of $\mathcal{D}_t$, ii) by considering the full feature space we are able to accurately model the PDFs of $\mathcal{D}_t$ for the invariant classes, and iii) reliable classification result can be obtained for the set of invariant classes of $\mathcal{D}_t$ without requiring any labeled sample for it. Moreover, the adaptation of the invariant classes introduces constraints to the general structure of the entire problem facilitating the adaptation of the remaining variant classes. Due to the adaptation performed by means of the sensor-driven inference method we are in the condition of focusing the attention only on the $\mathcal{C}_v$ classes, thus reducing the number of needed target samples to achieve accurate classification maps. Indeed, a standard DA-AL technique selects the most informative samples for all the land-cover classes, whereas the proposed method requires only the labels of samples belonging to the $\mathcal{C}_v$ classes. This condition allows us to rapidly increase both the accuracy on the $\mathcal{C}_v$ classes and on the entire DA problem.

As a final remark, we point out that due to the increasing availability of multisensor data, the proposed sensor-driven DA method is promising from the operational view point. Indeed, the possibility of acquiring complementary measures of the scene makes it possible to take advantage from the physical properties of the classes to drive the adaptation, thus increasing the reliability of the adaptation process. Moreover, differently

form the statistical DA method present in the literature, in this paper we propose a simple and effective approach based on the physical meaning of the measures collected by different sensors, thus performing an accurate adaptation which does not require the tuning of critical parameters.

As future work, we plan to further analyze the sensor-driven inference strategy by testing the proposed technique on datasets where a different combination of sensors is available. Moreover, we aim to replace the AL technique with an unsupervised DA method in order to adapt the set of variant classes without any labeling cost. This can be achieved by exploiting the constraints introduced on the set of invariant classes to drive the unsupervised adaptation of the remaining land-cover classes.

REFERENCES

[1] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 770 –787, May 2010.

[2] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] John R Schott, Carl Salvaggio, and William J Volchok, "Radiometric scene normalization using pseudoinvariant features," *Remote Sensing of Environment*, vol. 26, no. 1, pp. 1–16, 1988.

[4] Curtis E Woodcock, Scott A Macomber, Mary Pax-Lenney, and Warren B Cohen, "Monitoring large areas for forest change using landsat: Generalization across space, time and landsat sensors," *Remote Sensing of Environment*, vol. 78, no. 1, pp. 194–203, 2001.

[5] Mary Pax-Lenney, Curtis E Woodcock, Scott A Macomber, Sucharita Gopal, and Conghe Song, "Forest mapping with a generalized classifier and landsat tm data," *Remote Sensing of Environment*, vol. 77, no. 3, pp. 241–250, 2001.

[6] Shilpa Inamdar, Francesca Bovolo, Lorenzo Bruzzone, and Subhasis Chaudhuri, "Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 4, pp. 1243–1252, 2008.

[7] Zhengwei Yang, Wei Zhang, Wei Wang, and Qiongcheng Xu, "Change detection based on iterative invariant area histogram matching," in *Geoinformatics, 2011 19th International Conference on*. IEEE, 2011, pp. 1–6.

[8] Joon Heo and Thomas W FitzHugh, "A standardized radiometric normalization method for change detection using remotely sensed imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 66, no. 2, pp. 173–181, 2000.

[9] Lorenzo Bruzzone and Diego Fernàndez Prieto, "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 2, pp. 456–460, 2001.

[10] Lorenzo Bruzzone and Roberto Cossu, "A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 9, pp. 1984–1996, 2002.

[11] Begüm Demir, Francesca Bovolo, and Lorenzo Bruzzone, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 1, pp. 300–312, 2013.

[12] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.

[13] Kai Zhang, Vincent Zheng, Qiaojun Wang, James Kwok, Qiang Yang, and Ivan Marsic, "Covariate shift in hilbert space: A solution via sorrogate kernels," in *International Conference on Machine Learning*, 2013, pp. 388–395.

[14] Claudio Persello, "Interactive domain adaptation for the classification of remote sensing images using active learning," *Geoscience and Remote Sensing Letters, IEEE*, vol. 10, no. 4, pp. 736–740, 2013.

[15] Giona Matasci, Devis Tuia, and Mikhail Kanevski, "Svm-based boosting of active learning strategies for efficient domain adaptation," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 5, pp. 1335–1343, 2012.

[16] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 3, pp. 1014–1031, 2011.

[17] Sumit Chopra, Suhrid Balakrishnan, and Gopalan Raghuraman, "Dlid: Deep learning for domain adaptation by interpolating between domains," in *ICML Workshop on Challenges in Representation Learning*, 2013.

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[19] Lorenzo Bruzzone, Mingmin Chi, and Mattia Marconcini, "A novel transductive svm for semisupervised classification of remote-sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 11, pp. 3363–3373, 2006.

[20] Steffen Bickel, Michael Brückner, and Tobias Scheffer, "Discriminative learning for differing training and test distributions," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 81–88.

[21] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, 2006, pp. 601–608.

[22] Goo Jun and Joydeep Ghosh, "Spatially adaptive classification of land cover with remote sensing data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 7, pp. 2662–2673, 2011.

[23] Anshu Singla, Swarnajyoti Patra, and Lorenzo Bruzzone, "A novel classification technique based on progressive transductive svm learning," *Pattern Recognition Letters*, vol. 42, pp. 101–106, 2014.

[24] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko, "Efficient learning of domain-invariant image representations," *arXiv preprint arXiv:1301.3224*, 2013.

[25] I-Hong Jhuo, Dong Liu, DT Lee, and Shih-Fu Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2168–2175.

[26] Brian Kulis, Kate Saenko, and Trevor Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1785–1792.

[27] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," *Computer Vision–ECCV 2010*, pp. 213–226, 2010.

[28] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 10, pp. 3044–3054, 2007.

[29] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[30] Luis Gómez-Chova, Gustavo Camps-Valls, Jordi Munoz-Mari, and Javier Calpe, "Semisupervised image classification with laplacian support vector machines," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 3, pp. 336–340, 2008.

[31] Tiberio S Caetano, Terry Caelli, Dale Schuurmans, and Dante Augusto Couto Barone, "Graphical models and point pattern matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1646–1663, 2006.

[32] Tibério S Caetano, Julian J McAuley, Li Cheng, Quoc V Le, and Alex J Smola, "Learning graph matching," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 6, pp. 1048–1058, 2009.

[33] Bin Luo and Edwin R. Hancock, "Structural graph matching using the em algorithm and singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1120–1136, 2001.

[34] Biplab Banerjee, Francesca Bovolo, Avik Bhattacharya, Lorenzo Bruzzone, Subhasis Chaudhuri, and Krishna Mohan Buddhiraju, "A novel graph-matching-based approach for domain adaptation in classification of remote sensing image pair," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 7, pp. 4045–4062, 2015.

[35] Vishal M Patel, Raghuraman Gopalan, and Ruonan Li, "1visual domain adaptation: An overview of recent advances," .

[36] Devis Tuia and Gustau Camps-Valls, "Kernel manifold alignment for domain adaptation," *PloS one*, vol. 11, no. 2, pp. e0148655, 2016.

[37] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.

[38] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 769–776.

[39] Si Si, Dacheng Tao, and Bo Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 929–942, 2010.

[40] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1134–1148, 2014.

[41] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2066–2073.

[42] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 999–1006.

[43] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2288–2302, 2014.

[44] Hal Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[45] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[46] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 9, pp. 3180–3191, 2009.

[47] Devis Tuia, Jordi Munoz-Mari, Luis Gomez-Chova, and Jesus Malo, "Graph matching for adaptation in remote sensing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 1, pp. 329–341, 2013.

[48] Chang Wang and Sridhar Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, vol. 22, p. 1541.

[49] Giona Matasci, Michele Volpi, Mikhail Kanevski, Lorenzo Bruzzone, and Devis Tuia, "Semisupervised transfer component analysis for domain adaptation in remote sensing image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 7, pp. 3550–3564, 2015.

[50] Jihun Ham, Daniel Lee, and Lawrence Saul, "Semisupervised alignment of manifolds," in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, 2005, vol. 10, pp. 120–127.

[51] Devis Tuia, Michele Volpi, Maxime Trolliet, and Gustau Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, 2014.

[52] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál, "Impossibility theorems for domain adaptation.," in *AISTATS*, 2010, pp. 129–136.

[53] Shai Ben-David and Ruth Urner, "On the hardness of domain adaptation and the utility of unlabeled target samples," in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 139–153.

[54] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, June 2016.

[55] Jerónimo Arenas-García, K Petersen, Gustavo Camps-Valls, and Lars Kai Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 16–29, 2013.

[56] Matthew Brand, "Charting a manifold," in *Advances in neural information processing systems*, 2002, pp. 961–968.

[57] Yee W Teh and Sam T Roweis, "Automatic alignment of local representations," in *Advances in neural information processing systems*, 2002, pp. 841–848.

[58] Hsiuhan Lexie Yang and Melba M Crawford, "Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 54, no. 1, pp. 51–64, 2016.

[59] Lei Zhu and Li Ma, "Class centroid alignment based domain adaptation for classification of remote sensing images," *Pattern Recognition Letters*, 2016.

[60] Wonkook Kim and Melba M Crawford, "Adaptive classification for hyperspectral image data using manifold regularization kernel machines," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 11, pp. 4110–4121, 2010.

[61] Begum Demir, Francesca Bovolo, and Lorenzo Bruzzone, "Classification of time series of multispectral images with limited training data," *Image Processing, IEEE Transactions on*, vol. 22, no. 8, pp. 3219–3233, 2013.

[62] D Tuia, E Pasolli, and WJ Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2232–2242, 2011.

[63] Xin Huang, Qikai Lu, and Liangpei Zhang, "A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 90, pp. 36–48, 2014.

[64] Xin Huang, Chunlei Weng, Qikai Lu, Tiantian Feng, and Liangpei Zhang, "Automatic labelling and selection of training samples for high-resolution remote sensing image classification over urban areas," *Remote Sensing*, vol. 7, no. 12, pp. 16024–16044, 2015.

[65] Lorenzo Bruzzone, Mattia Marconcini, U Wegmuller, and Andreas Wiesmann, "An advanced system for the automatic classification of multitemporal sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1321–1334, 2004.

[66] JA Gualtieri, Samir R Chettri, RF Cromp, and LF Johnson, "Support vector machine classifiers as applied to aviris data," in *Proc. Eighth JPL Airborne Geoscience Workshop*. Citeseer, 1999.

[67] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 8, pp. 1778 – 1790, Aug. 2004.

[68] Mingmin Chi and Lorenzo Bruzzone, "A semilabeled-sample-driven bagging technique for ill-posed classification problems," *Geoscience and Remote Sensing Letters, IEEE*, vol. 2, no. 1, pp. 69–73, 2005.

[69] Gustavo Camps-Valls and Lorenzo Bruzzone, "Kernel-based methods for hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 6, pp. 1351–1362, 2005.

[70] Swarnajyoti Patra and Lorenzo Bruzzone, "A fast cluster-assumption based active-learning technique for classification of remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 5, pp. 1617–1626, 2011.

[71] Suju Rajan, Joydeep Ghosh, and Melba M Crawford, "An active learning approach to hyperspectral data classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 4, pp. 1231–1242, 2008.

[72] Claudio Persello and Lorenzo Bruzzone, "A novel active learning strategy for domain adaptation in the classification of remote sensing images," in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*. IEEE, 2011, pp. 3720–3723.

[73] Devis Tuia, Frédéric Ratle, Fabio Pacifici, Mikhail F Kanevski, and William J Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[74] Pabitra Mitra, B Uma Shankar, and Sankar K Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern recognition letters*, vol. 25, no. 9, pp. 1067–1074, 2004.

[75] Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, et al., "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.

[76] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.