Semantics-Aware Hierarchical Consensus Learning for Remote Sensing Image Classification

Giulio Weikmann, *Member, IEEE*, Gianmarco Perantoni, *Member, IEEE*, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Deep learning (DL) has become increasingly important in remote sensing (RS) image classification due to its ability to extract semantic information from complex data. Classification tasks often include predefined label hierarchies that represent the semantic relationships among classes. However, these hierarchies are frequently overlooked, and most approaches focus only on fine-grained classification schemes. In this article, we present a novel Semantics-Aware Hierarchical Consensus (SAHC) approach that integrates hierarchical-level-specific classification heads within a deep network architecture and combines their output through cross-level probability projectors. Direct and projected predictions are fused into a geometric consensus distribution, which is used for self-consistent training and optional hierarchy-aware inference. This mechanism acts as a geometric ensemble that leverages the inherent structure of the hierarchical classification task. The projectors are initialized from the *user-defined* taxonomy (i.e., the hierarchical label structure) and can be adaptively refined during optimization. The proposed SAHC method is evaluated on two benchmark datasets with different degrees of hierarchical complexity on different tasks, considering varying spectral and spatial resolutions. Experimental results show both the effectiveness of the proposed approach in guiding network learning and the robustness of the hierarchical consensus (HC) for RS image classification tasks. The source code is available at: https://github.com/rslab-unitrento/sahc**

***Index Terms*—Consensus, consistency, deep learning (DL), hierarchical, multigranularity, multispectral (MS), remote sensing (RS), semantics-aware, Sentinel-2 (S2), time series (TS), very-high resolution (VHR).**

I. Introduction

IMAGE classification and semantic segmentation are central tasks in remote sensing (RS) image processing. In recent years, the proliferation of RS data and open distribution policies have significantly increased the availability of land-use and land-cover (LULC) products, along with their corresponding classification schemes [1]. These products play a critical role in various applications and are widely used by government agencies, municipalities, and ministries. However, global and regional LULC maps often differ in terms of resolution and are affected by semantic inconsistencies in class definitions and production algorithms [2]. While deep learning (DL) methods have revolutionized LULC mapping by capturing complex spatial and spectral dependencies, standard approaches typically treat classification as a flat, single-level task. In practice, land-cover (LC) classes are naturally organized into hierarchical label structures that reflect intrinsic levels of semantic granularity. This multilevel hierarchy is a property of the physical landscape. For instance, within a single dataset, a specific target might be unambiguously identified at a coarse-grained level (e.g., forest), while its finer grained classification (e.g., deciduous versus coniferous) remains uncertain due to inherent label ambiguity, complex scene compositions, or the physical limits of the sensor. Neglecting this inherent hierarchical structure discards valuable semantic information and can lead to suboptimal or inconsistent predictions, particularly when dealing with label ambiguity or complex RS scenes. An example of this hierarchical problem definition can be seen in Fig. 1.

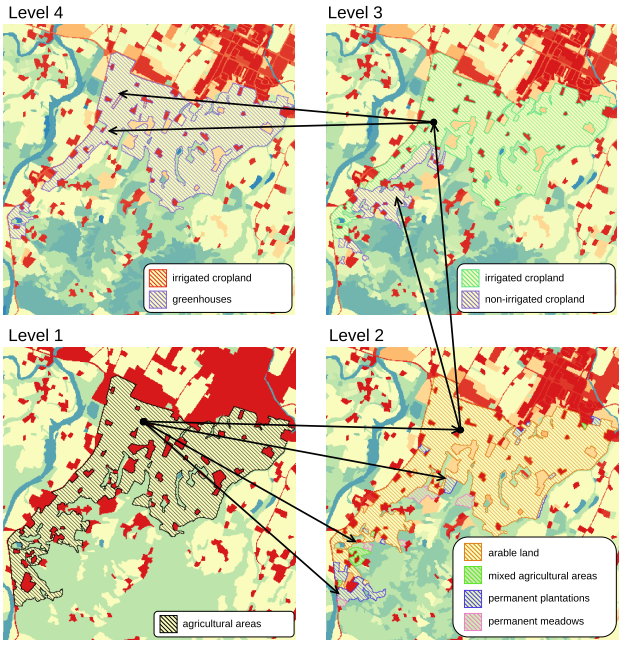


Fig. 1. Hierarchical classification task example. Level 1 represents the coarsest class definition, where each class is progressively broken down into its fine-grained representations.

Received 9 April 2026; revised 14 July 2026 and 21 August 2026; accepted 2 September 2026. Date of publication 7 September 2026; date of current version 22 September 2026. This work was supported in part by European Space Agency (ESA) through the project "CCI High Resolution Land Cover ECV" under Contract 4000125259718/I-NB and in part by Italian Space Agency (ASI) and the Ministry of University and Research (MUR) through the Project "Space It Up!" under Contract 2024-5-E.0—CUP n. I53D24000060005. *(Corresponding author: Lorenzo Bruzzone.)*

The authors are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: giulio.weikmann@unitn.it; gianmarco.perantoni@unitn.it; lorenzo.bruzzone@unitn.it).

Digital Object Identifier 10.1109/TGRS.2026.3731443

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

In the literature, most of the works related to RS image classification focus on single-level approaches, discarding the hierarchical information contained in the data. Hierarchical classification schemes, such as the CORINE Land Cover (CLC) maps, have mostly been adopted in object-based image classification methods based on very-high resolution (VHR) satellite images [3], often considering a limited number of classes to simplify the classification task. To validate the hierarchical structure, few works analyzed the interlevel relationships, considering decision trees in object-based classification scenarios [4], or validation approaches based on intercomparison of different maps [5]. Instead, the wide majority of works in the RS domain focus on single-level approaches, discarding the hierarchical information contained in the data. While these methods show promising results at individual hierarchical levels, they fail to incorporate the hierarchical structure of the data into the decision-making process, overlooking important information included in the source data.

Only recently, hierarchical and multigranularity classification (MGC) methods have emerged to address these challenges in RS. These methods leverage the hierarchical structure to encode semantic relationships among labels, enabling knowledge transfer across different levels of granularity. Recent studies have demonstrated the benefits of adaptive label granularity selection (ALGS) to balance classification accuracy and specificity [6], as well as the integration of parent–child relationships and sibling mutual exclusivity to improve hierarchical representation learning [7]. Despite these advancements, integrating hierarchical information effectively into the learning process remains an open challenge. Many existing methods rely on complex, task-specific architectures or rigid rule-based postprocessing steps, which limit their flexibility across different DL backbones and fail to explicitly enforce probability consistency across different hierarchical levels in a generalized manner.

To address these limitations, we propose the novel and modular Semantics-Aware Hierarchical Consensus (SAHC) method. SAHC integrates hierarchical-level-specific classification heads within standard DL architectures, projects multilevel predictions to a common hierarchical level, and combines them through a consensus distribution. The corresponding projectors can be derived from the *user-defined* taxonomy (i.e., the hierarchical label structure), while trainable hierarchy projectors provide an optional adaptive refinement around this semantic prior.

The key contributions of this work can be summarized as follows.

1) We introduce the SAHC framework that treats the predictions produced at different hierarchical levels as a committee of probabilistic estimates. The predictions are projected to a common target level and fused into a consensus distribution.
2) We propose a self-consistent learning strategy that aligns direct and projected predictions with the consensus during training. The same mechanism supports hierarchy-aware consensus prediction at inference, while direct-head predictions remain available when fine-level discrimination is the primary objective.
3) We introduce taxonomy-initialized trainable projection matrices as an adaptive parameterization of the cross-level projectors. This parameterization enables *data-driven* calibration of cross-level compatibility while preserving the *user-defined* taxonomy as a semantic prior.
4) We demonstrate the effectiveness and robustness of the proposed approach on two RS datasets with different hierarchy complexities, spatial resolutions, and classification settings.

The remainder of this article is organized as follows. Section II reviews related works on hierarchical semantic segmentation. Section III details the proposed hierarchical consensus (HC) approach. The datasets and experimental setup are described in Section IV, while Section V presents and discusses the experimental results. Finally, Section VI concludes this article and outlines future research directions.

II. Related Works

Image classification and semantic segmentation are widely studied tasks in RS image processing, with DL models achieving state-of-the-art performance across various domains. Standard approaches typically treat classification as a "flat," single-granularity problem, predicting probabilities over a set of mutually exclusive classes without explicitly considering potential semantic relationships or hierarchical structures among them. However, many real-world RS datasets possess inherent taxonomies where classes can be naturally organized into hierarchies (e.g., standardized LC systems and target recognition taxonomies with super-categories and subcategories). Ignoring this structure can lead to suboptimal performance, as standard loss functions treat all misclassifications equally. Without an explicit hierarchical representation, models fail to differentiate between errors involving closely related classes and those involving semantically distant ones.

Leveraging hierarchical label structures during model training and inference has emerged as a promising direction to improve accuracy and semantic understanding. Early work demonstrated that standard convolutional neural networks (CNNs) implicitly capture some hierarchical information, with shallower layers grouping high-level categories and deeper layers specializing in finer distinctions [8]. Building on this, researchers have explored explicit methods to incorporate hierarchy. Typically, this is achieved by mapping the label hierarchy to specific network architectures or by introducing loss functions that handle the correlation between different granularity tasks [9], [10], [11]. For instance, several methods propose levelwise hierarchical cross-entropy losses based on conditional probabilities or introduce custom features to minimize the divergence between aggregated predictions at different hierarchical levels, aiming to reduce the severity of misclassification. Bertinetto et al. [12] proposed a levelwise hierarchical cross-entropy loss based on conditional probabilities within the hierarchy and used mapping functions to embed class relationships. Similarly, Garg et al. [13] introduced the Hierarchy-Aware Features (HAF), incorporating multiple losses, including minimizing the Jensen–Shannon divergence (JSD) between aggregated predictions at different hierarchical levels and enforcing geometric consistency in the feature space, aiming to reduce the severity of misclassifications. In [14], the authors proposed hierarchy and exclusion graphs to model semantic relations between two labels assigned to

the same objects. They considered exclusions, overlaps, and subsumptions, and provided a detailed theoretical analysis of the structure of the hierarchical problem. Building on this work, a combinatorial loss is proposed by Chen et al. [15], which addresses the scenario where samples might be labeled at different levels of granularity. The loss is designed to maximize the marginal probability of the observed ground-truth label by aggregating information from related labels in the hierarchy. The authors introduced an hierarchical residual network (HRN) architecture, where features from parent levels act as residual connections to inform child-level features, explicitly facilitating hierarchical feature interaction. Following this work, Zhou et al. [16] proposed a multi-granularity archeological dating method that depends on an explicit knowledge-guided relation graph, employing a focal-type probability classification loss to mine the relationship between specific object attributes and their broader eras. Their network leverages a bidirectional multigranularity module (MGM) with gradient-truncated addition to explicitly fuse coarse- and fine-grained features. Similarly, Cai et al. [17] introduced HierSwin, a hierarchical Vision Transformer, which explicitly models the semantic correlations embedded within a predefined three-level hierarchical tree to gradually improve accuracy from fine- to coarse-grained classifications.

More recently, some works addressed the MGC in RS to handle targets that are discernible only at specific variable semantic levels due to differences in spatial resolution, sensor modalities, and data quality. Chen et al. [6] introduced an accuracy-specificity curve to adaptively select task-dependent prediction thresholds. Their approach also employs a penalty term based on the global semantic distance across the label tree to adjust class logits. Addressing a similar challenge in ship classification, Chen et al. [7] proposed a hierarchical contrastive learning (HCL) framework. By explicitly incorporating parent–child dependencies and mutual exclusivity among sibling categories, their method aligns class-specific representations with semantic relationships and enforces probability coherence across levels through a consistency loss.

Despite these significant advancements, many existing multigranularity and hierarchical learning methods still rely on the assumption that a predefined and accurate hierarchy is available, where no hierarchy violation is possible. This requires rigid postprocessing rules, fixed decision thresholds, or highly specific nonmodular architectures. Furthermore, many existing techniques require explicit guidance or architectural modifications tailored to a specific fixed hierarchy [18], potentially limiting their flexibility and ability to adapt if the underlying hierarchical understanding evolves or is uncertain. For example, some approaches decompose the problem into sequential decisions or binary classifications guided by the known structure [19], [20], rather than allowing the model to learn or refine hierarchical relationships implicitly from the data.

The challenge remains to exploit predefined hierarchical semantics through a modular probability-driven mechanism, while allowing limited *data-driven* calibration of cross-level relationships around the taxonomy prior.

III. Proposed Methodology

In this work, we address the above-mentioned challenges by proposing a novel DL framework for hierarchy-aware image classification. Our approach models the hierarchical semantic relationships within image data by integrating hierarchical loss components with a robust consensus mechanism.

The main methodological component of SAHC is an all-level HC approach. For each target hierarchical level, the predictions produced by all hierarchical levels are projected into the corresponding target label space through cross-level projectors. These projectors can be directly derived from the *user-defined* taxonomy and, in the adaptive implementation, are initialized from this taxonomy and jointly optimized with the network. Then, the direct prediction of the target-level head and the projected predictions from all remaining levels are fused into a consensus probability distribution. SAHC jointly learns level-specific classifiers and taxonomy-anchored probabilistic mappings that align their predictions across the hierarchy.

The proposed SAHC differs from other consistency-based approaches in the way hierarchical agreement is modeled. Previous methods generally enforce local parent–child coherence through fixed child-to-parent probability aggregation, pairwise consistency constraints, or semantic-distance-based penalties. In contrast, SAHC projects the predictions produced at all hierarchical levels into a common target label space and jointly combines them with the direct target-level prediction through a consensus distribution. This consensus is used both as a regularization target during training and as an optional hierarchy-aware prediction at inference.

A. Hierarchical Problem Formulation

Let us denote column vectors by lowercase bold letters (e.g., $\mathbf{v}$) and matrices by uppercase bold letters (e.g., $\mathbf{B}$). The ith component of a vector $\mathbf{v}$ is denoted by v_i, while the ith row and jth column of a matrix $\mathbf{B}$ are denoted by $\mathbf{B}_{i,\cdot}$ and $\mathbf{B}_{\cdot,j}$, respectively. Then, let us consider a user-defined hierarchy label tree with H levels, where the leaves represent the finest level of detail. We denote each hierarchical level as $h \in \mathcal{H} = [1, 2, \ldots, H]$, ranging from the coarsest level $h = 1$ to the finest one $h = H$. The root node $h = 0$ is not considered, as it represents the trivial case in which all classes are mapped into a single class. Let $\mathbf{x} \in \mathbb{R}^b$ be the input vector, where b is the number of input features. Its associated labels are $y^h \in \Omega^h$, where $\Omega^h = \{\omega_1^h, \ldots, \omega_{|\Omega^h|}^h\}$ is the set of labels at hierarchical level h, with $|\Omega^h|$ number of labels. For every $h_1 < h_2$, we have that each fine-grained class $\omega_j^{h_2}$ is mapped to a coarse-grained class $\omega_i^{h_1}$, and $|\Omega^{h_1}| \leq |\Omega^{h_2}|$. We define this mapping relationship by $\omega_j^{h_2} \implies \omega_i^{h_1}$. Let $\mathcal{D} = \{(\mathbf{x}_n, Y_n) | n = 1, \ldots, N\}$ be the training set, where for each input feature vector $\mathbf{x}_n$, we have an associated set $Y_n = \{y_n^1, \ldots, y_n^h, \ldots, y_n^H\}$, which denotes the ground-truth labels at all H hierarchical levels, for a total number of N training samples. The labels in the dataset satisfy a nested label structure, where each class at a finer level is contained within a class at a coarser level.

B. Fully Supervised Multilevel Learning

The proposed methodology can be applied to any backbone network by simply introducing the required additional hierarchical classification modules. We refer to each of the hierarchical classification modules $\mathbf{g}^h : \mathbb{R}^d \to \mathbb{R}^{|\Omega^h|}$ on top of a given backbone network as the composition $\mathbf{g}^h \circ \mathbf{f} : \mathbb{R}^b \to \mathbb{R}^{|\Omega^h|}$,

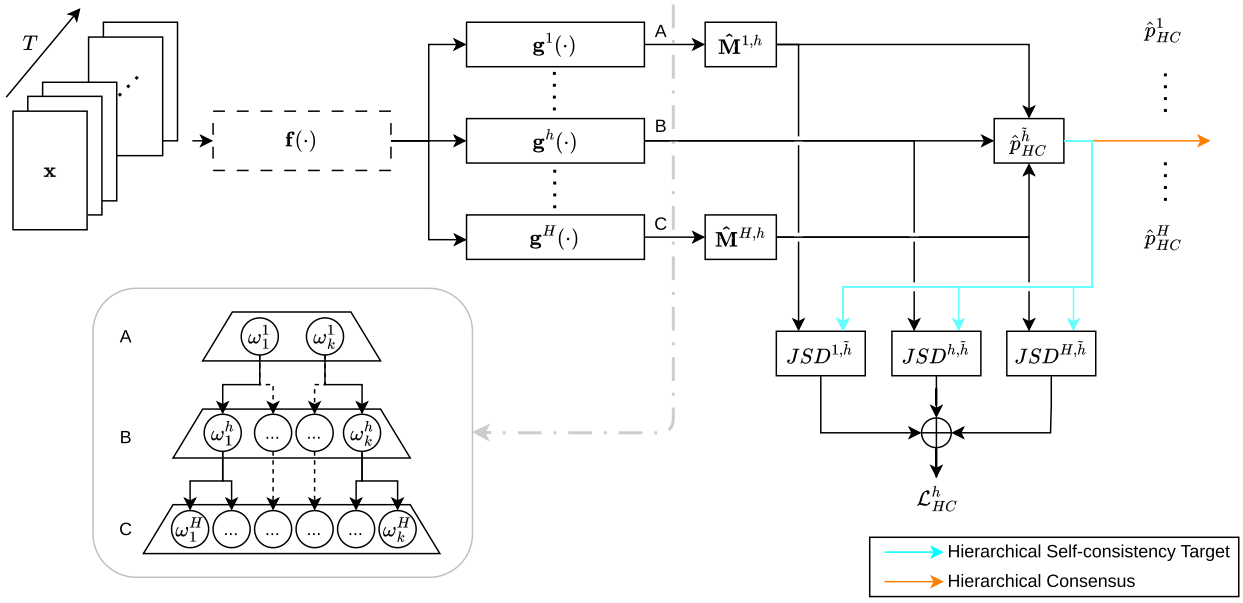


Fig. 2. Illustration of the proposed hierarchical architecture with the HC calculated on $h = \tilde{h}$. Heads tailored to different granularity levels are reprojected on the target level and geometrically averaged to extract the consensus probability $\hat{p}^{\tilde{h}}_{HC}$. This consensus probability is used to calculate the hierarchical self-consistency loss $\mathcal{L}^h_{HC}$. Optionally, this consensus probability can be used to generate a robust hierarchy-aware prediction during inference.

where $\mathbf{f} : \mathbb{R}^b \to \mathbb{R}^d$ is the backbone feature extractor. Fig. 2 provides a high-level overview of the considered multilevel classification approach.

Each classifier $\mathbf{g}^h$ estimates class-posterior logits. Let $\hat{\mathbf{p}}^h(\mathbf{x}) \in \mathbb{R}^{|\Omega^h|}$ be the vector of estimated class-posterior probabilities at level h given $\mathbf{x}$, i.e., $\hat{\mathbf{p}}^h(\mathbf{x}) = [\hat{P}(\omega^h_1|\mathbf{x}), \ldots, \hat{P}(\omega^h_{|\Omega^h|}|\mathbf{x})]^\mathsf{T}$, where T refers to the transpose operator. Then, the class-posterior probabilities $\hat{\mathbf{p}}^h$ for the input sample $\mathbf{x}$ at level h can be retrieved as follows:

$$\hat{\mathbf{p}}^h(\mathbf{x}) = \boldsymbol{\sigma}\left[\mathbf{g}^h(\mathbf{f}(\mathbf{x}))\right] \tag{1}$$

where $\boldsymbol{\sigma} : \mathbb{R}^d \to [0,1]^d$ is the *softmax* operator, with $\sigma_i[\cdot]$ computed as follows:

$$\sigma_i\left[\mathbf{g}\right] = \frac{\exp g_i}{\sum_j \exp g_j} \in [0,1]. \tag{2}$$

The training of $\mathbf{g}^h$ is done by computing the categorical cross-entropy (CCE) loss as follows:

$$\begin{aligned}\mathcal{L}^h_{CCE} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|\Omega^h|}\mathbb{1}\left[y^h_n = \omega^h_i\right]\log\hat{P}^h\left(\omega^h_i|\mathbf{x}_n\right)\\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|\Omega^h|}\mathbb{1}\left[y^h_n = \omega^h_i\right]\log\sigma_i\left[\mathbf{g}^h(\mathbf{f}(\mathbf{x}_n))\right]\end{aligned} \tag{3}$$

where $\mathbb{1}[\cdot]$ serves as an indicator function and takes a value of 1 when the argument is true, else 0. We apply the CCE loss function to all hierarchical levels, obtaining the total CCE loss $\mathcal{L}_{CCE}$ as the sum of the levelwise losses

$$\mathcal{L}_{CCE} = \sum_{h=1}^{H}\lambda^h\mathcal{L}^h_{CCE}. \tag{4}$$

As it has also been noted in previous works [21], adopting several CCE loss functions for diverse classification levels may negatively affect performance at the most granular classification level. To mitigate this behavior, we introduced the weighting terms $\lambda = \{\lambda^1, \ldots, \lambda^H\}$ to limit the impact of coarser levels, while still accounting for the additional losses.

C. Cross-Level Projectors

SAHC requires cross-level projection operators to map predictions between hierarchical levels before consensus formation. These projectors can be directly derived from the *user-defined* taxonomy or adaptively parameterized. The implementation considered in this work uses trainable, taxonomy-initialized matrices to refine the projectors during optimization.

The CCE loss functions at each hierarchical level lead to independent classifications, which are not constrained to follow the hierarchy definition formulated in the problem statement. To promote cross-level agreement, additional consensus predictions and self-consistent penalty terms are introduced. These penalty terms are defined through hierarchy projection matrices designed to map the taxonomy-specified relationships between the different layers.

Let us consider for simplicity the case of $H = 2$ levels, with h_1 representing the coarse-grained labels and h_2 the fine-grained labels, i.e., $h_1 < h_2$. We refer to Ω^{h_1} as the set of the *macroclasses* (i.e., the coarse-grained classes) and to Ω^{h_2} as the set of the *microclasses* (i.e., the fine-grained classes). We can define the hierarchy indicator matrix $\mathbf{I}^{h_2,h_1} \in \{0,1\}^{|\Omega^{h_2}|\times|\Omega^{h_1}|}$, which represents the *user-defined* hierarchical relationship between level h_2 and h_1

$$\mathbf{I}^{h_2,h_1} = \begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1,|\Omega^1|} \\ I_{2,1} & I_{2,2} & \cdots & I_{2,|\Omega^1|} \\ \vdots & \vdots & \ddots & \vdots \\ I_{|\Omega^2|,1} & I_{|\Omega^2|,2} & \cdots & I_{|\Omega^2|,|\Omega^1|} \end{bmatrix}. \tag{5}$$

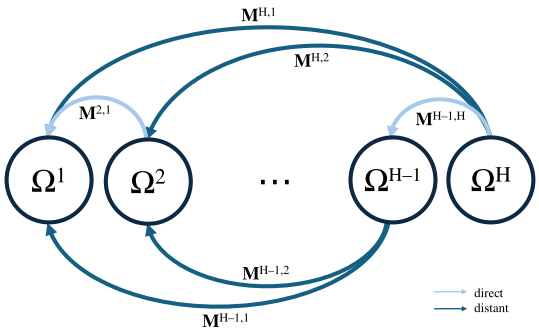


Fig. 3. Illustration of both direct and distant hierarchical mappings, from fine-grained classes to coarse-grained classes.

This approach can be easily extended to a multilevel scenario, considering any given number H of hierarchical levels. In this case, the total number of hierarchy matrices that can be determined is $((H-1)\cdot H)/2$, as also mappings of distant relationship levels can be defined (e.g., from the root of the label tree to the leaves, skipping intermediate levels) if needed. A generalized example is illustrated in Fig. 3.

In a strict hard-hierarchy, $I_{i,j}^{h_2,h_1}$ assumes a value of 1 when the microclass $\omega_i^{h_2}$ is contained in the macroclass $\omega_j^{h_1}$ through the relationship $\omega_i^{h_2} \implies \omega_j^{h_1}$, and 0 otherwise. However, this rigid formulation often fails to account for semantic overlaps in real-world hierarchies. Therefore, we distinguish between *user-defined* aggregations, established a priori for a given task, and *data-driven* adaptive projection weights, which calibrate dataset-specific associations and model uncertainties around the taxonomy prior. To combine the *user-defined* prior with a *data-driven* optimization, we parameterize the hierarchical relationships in the unconstrained logit space. Instead of strictly fixing the matrices to binary values or directly bounding them in $[0,1]$, we define learnable parameter matrices $\mathbf{W}^{h_2,h_1} \in \mathbb{R}^{|\Omega^{h_2}|\times|\Omega^{h_1}|}$ used to model the log joint distribution matrices $\log \mathbf{J}^{h_2,h_1} \in \mathbb{R}^{|\Omega^{h_2}|\times|\Omega^{h_1}|}$. $\mathbf{W}^{h_2,h_1}$ are initialized based on the *user-defined* indicator matrix $\mathbf{I}^{h_2,h_1}$ to provide a stronger prior

$$W_{i,j}^{h_2,h_1} = \begin{cases} \epsilon, & \text{if } I_{i,j}^{h2,h_1} = 1 \\ -\Delta + \epsilon, & \text{if } I_{i,j}^{h2,h_1} = 0 \end{cases} \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, s^2)$ is a small random elementwise noise added to break network symmetry while still maintaining the injected structural prior during initial phases of optimization. The scalar Δ defines the logit gap between *user-defined* connections and unmapped relations. The two values act as positive and negative biases toward valid connections and unmapped relationships. The value of Δ is selected in the unconstrained logit space to encode a soft structural prior over the hierarchy and is used to provide small nonzero probability values to unmapped relations in the *user-defined* taxonomy. Indeed, during the forward pass, these unconstrained parameters are dynamically mapped into valid hierarchy log joint distribution matrices $\log \mathbf{J}^{h_2,h_1}$ via a log-softmax operation on the flattened parameter matrices $\mathbf{W}^{h_2,h_1}$. Each term of the matrix $J_{i,j}^{h_2,h_1} = P(\omega_j^{h_1}, \omega_i^{h_2})$ models the probability of observing a relationship between microclass ω^{h_2} and macroclass ω^{h_1}. This constraint relaxation allows us to better combine the *user-defined* aggregation with *data-driven* calibration of cross-level compatibility, while retaining the initial *user-defined* mappings as the structural anchor.

D. Self-Consistent HC Learning

Let us return to the general case of H hierarchical levels, considering projections from source levels h to a target hierarchical level $\tilde{h}$. Each hierarchical level acts independently of the others. To ensure that the network is *hierarchy-aware*, an additional constraint based on the prediction associated with the target hierarchical level and the projected predictions from the other levels is defined. In particular, we take inspiration from the disagreement-based active learning field [22] to define the disagreement between the predictions calculated at each hierarchical level. For each level $\tilde{h}$, we consider a committee composed of all H levels. All predictions and their projected versions can be viewed as "soft" votes, which define each committee member's confidence. Then, we introduce a self-consistent loss that penalizes the committee members' disagreement with the overall committee consensus, calibrating the mapping weights in $\mathbf{J}^{h,\tilde{h}}$ when adaptive projectors are used, and providing further supervision to the network based on the predictions at the different levels.

1) Hierarchy-Based Projected Predictions: In order to generate the committee at each hierarchical level, the predictions first need to be projected to the other levels. For simplicity, let us consider the projection from a source-level h to a target level $\tilde{h}$. Let $\hat{\mathbf{J}}^{h,\tilde{h}} \in [0,1]^{|\Omega^h|\times|\Omega^{\tilde{h}}|}$, with $h, \tilde{h} \in \mathcal{H}, h \neq \tilde{h}$, denote the joint distribution matrices. From $\hat{\mathbf{J}}^{h,\tilde{h}}$, we derive the conditional projection matrix $\hat{\mathbf{M}}^{h,\tilde{h}} \in [0,1]^{|\Omega^h|\times|\Omega^{\tilde{h}}|}$ by row-normalization

$$\hat{M}_{i,j}^{h,\tilde{h}} = \hat{P}\left(\omega_j^{\tilde{h}} | \omega_i^h\right) = \frac{\hat{J}_{i,j}^{h,\tilde{h}}}{\sum_{k=1}^{|\Omega^{\tilde{h}}|} \hat{J}_{i,k}^{h,\tilde{h}}}. \tag{7}$$

Therefore, the projected probability distribution $\hat{P}^{h\to\tilde{h}}(\omega_j^{\tilde{h}}|\mathbf{x})$ is obtained from the source-level posterior $\hat{P}^h(\omega_i^h|\mathbf{x})$ and the conditional projection probabilities $\hat{P}(\omega_j^{\tilde{h}}|\omega_i^h)$ as follows:

$$\hat{P}^{h\to\tilde{h}}\left(\omega_j^{\tilde{h}}|\mathbf{x}\right) = \sum_{i=1}^{|\Omega^h|} \hat{P}\left(\omega_j^{\tilde{h}}|\omega_i^h\right) \cdot \hat{P}^h\left(\omega_i^h|\mathbf{x}\right). \tag{8}$$

This corresponds to the input-independent hierarchy projection assumption $\hat{P}(\omega_j^{\tilde{h}}|\omega_i^h, \mathbf{x}) \approx \hat{P}(\omega_j^{\tilde{h}}|\omega_i^h)$. The dependency on the input $\mathbf{x}$ is preserved through the source-level posterior $\hat{P}^h(\omega_i^h|\mathbf{x})$, while the conditional hierarchy transition is modeled as a global projection probability shared across the dataset. In general vector form, we can write the class-posterior probability vector $\hat{\mathbf{p}}^{h\to\tilde{h}}(\mathbf{x})$ projected from level h to level $\tilde{h}$ as follows:

$$\hat{\mathbf{p}}^{h\to\tilde{h}}(\mathbf{x}) = \left(\hat{\mathbf{M}}^{h,\tilde{h}}\right)^{\mathsf{T}} \cdot \hat{\mathbf{p}}^h(\mathbf{x}). \tag{9}$$

The row-normalization in (7) constrains each row in $\hat{\mathbf{M}}^{h,\tilde{h}}$ to sum to one. Therefore, although the hierarchy parameters $\mathbf{W}^{h,\tilde{h}}, h, \tilde{h} \in \mathcal{H}, h \neq \tilde{h}$, are unconstrained in the logit space, the quantities used to project predictions across hierarchical levels are valid conditional probability distributions.

To ensure numerical stability and maintain better control over gradient magnitudes, we consider the computation of all the losses directly in the log domain. First, we compute the

class-posterior log probabilities of classification modules $\mathbf{g}^h$ from (2)

$$\log \hat{P}^h\left(\omega_i^h|\mathbf{x}\right) = g_i^h - \mathrm{LSE}\left(\mathbf{g}^h\right) \tag{10}$$

where $\mathrm{LSE}: \mathbb{R}^{|\Omega^h|} \to \mathbb{R}$ is the log–sum–exp (LSE) operator

$$\mathrm{LSE}\left(\mathbf{g}^h\right) = \log \sum_{i=1}^{|\Omega^h|} \exp\left(g_i^h\right). \tag{11}$$

Then, we can also define the projected predictions $\hat{P}^{h\to\tilde{h}}(\omega_j^{\tilde{h}}|\mathbf{x})$ from level h to level $\tilde{h}$ in the log domain, starting from the definition in the natural domain. Applying the logarithm to both sides of (8), we obtain the following:

$$\begin{aligned}
&\log\left(\hat{P}^{h\to\tilde{h}}\left(\omega_j^{\tilde{h}}|\mathbf{x}\right)\right) \\
&= \log\left(\sum_i \hat{M}_{i,j}^{h,\tilde{h}} \cdot \hat{P}^h\left(\omega_i^h|\mathbf{x}\right)\right) \\
&= \log\left(\sum_i \hat{M}_{i,j}^{h,\tilde{h}} \cdot \frac{\exp g_i^h}{\sum_k \exp g_k^h}\right) \\
&= \log\left(\sum_i \exp\left(g_i^h + \log \hat{M}_{i,j}^{h,\tilde{h}}\right)\right) - \log \sum_k \exp g_k^h.
\end{aligned} \tag{12}$$

This can finally be rewritten as the combination of two LSE operations, yielding the projected log-probability distribution at level $\tilde{h}$

$$\log\left(\hat{P}^{h\to\tilde{h}}\left(\omega_j^{\tilde{h}}|\mathbf{x}\right)\right) = \mathrm{LSE}\left(\mathbf{g}^h + \log\left(\hat{\mathbf{M}}_{\cdot,j}^{h,\tilde{h}}\right)\right) - \mathrm{LSE}\left(\mathbf{g}^h\right). \tag{13}$$

Hierarchy projection matrices can be defined directly in the log domain, i.e., we retrieve log joint probabilities $\log \hat{\mathbf{J}}^{h,\tilde{h}} \in \mathbb{R}^{|\Omega^h|\times|\Omega^{\tilde{h}}|}$ via learned parameter matrices $\mathbf{W}^{h,\tilde{h}}$

$$\log \hat{\mathbf{J}}^{h,\tilde{h}} = \mathbf{W}^{h,\tilde{h}} - \mathrm{LSE}\left(Flatten\left(\mathbf{W}^{h,\tilde{h}}\right)\right). \tag{14}$$

Then, we compute $\log \hat{\mathbf{M}}^{h,\tilde{h}}$ as follows:

$$\log \hat{\mathbf{M}}_{i,\cdot}^{h,\tilde{h}} = \log \hat{\mathbf{J}}_{i,\cdot}^{h,\tilde{h}} - \mathrm{LSE}\left(\log \hat{\mathbf{J}}_{i,\cdot}^{h,\tilde{h}}\right) \tag{15}$$

which is equivalent to the row-normalization of $\hat{\mathbf{J}}^{h,\tilde{h}}$ in the log domain. The projected logits $\log \hat{\mathbf{p}}^{h\to\tilde{h}}$ can be used to calculate additional CCE losses that can contribute to the final loss [23]. However, this does not explicitly model the HC at different hierarchical levels, requiring an additional step to ensure the joint classification of the classes.

This modeling choice implies that the projection matrices encode dataset cross-level relationships rather than sample-specific transitions. When projecting from a coarse class to finer classes, the projector models global relations in the training distribution, while the direct target-level classifier retains the spectral–temporal characteristics and uncertainty of the individual sample. This formulation provides a compact and interpretable hierarchy-aware regularization mechanism, avoiding the modeling of sample-dependent projectors.

2) Semantics-Aware Consensus: Given the projected prediction at each hierarchical level, we can now define a HC class probability vector $\hat{\mathbf{p}}_{HC}^{\tilde{h}}(\mathbf{x})$

$$\begin{aligned}
\hat{\mathbf{p}}_{HC}^{\tilde{h}}(\mathbf{x}) &= \frac{1}{Z^{\tilde{h}}} \sqrt[H]{\hat{\mathbf{p}}^{\tilde{h}}(\mathbf{x}) \prod_{h\neq\tilde{h}} \hat{\mathbf{p}}^{h\to\tilde{h}}(\mathbf{x})} \\
&= \frac{1}{Z^{\tilde{h}}} \sqrt[H]{\hat{\mathbf{p}}^{\tilde{h}}(\mathbf{x}) \prod_{h\neq\tilde{h}} (\hat{\mathbf{M}}^{h,\tilde{h}})^{\mathsf{T}} \cdot \hat{\mathbf{p}}^{h}(\mathbf{x})}
\end{aligned} \tag{16}$$

where $\hat{\mathbf{p}}_{HC}^{\tilde{h}}$ is the HC class-posterior probability defined as the geometric mean of the predictions at the target hierarchical level $\tilde{h}$ and the projected predictions of all the other h levels, and $Z^{\tilde{h}}$ is a normalization constant to ensure that $\mathbf{p}_{HC}^{\tilde{h}}$ sums to one. This is equivalent to computing the average in the log domain [24]

$$\begin{aligned}
g_{HC_j}^{\tilde{h}} &= \frac{1}{H}\left[\log \hat{P}^{\tilde{h}}\left(\omega_j^{\tilde{h}}|\mathbf{x}\right) + \sum_{h\neq\tilde{h}} \log \hat{P}^{h\to\tilde{h}}\left(\omega_j^{\tilde{h}}|\mathbf{x}\right)\right] \\
&= \frac{1}{H}\left\{\left[g_j^{\tilde{h}} - \mathrm{LSE}\left(\mathbf{g}^{\tilde{h}}\right)\right] \right. \\
&\qquad \left. + \sum_{h\neq\tilde{h}}\left[\mathrm{LSE}\left(\mathbf{g}^h + \log \hat{\mathbf{M}}_{\cdot,j}^{h,\tilde{h}}\right) - \mathrm{LSE}\left(\mathbf{g}^h\right)\right]\right\} \\
\log \hat{\mathbf{p}}_{HC}^{\tilde{h}}(\mathbf{x}) &= \mathbf{g}_{HC}^{\tilde{h}} - \mathrm{LSE}\left(\mathbf{g}_{HC}^{\tilde{h}}\right)
\end{aligned} \tag{17}$$

where $\mathrm{LSE}(\mathbf{g}_{HC}^{\tilde{h}}) = \log Z^{\tilde{h}}$. This HC represents the central SAHC consensus found by the voting committee and can serve as an optional hierarchy-aware prediction during inference. The direct prediction $\hat{\mathbf{p}}^{\tilde{h}}$ preserves sample-specific evidence from the target-level classifier, while the consensus promotes agreement with the remaining levels. The normalized geometric mean is selected because the HC is intended to emphasize agreement among the committee members rather than to provide an average. This differs from an arithmetic mean, where a class can maintain high probability also when it is strongly supported by only one committee member despite disagreement from the others.

3) Consensus Loss: After calculating the consensus, both the prediction at the target level and the projected predictions are compared with the HC, defining an additional loss based on the divergence from the HC. To this purpose, we employ a JSD loss rather than a standard Kullback–Leibler divergence (KLD). While the KLD is widely used to measure the difference between two probability distributions, its asymmetry makes it better suited for scenarios where one distribution is a fixed "ground truth" or prior. In this case, all hierarchical levels provide dynamically trained "soft" votes, with no reference distribution. The symmetry in the JSD ensures balanced regularization between the projected predictions and the consensus, also providing stabler gradients due to its bounded nature. In the proposed framework, the JSD loss is computed as follows:

$$\mathrm{JSD}^{h,\tilde{h}}\left(\hat{\mathbf{p}}_{HC}^{\tilde{h}} \| \hat{\mathbf{p}}^{h\to\tilde{h}}\right) = \frac{1}{2}\left(\mathrm{KLD}\left(\hat{\mathbf{p}}_{HC}^{\tilde{h}} \| \mathbf{q}\right) + \mathrm{KLD}\left(\hat{\mathbf{p}}^{h\to\tilde{h}} \| \mathbf{q}\right)\right) \tag{18}$$

where $\mathbf{q}$ is a mixture distribution of $\hat{\mathbf{p}}_{HC}^{\tilde{h}}$ and $\hat{\mathbf{p}}^{h\to\tilde{h}}$ calculated as $\mathbf{q} = 1/2(\hat{\mathbf{p}}_{HC}^{\tilde{h}} + \hat{\mathbf{p}}^{h\to\tilde{h}})$, KLD represents the Kullback–Leibler divergence, and we omit the dependency on $\mathbf{x}$ for simplicity. Please note that when $h = \tilde{h}$, we have that $\hat{\mathbf{p}}^{h\to\tilde{h}} = \hat{\mathbf{p}}^{\tilde{h}}$. Finally,

we can retrieve the JSD losses calculated at the different hierarchical levels and compute the final $\mathcal{L}_{HC}$ penalty loss

$$\mathcal{L}_{HC} = \frac{1}{N}\sum_{n=1}^{N}\sum_{\tilde{h}=1}^{H}\frac{1}{\log|\Omega^{\tilde{h}}|}\sum_{h=1}^{H}\mathrm{JSD}^{h,\tilde{h}}(\mathbf{x}_n) \tag{19}$$

where the logarithm of the number of classes at level $\tilde{h}$, i.e., $|\Omega^{\tilde{h}}|$, allows the balancing of the different numbers of classes at each hierarchical level. This prevents skewness toward levels with a larger number of classes. The $\mathcal{L}_{HC}$ measures the discrepancy between the HC prediction and the projected predictions, encouraging the different layers to converge to the committee's vote.

Finally, we also apply full supervision on the consensus predictions at the different levels, considering a standard CCE loss

$$\mathcal{L}_{CCE}^{HC} = \sum_{h=1}^{H} -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|\Omega^h|}\mathbb{1}\left[y_n^h = \omega_i^h\right]\log\hat{P}_{HC}^h\left(\omega_i^h|\mathbf{x}_n\right). \tag{20}$$

E. Overall Loss Function

Let $\theta_{\mathbf{f}}$ be the parameters of the backbone feature extractor $(\mathbf{f};\theta_{\mathbf{f}})$, and let $\theta_{\mathbf{g}^h}$ be the parameters of the classification head $(\mathbf{g}^h;\theta_{\mathbf{g}^h})$ associated with level h. Moreover, let $\mathcal{W} = \{\mathbf{W}^{a,b}\}_{1\le a<b\le H}$ be the set of trainable hierarchy-parameter matrices. The complete set of trainable parameters can be defined as follows:

$$\Theta = \{\theta_{\mathbf{f}}\}\cup\left\{\theta_{\mathbf{g}}^h\right\}_{h=1}^{H}\cup\mathcal{W}. \tag{21}$$

The final optimization problem is defined as

$$\Theta^\star = \arg\min_{\Theta}\mathcal{L}(\Theta) \tag{22}$$

where the final loss function $\mathcal{L}$ is derived from the sum of the losses computed in (4), (19), and (20)

$$\begin{aligned}\mathcal{L}(\Theta) = {} & \mathcal{L}_{CCE}\left(\theta_{\mathbf{f}},\left\{\theta_{\mathbf{g}}^h\right\}_{h=1}^{H}\right)\\ & +\lambda_{HC}\left(\mathcal{L}_{CCE}^{HC}\left(\theta_{\mathbf{f}},\left\{\theta_{\mathbf{g}}^h\right\}_{h=1}^{H},\mathcal{W}\right)\right.\\ & \left.+\mathcal{L}_{HC}\left(\theta_{\mathbf{f}},\left\{\theta_{\mathbf{g}}^h\right\}_{h=1}^{H},\mathcal{W}\right)\right)\end{aligned} \tag{23}$$

where λ_{HC} is a metaparameter tuning the contribution of the HC losses. It is important to note that $\mathcal{L}_{HC}$ does not require direct supervision, as it is simply a loss representing the divergence from the HC. The HC prediction is used both within the penalty loss and the additional supervised CCE loss during the training phase. The HC prediction at each hierarchical level can also be used at inference to exploit the cross-level agreement learned during self-consistent training, supporting coherent hierarchical predictions. The pseudocode of the proposed framework is detailed in Algorithm 1.

F. Computational Complexity and Scalability

The SAHC formulation learns a hierarchy parameter matrix for every unordered pair of hierarchical levels. Let $|\Omega^h|$ be the number of classes at level h. Although the number of hierarchy matrices is $H(H-1)/2$, their practical memory requirement depends on the cardinality of the corresponding label spaces. The total number of trainable hierarchy parameters is

$$P_{\text{hier}} = \sum_{1\le h_1<h_2\le H}\left|\Omega^{h_1}\right|\left|\Omega^{h_2}\right|. \tag{24}$$

The matrices $\log\mathbf{J}^{h_2,h_1}$ and $\log\mathbf{M}^{h,\tilde{h}}$ are differentiable quantities derived from the trainable matrices $\mathbf{W}^{h_2,h_1}$. Thus, they do not introduce additional independent parameters. The additional parameters associated with the level-specific linear heads scale as $\mathcal{O}\left(d\sum_{h=1}^{H}|\Omega^h|\right)$, where d is the backbone feature dimension. For a mini-batch of size B, the consensus requires a projection from each source level to each distinct target level. Since the projection in (9) has complexity of $\mathcal{O}(B|\Omega^h||\Omega^{\tilde{h}}|)$, the overall cross-level projection cost is

$$\mathcal{O}\left(2BP_{\text{hier}}\right). \tag{25}$$

The consensus fusion and JSD operate on probability vectors and scale as $\mathcal{O}\left(BH\sum_{h=1}^{H}|\Omega^h|\right)$. Therefore, when all

Algorithm 1 Proposed SAHC Training Approach

Input: Training set $\mathcal{D}$, hierarchical levels $\{\Omega^h\}_{h=1}^H$, indicator matrices $\{\mathbf{I}^{a,b}\}_{1\le a<b\le H}$, loss weights $\{\lambda^h\}_{h=1}^H$, λ_{HC}, warm-up schedule $\alpha(t)$.

Output: Optimized parameters $\Theta = \{\theta_{\mathbf{f}}\}\cup\{\theta_{\mathbf{g}}^h\}_{h=1}^H\cup\mathcal{W}$.

Initialize the backbone parameters $\theta_{\mathbf{f}}$ and the classification-head parameters $\{\theta_{\mathbf{g}}^h\}_{h=1}^H$.
Initialize each hierarchy-parameter matrix $\mathbf{W}^{a,b}\in\mathcal{W}$ from the corresponding fixed indicator matrix $\mathbf{I}^{a,b}$ according to (6).
for each training epoch t **do**
  for each mini-batch $\mathcal{B}\subset\mathcal{D}$ **do**
    Compute features $\mathbf{z} = \mathbf{f}(\mathbf{x};\theta_{\mathbf{f}}),\quad \mathbf{x}\in\mathcal{B}$.
    For each level h, compute logits $\mathbf{r}^h$ and probabilities $\hat{\mathbf{p}}^h$:
    $$\mathbf{r}^h = \mathbf{g}^h(\mathbf{z};\theta_{\mathbf{g}}^h),\qquad \hat{\mathbf{p}}^h = \sigma(\mathbf{r}^h).$$
    Compute the multi-level supervised loss $\mathcal{L}_{\text{CCE}}$ using (4).
    For each hierarchy pair (a,b), compute $\log\mathbf{J}^{a,b}$ and the corresponding conditional projection matrices $\log\mathbf{M}^{h,\tilde{h}}$.
    for each target level $\tilde{h} = 1,\dots,H$ **do**
      Project the predictions from all source levels $h\neq\tilde{h}$ to level $\tilde{h}$ using (13).
      Compute the hierarchical consensus prediction $\hat{\mathbf{p}}_{\text{HC}}^{\tilde{h}}$ using (17).
    end for
    Compute the consensus loss $\mathcal{L}_{\text{HC}}$ using (19).
    Compute the supervised consensus loss $\mathcal{L}_{\text{CCE}}^{\text{HC}}$ using (20).
    Minimize
    $$\mathcal{L} = \mathcal{L}_{\text{CCE}} + \alpha(t)\lambda_{\text{HC}}(\mathcal{L}_{\text{CCE}}^{\text{HC}} + \mathcal{L}_{\text{HC}}).$$
    Update $\theta_{\mathbf{f}}$, $\{\theta_{\mathbf{g}}^h\}_{h=1}^H$, and $\mathcal{W}$ by back-propagation.
  end for
end for

levels have an average cardinality $|\bar{\Omega}| = C$, the implementation has a quadratic dependence on hierarchy depth, i.e., $\mathcal{O}(H^2C^2)$ for the hierarchy parameters and cross-level projections. For a deep RS hierarchy, as the one analyzed later in this study before class filtering, a hierarchy with $H = 4$ and $(|\Omega^1|, |\Omega^2|, |\Omega^3|, |\Omega^4|) = (5, 15, 40, 90)$ classes per-level, the complete formulation contains six trainable matrices and 6275 parameters, usually negligible relative to the backbone parameters and the added projection cost remains modest.

IV. Dataset Description and Experimental Setup

This section provides a description of the datasets employed in this study. Then, the experimental setup, including the preprocessing phase and the benchmark methodologies adopted for evaluating the DL models, is described.

A. NWPU-RESISC45

To assess the performance of the proposed methodology on VHR images, where spatial features play a prominent role, we considered the NWPU-RESISC45 [25] benchmark as a first dataset. NWPU-RESISC45 is composed of 31 500 RGB multiresolution images for scene classification of VHR images categorized into 45 different classes at the finest level of detail, and exhibits a high within-class diversity and between-class similarity. The dataset is completely balanced, i.e., each class has the same number of samples at the *fine-grained* level. We defined a custom hierarchy based on the subdivision defined in MillionAID [26], resulting in 8, 24, and 45 classes, corresponding to *coarse-grained*, *intermediate*, and *fine-grained* classes, respectively.

B. ELU Dataset

To further validate the performance of the proposed methodology and demonstrate its generalizability on different LC mapping tasks, we additionally considered a multitemporal multispectral (MS) dataset differing in terms of complexity in the hierarchical structure. This dataset[1] consists of a publicly available land-use (LU) map of 2020 over Emilia Romagna region, Italy. The maps were defined considering the standard CLC aggregation at four levels, showcasing a highly detailed semantic dataset. For the purpose of this analysis, we evaluated the hierarchical approach on four levels of hierarchy, consisting of the first three levels of the CLC and the additional finest level of detail defined by the LULC, here referred to as *macro*, *coarse-grained*, *intermediate*, and *fine-grained*, respectively. The dataset has been built considering orthophotographs and infrared images, and shows a minimum mapping unit of 0.16 hectares, with a minimum of 7 m for linear structures. In addition, it represents LU classes, where the same LC class could be associated with different semantic classes, such as the fine-grained classes "commercial airports," "sport flight airports and heliports," and "military airports." We considered Emilia land use (ELU) dataset corresponding to the Military Grid Reference System (MGRS) Sentinel-2 (S2) granule *32TPQ* (10 980 × 10 980 pixels, for a total area of 110 × 110 km^2). In the analyzed granule, the four hierarchical levels show 5, 15, 40, and 90 classes at the macro, coarse-grained, intermediate, and fine-grained levels, respectively. Please note that such an elevated number of fine-grained classes is not commonly approached in Earth observation (EO) classification problems, where the majority of studies focus on the intermediate or coarse-grained classification task. Nonetheless, this level of precision allows better monitoring of the class of interest and the extraction of information-rich content. Due to the high number of classes and, consequently, the large size of the hierarchical tree, we refer to this hierarchy as *a complex* hierarchy. A semicircular tree representing the hierarchical structure of the dataset is shown in Fig. 4.

[1][Online]. Available: (https://geoportale.regione.emilia-romagna.it/catalogo/dati-cartografici/pianificazione-e-catasto/uso-del-suolo/layer-14). Accessed April 9, 2026.

C. S2 Preprocessing and Dataset Preparation

1) VHR Dataset: The training, validation, and test splits were selected according to [27], considering a split ratio of 0.6, 0.2, and 0.2, respectively. During training, we applied data augmentation techniques consisting of resize, random crops, and random rotations of the images.

2) Multitemporal Ms Datasets: The input data consists of S2 MS satellite images provided in the Copernicus EO program. The S2 acquisitions considered have been preprocessed at Level-2A (L2A) and reprojected at 10-m/pixel resolution through a nearest-neighbor approach. In the experiments, all the bands except the 60-m resolution band were used. Due to the high complexity of this classification problem, we collected a time series (TS) of S2 images, fully exploiting the temporal signature and allowing the detection of seasonal targets. The S2 TS considered for Emilia dataset consists of acquisitions ranging from January 1, 2020 to December 31, 2020. The S2 TS has been processed to achieve a more regular TS, resulting in 12 medoid monthly composites [28].

To extract a balanced training set, we considered an approach based on the heterogeneity of the input training data. In particular, we divided the initial S2 monthly composite images into nonoverlapping tiles of 1000 × 1000 pixels (100 km^2). After the initial subdivision, we randomly subdivided the tiles into training, validation, and test sets by considering a split ratio of 0.7, 0.15, and 0.15, respectively. We extracted $N = 2000$ representative points for each tile and class in the analyzed classification scheme. Classes not present in all sets were excluded from the analysis, resulting in 74, 33, 13, and 5 classes for the different hierarchical levels. In the case of minority classes having less than N samples, all the available samples were extracted. The training set was generated by extracting patches of 9 × 9 pixels around the selected points in the training tiles. To minimize spatial autocorrelation between the training patches, the corresponding reference data was extracted only for the central 3 × 3 pixels. In this way, we collected a training dataset representative of the entirety of the classification scheme, corresponding to ~1.17% (1 413 522 pixels) of the total number of pixels available in the MGRS S2 granule.

D. Design of Experiments

Validation has been conducted on both datasets considering different multigranularity approaches. The reported results include the values of overall accuracy (OA) and

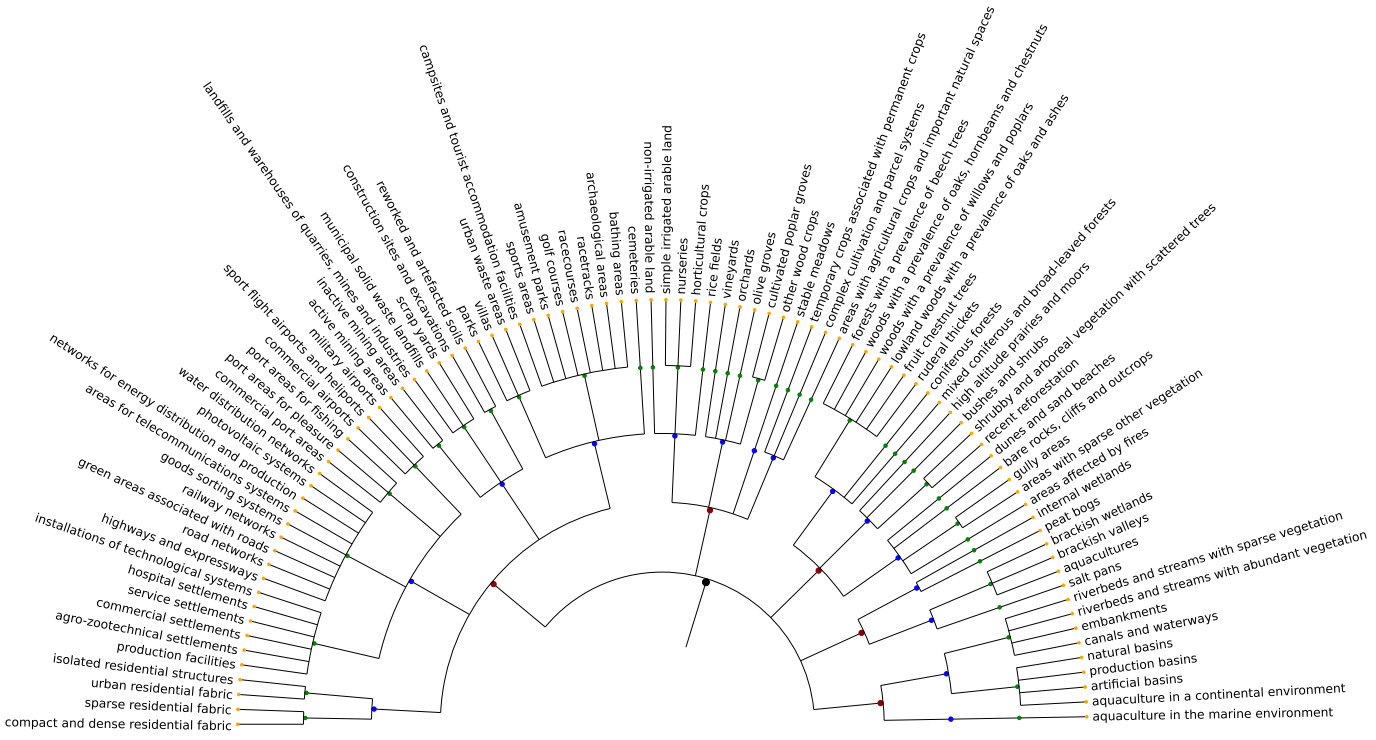


Fig. 4. Semicircular hierarchical label tree representation of the ELU dataset at the fine-grained level of the hierarchy. The leaves corresponding to the fine-grained classes are shown in orange, while the nodes for intermediate classes are depicted in green, those for coarse-grained classes are shown in blue, and the nodes for macro in brown. The root node is represented in black.

mean $F1$ (m$F1$) score for both tasks, while for the MS dataset also Top-3 and Top-5 metrics are reported. We considered different MGMs for comparison, all built on the same backbone architecture for fairness on all datasets. We considered the following approaches.

1) *Baseline* trained only with the level-specific hierarchical head through standard CCE minimization. Selected to evaluate the improvement gained by incorporating the hierarchy.
2) *HAF* [13] selected as a regularization approach adopting soft labels and geometric constraints that use auxiliary classifiers at each level to ensure consistent predictions with the hierarchy, specifically aiming to reduce the severity of mistakes.
3) *HRN* [15] included as a state-of-the-art benchmark that facilitates hierarchical feature interaction by forcing finer classes to inherit attributes from coarser classes through residual skip connections.
4) *MGM* [16] included to assess the effectiveness of using knowledge-guided relation graphs and bidirectional interaction structures that enhance coarse-grained features through gradient truncations.
5) *HiCervix* [17] chosen as a benchmark for methods that model semantic correlations within an explicit tree structure using soft labels and recursive probability summation across hierarchical paths.
6) *ALGS* [6] selected to evaluate the impact of a global semantic distance-based penalty term for adjusting class predictions, ensuring hierarchical consistency when classification is enforced at every level. Specifically designed for RS tasks.
7) *HCL* [7] selected as a comparison for representation-based learning that disentangles class-specific embeddings at multiple levels to align them with their semantic relationships. Specifically designed for RS tasks.

For the VHR task, we considered a ResNet50 backbone pretrained with ImageNet weights [29]. This choice is due to its deeply stacked convolutional layers and residual connections, making it well-suited for extracting highly localized, fine-grained spatial and morphological features necessary for VHR analysis. For the MS satellite image time series (SITS) classification task, we adopted a different backbone architecture designed to process MS and temporal information. This backbone consists of an adaptation to the RS scenario of a *Swin Transformer* [30] and is trained from scratch. For a detailed description of the layer topology, please refer to the cited paper [30]. The choice of this backbone is due to its shifted-window self-attention mechanism that efficiently captures both local details and long-range global contextual dependencies, making it ideal for processing complex multiscale structural variations.

We adopted the *AdamW* optimizer for all the architectures and configurations. The metaparameters of each different configuration have been validated considering a tree-structured Parzen estimator (TPE) algorithm to efficiently explore the metaparameter space. The learning rate and the weight decay were sampled from log-uniform distributions $U_{\log}([10^{-6}, 10^{-3}])$. We considered 50 training epochs on both datasets, together with stochastic weight averaging (SWA) to enhance model generalization [31]. Considering the proposed SAHC approach, the weight λ_{HC} associated with the HC loss has been sampled from a log-uniform distribution

TABLE I

COMPARISON OF THE ACCURACIES OBTAINED BY DIFFERENT HIERARCHICAL METHODOLOGIES ON THE NWPU-RESISC45 DATASET. THE BEST RESULTS ARE REPORTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE RELATED STANDARD DEVIATION IS REPORTED IN PARENTHESES

Metric	Hierarchy	Baseline [29]	HRN [15]	HAF [13]	ALGS [6]	HCL [7]	MGM [16]	HiCervix [17]	SAHC
OA (%)	Fine-grained	96.12 (0.12)	<u>96.34 (0.19)</u>	96.05 (0.13)	96.28 (0.06)	95.70 (0.16)	96.20 (0.10)	95.55 (0.48)	**96.55 (0.08)**
	Intermediate	97.29 (0.11)	97.30 (0.14)	97.13 (0.18)	<u>97.41 (0.05)</u>	97.00 (0.22)	97.27 (0.23)	96.88 (0.27)	**97.48 (0.07)**
	Coarse-grained	97.43 (0.08)	97.82 (0.11)	97.76 (0.09)	<u>97.85 (0.11)</u>	97.57 (0.16)	97.79 (0.25)	97.50 (0.20)	**97.90 (0.07)**
mF1(%)	Fine-grained	96.11 (0.11)	<u>96.34 (0.19)</u>	96.05 (0.13)	96.28 (0.06)	95.70 (0.15)	96.20 (0.11)	95.55 (0.48)	**96.56 (0.07)**
	Intermediate	96.59 (0.10)	96.56 (0.19)	96.20 (0.23)	<u>96.65 (0.04)</u>	96.05 (0.19)	96.52 (0.42)	95.97 (0.35)	**96.75 (0.16)**
	Coarse-grained	97.07 (0.10)	97.50 (0.17)	97.39 (0.11)	<u>97.55 (0.10)</u>	97.16 (0.18)	97.41 (0.31)	97.10 (0.26)	**97.58 (0.09)**
Params (M)		25.8	35.5	25.9	25.9	27.0	35.5	25.9	25.9
FLOP (T)		6.22	7.03	6.22	6.22	6.51	6.88	6.22	6.22
Inference time[a] (ms)		135.48	139.23	135.68	135.62	138.21	139.30	135.73	135.62
Peak VRAM[a] (MB)		2811.16	2851.75	2812.60	2812.42	2817.01	2851.88	2812.91	2812.24

[a] Results for a single forward pass over a batch of 256 images on an NVIDIA L40S GPU. Inference time is averaged over 50 forward passes.

$U_{\log}([10^{-2}, 10^{2}])$, while the weights λ^h have been set to 0.5, 0.2, and 0.3 for the VHR dataset, and to 0.4, 0.2, 0.1, and 0.3 for the MS dataset, to prioritize the fine-grained classification and reduce the influence of the intermediate, coarse-grained, and macro levels. For the NWPU-RESISC45 dataset, these weights were selected on the validation set through a simplex grid search. For the ELU dataset, we heuristically fixed the weights according to the trend observed in the NWPU-RESISC45. In both datasets, the learnable parameter matrices $\mathbf{W}^{h,\tilde{h}}$ were initialized with $\Delta = 5$ and $s = 0.01$. We implemented a dynamic consistency warmup schedule for the HC loss weight, λ_{HC}. Specifically, during the training phase, λ_{HC} was initially fixed to 0 for the first 5 epochs to allow for stable feature extraction. Subsequently, the weight value was linearly ramped up from epoch 5–15 until reaching its sampled maximum value, where it remained constant for the duration of the training. We report the average scores on three runs, together with their unbiased estimation of standard deviation.

V. EXPERIMENTAL RESULTS

The experiments focus on validating the proposed SAHC approach (with HC predictions at inference time) over different datasets, comparing it to other state-of-the-art hierarchical approaches on different levels of the hierarchy. We first focus on the results on the specific tasks, i.e., scene classification on the VHR dataset and semantic segmentation of MS SITS. After this, we focus on ablation studies on the divergence module and the fusion operator. Then, we analyze the importance of the hierarchical weights selection λ^h on the NWPU dataset. Subsequently, a complete ablation study on both datasets on each component of the proposed SAHC is conducted, as well as a sensitivity analysis on the initializations of the projectors. Finally, we discuss a visualization of the trained hierarchy projection matrices to identify the key hierarchical relationships learned by the proposed approach.

A. Case I: VHR Scene Classification

The performance of the proposed SAHC method on the NWPU-RESISC45 dataset is summarized in Table I, providing a comparative analysis against state-of-the-art methodologies and the nonhierarchical baseline. Given the balanced class distribution characteristic of the NWPU-RESISC45 dataset, the OA and m$F1$ are aligned across all evaluated models. From the table, one can see that the integration of hierarchical information consistently improves classification performance over the baseline at each hierarchical level. Indeed, the evaluated hierarchical architectures outperform their nonhierarchical counterpart, confirming that the exploitation of semantic relationships allows for a more robust distinction between fine-grained RS classes. Among the compared methodologies, the proposed SAHC achieves the highest performance at the fine-grained level, with an OA of 96.55% and an m$F1$ of 96.56%. Notably, the good performance of SAHC extends across the entire taxonomic hierarchy. At the intermediate and coarse-grained levels, the proposed method achieves OAs of 97.48% and 97.90%, respectively, consistently outperforming the other frameworks. The second-best overall performance is obtained by the HRN method at the most detailed hierarchical level, which reaches a fine-grained OA of 96.34%. At the coarsest levels, ALGS outperforms the other state-of-the-art methodologies, while still performing slightly lower compared with the proposed SAHC. This comparison also highlights the computational overhead introduced by the different architectures. The HRN method achieves its high accuracy at the cost of architectural efficiency. Indeed, the introduction of the granularity-specific block (GSB) leads to a significant increase in parameters and computational overload. In contrast, the proposed SAHC maintains a nearly identical computational footprint as the baseline architecture. This efficiency is shared by other methods such as ALGS, HAF, and HiCervix, where the bottleneck remains the backbone itself. In the specific case of SAHC, the architectural overhead is due to the three auxiliary heads and their associated projection matrices, both of which are negligible relative to the dimension of the backbone. When focusing strictly on architectures that do not introduce additional complexity, the ALGS method provides the highest performance after SAHC, though it remains lower in terms of OA.

TABLE II
COMPARISON OF THE ACCURACIES OBTAINED BY DIFFERENT HIERARCHICAL METHODOLOGIES ON THE ELU DATASET. THE BEST RESULTS ARE REPORTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE RELATED STANDARD DEVIATION IS REPORTED IN PARENTHESES

Metric	Hierarchy	Baseline [29]	HRN [15]	HAF [13]	ALGS [6]	HCL [7]	MGM [16]	HiCervix [17]	SAHC
OA (%)	Fine-grained	36.84 (0.37)	36.49 (0.19)	36.11 (0.69)	35.94 (0.07)	37.58 (0.30)	35.54 (0.27)	37.30 (0.83)	**38.38 (0.31)**
	Intermediate	47.84 (0.46)	46.57 (0.03)	39.72 (1.73)	46.18 (0.23)	49.20 (0.37)	46.45 (0.43)	48.18 (0.65)	**50.20 (0.13)**
	Coarse-grained	57.64 (0.07)	56.79 (0.30)	57.94 (0.50)	56.29 (0.12)	59.86 (0.25)	56.53 (0.28)	58.55 (0.51)	**60.21 (0.21)**
	Macro	77.55 (0.04)	75.95 (0.44)	**78.88 (0.14)**	75.63 (0.12)	78.47 (0.19)	75.76 (0.39)	77.52 (0.11)	78.33 (0.17)
Top-3 (%)	Fine-grained	56.78 (0.03)	56.50 (0.94)	54.84 (1.37)	54.83 (0.08)	57.29 (0.31)	56.12 (0.27)	57.97 (1.42)	**59.33 (0.35)**
	Intermediate	70.86 (0.50)	70.27 (0.64)	65.29 (1.48)	69.28 (0.32)	72.00 (0.36)	70.17 (0.24)	71.82 (0.80)	**73.71 (0.23)**
	Coarse-grained	81.35 (0.11)	81.75 (0.67)	81.12 (0.13)	80.79 (0.21)	83.00 (0.27)	81.47 (0.32)	82.71 (0.68)	**84.49 (0.12)**
	Macro	96.58 (0.08)	96.89 (0.11)	95.97 (0.51)	97.04 (0.15)	97.21 (0.15)	96.79 (0.20)	**97.29 (0.19)**	97.15 (0.17)
Top-5 (%)	Fine-grained	65.23 (0.13)	65.25 (1.18)	63.66 (1.07)	63.15 (0.03)	65.82 (0.31)	65.12 (0.22)	66.69 (1.65)	**68.00 (0.32)**
	Intermediate	80.01 (0.27)	80.06 (0.55)	75.49 (0.83)	78.76 (0.40)	81.05 (0.40)	79.76 (0.20)	81.08 (0.80)	**82.66 (0.14)**
	Coarse-grained	89.45 (0.15)	90.62 (0.24)	88.44 (0.55)	89.93 (0.22)	90.89 (0.38)	89.93 (0.29)	91.08 (0.44)	**91.88 (0.10)**
	Macro	—	—	—	—	—	—	—	—
mF1 (%)	Fine-grained	32.57 (0.23)	32.36 (0.44)	31.40 (1.12)	31.73 (0.17)	33.48 (0.22)	31.63 (0.19)	33.12 (0.93)	**34.56 (0.32)**
	Intermediate	40.64 (0.38)	39.33 (0.09)	26.34 (1.66)	38.75 (0.45)	41.23 (0.33)	39.27 (0.18)	40.32 (1.05)	**42.97 (0.37)**
	Coarse-grained	52.89 (0.14)	52.11 (0.48)	48.02 (1.70)	51.42 (0.02)	54.94 (0.30)	51.79 (0.36)	53.57 (0.87)	**55.54 (0.15)**
	Macro	68.35 (0.18)	66.91 (0.43)	**70.56 (0.20)**	66.85 (0.40)	69.35 (0.58)	66.65 (0.54)	68.54 (0.50)	69.06 (0.16)
Params (M)		4.7	16.2	4.8	4.8	5.8	16.2	4.7	4.8
FLOP (T)		2.22	2.38	2.22	2.22	2.41	2.32	2.22	2.22
Inference time[a] (ms)		81.45	120.43	119.62	119.64	121.46	120.39	119.64	84.18
Peak VRAM[a] (MB)		1946.24	3404.32	3362.73	3362.14	3308.48	3349.73	3304.87	1958.99

[a] Results for a single forward pass over a batch of 256 patches on an NVIDIA L40S GPU. Inference time is averaged over 50 forward passes.

It is worth noting that NWPU-RESISC45 represents a highly saturated benchmark in this experimental setting, as the nonhierarchical baseline already achieves 96.12% OA and 96.11% m$F1$ at the fine-grained level. Consequently, large absolute gains are not expected. Nevertheless, SAHC achieves the best mean performance at all hierarchical levels in both OA and m$F1$, indicating that the HC improves the fine-grained prediction without degrading the intermediate and coarse-grained outputs. At the fine-grained level, SAHC improves the OA from 96.12% to 96.55% and the m$F1$ from 96.11% to 96.56%. Moreover, the standard deviation of SAHC at the fine-grained level is 0.08% for OA and 0.07% for m$F1$, showing that the improvement is stable across runs. These results suggest that, even when the flat baseline is already very strong, the proposed HC provides a consistent improvement while preserving computational efficiency.

B. Case II: MS SITS Semantic Segmentation

To further evaluate the robustness of the proposed framework, an extensive analysis was conducted on the highly complex ELU dataset. Unlike the previous scenario, this dataset consists of lower resolution MS images (10 m/pixel) mapped to a complex taxonomy of 74 classes distributed across four hierarchical levels. The results of this analysis are reported in Table II. The complexity of the ELU dataset, combined with the potential for semantic noise or inconsistencies in the CLC LU, makes this scenario particularly challenging. Standard hierarchical models that incorporate a rigid taxonomy might struggle to adapt to these inconsistencies. The proposed SAHC mitigates this limitation through consensus training and taxonomy-guided projectors that are adaptively tuned during optimization. This advantage can be observed in the OA and m$F1$ shown in the Table. Indeed, SAHC achieves the best performance at the fine-grained level with an OA of 38.38% and an m$F1$ of 34.56%. These results are also confirmed at the intermediate (OA of 50.20% and m$F1$ of 42.97%) and coarse-grained levels (OA of 60.21% and m$F1$ of 55.54%), consistently better performing with respect to other state-of-the-art methodologies like HiCervix and HCL. At the macro level, SAHC demonstrates competitive performance, ranking after only to the HCL and HAF methods by a marginal gap.

Although SAHC achieves the best fine-grained performance among the compared methods, the absolute OA at the finest level remains relatively low. This behavior is mainly due to the intrinsic complexity of the ELU dataset. Indeed, ELU contains 74 classes, making the Top-1 classification problem substantially more difficult than standard EO classification benchmarks. Then, the taxonomy represents LU classes, not only LC categories. Therefore, several fine-grained classes may be semantically distinct in the reference map while sharing similar spectral–temporal signatures at S2 resolution.

The Top-3 and Top-5 accuracy metrics evaluated at all the levels of the hierarchy further support this interpretation. These metrics indicate whether the ground-truth class is among the highest-confidence predictions of the model, measuring the

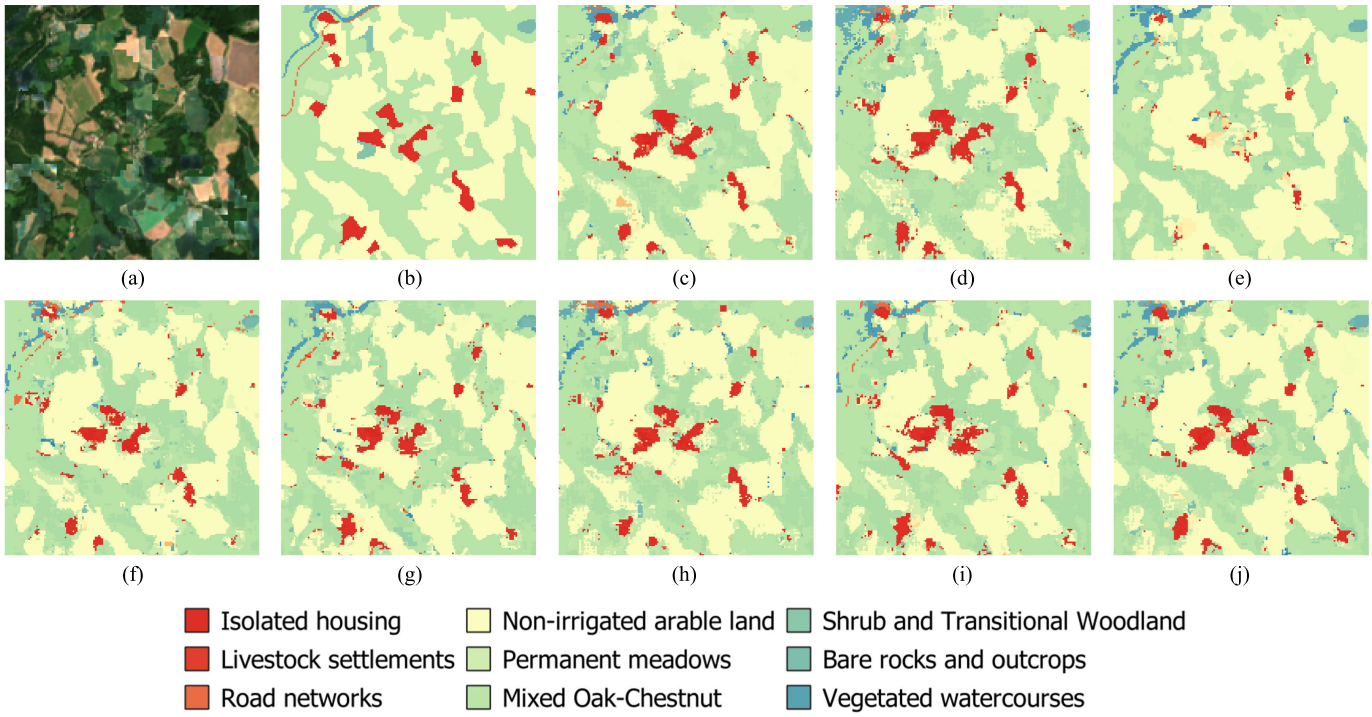


Fig. 5. Qualitative example of the predictions at the fine-grained level on the ELU dataset obtained by the considered hierarchical methodologies using the Swin Transformer backbone. (a) True color (RGB) monthly composite for August 2020. (b) CLC reference map. (c) Nonhierarchical baseline. (d) HRN method. (e) HAF method. (f) ALGS method. (g) HCL method. (h) MGM method. (i) HiCervix method. (j) Proposed SAHC method.

capacity of the network to maintain semantic proximity even when the absolute Top-1 prediction is incorrect. The proposed SAHC approach demonstrates the highest performance across almost all granularity levels. Specifically, SAHC achieves the highest Top-3 accuracies for the fine-grained, intermediate, and coarse-grained levels, reaching 59.33%, 73.71%, and 84.49%, respectively. This represents a large improvement against the HRN, which ranked second in the VHR case, showing a margin of 2.83% at the fine-grained level, while also outperforming the RS-oriented competitor HCL by 2.04%, 1.71%, and 1.49% in the fine, intermediate, and coarse levels, respectively. At the macro level, Top-3 performance saturates for all evaluated models. In this aggregated space, SAHC remains robust (97.15%), performing comparably to HCL (97.21%) and HiCervix (97.29%). Because the total number of target classes at the macro level is small, Top-3 classification becomes a less discriminative metric. For the Top-5 accuracy, SAHC maintains the overall best performance, outperforming the baseline and the compared methodologies across the evaluated granularities. At the fine-grained level, SAHC achieves 68.00%, surpassing HiCervix by 1.31%. This performance gap widens at the intermediate level, where SAHC reaches 82.66% compared with HiCervix at 81.08%. At the coarse-grained level, SAHC achieves 91.88%, outperforming HiCervix by 0.80%. Note that Top-5 metrics are omitted at the macro level, as the limited number of classes at this highest taxonomy level renders the Top-5 metric uninformative. The improvement observed in the Top-3 and Top-5 accuracies can be attributed to the design of the SAHC architecture. When the network struggles to isolate the exact fine-grained class, the HC mechanism, which promotes probabilistic alignment across different hierarchical levels, increases support for semantically related classes within the same hierarchical branch. This reduces the divergence between the projected predictions and the overall committee consensus.

Finally, consistent with the previous analysis, SAHC achieves the highest mean performance among the compared methods without sacrificing computational efficiency. While architectures like HRN and MGM impact computational efficiency, SAHC keeps the computational footprint nearly identical to the baseline model.

Finally, Fig. 5 provides a visual comparison of the evaluated hierarchical methodologies on a zoomed-in view portion of the ELU dataset. From the qualitative results, one can see that the SAHC approach retrieves the most spatially consistent LC maps, characterized by a reduction in classification artifacts compared with the competing approaches. This is particularly evident in the urbanized areas. Indeed, the proposed methodology appears to better characterize the "isolated housings," including dominant clusters and smaller patches, while still showing good agreement for the dominant agricultural and woodland classes. These advantages are further evident in Fig. 6. From a broader perspective, scaling an RS classification problem to an extensive taxonomy of 74 classes inherently introduces severe spectral confusion and extreme class imbalance. In this context, the observed improvements in the m$F1$ score are significant. Indeed, they demonstrate that the proposed SAHC mitigates the ambiguities typical of such a label space, maintaining robust and granular predictions where traditional approaches fail.

C. Sensitivity Analysis of the Consensus Formulation

The proposed SAHC includes two design choices for constructing and regularizing the HC: 1) the divergence measure

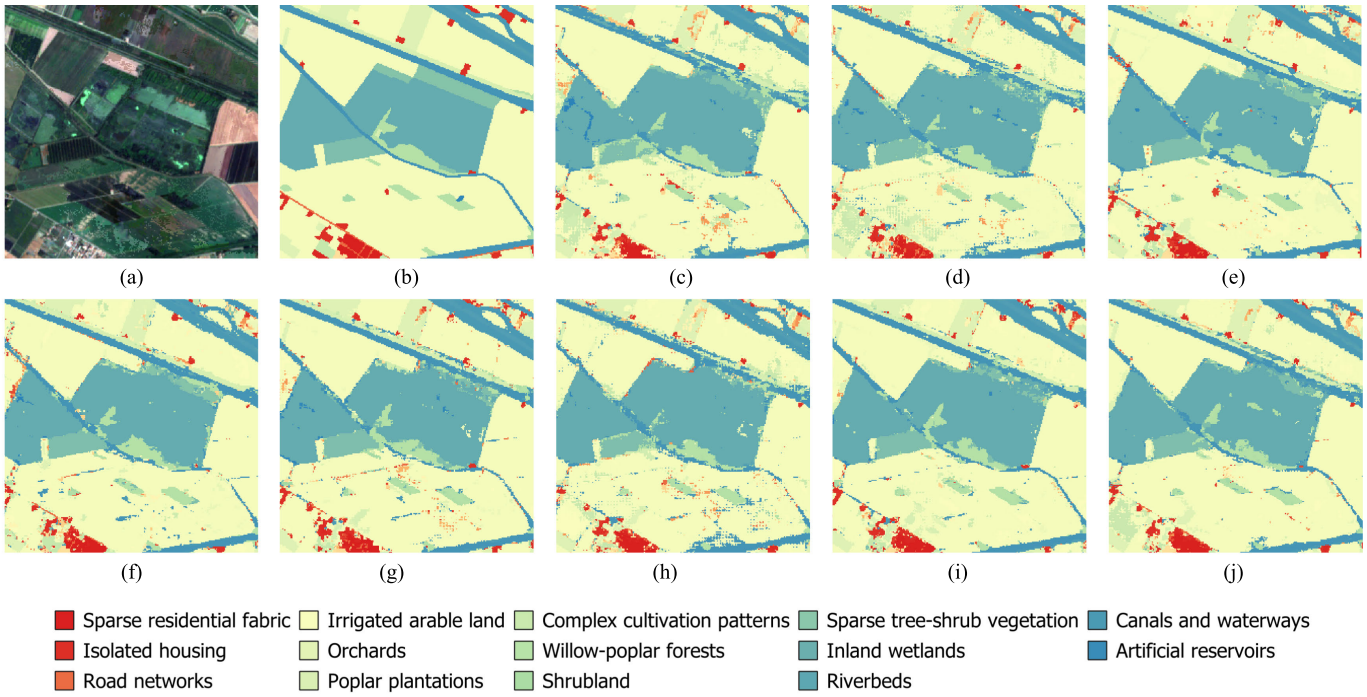


Fig. 6. Qualitative example of the predictions at the fine-grained level on the ELU dataset obtained by the considered hierarchical methodologies using the Swin Transformer backbone. (a) True color (RGB) monthly composite for August 2020. (b) CLC reference map. (c) Nonhierarchical baseline. (d) HRN method. (e) HAF method. (f) ALGS method. (g) HCL method. (h) MGM method. (i) HiCervix method. (j) Proposed SAHC method.

TABLE III

SENSITIVITY TO THE DIVERGENCE LOSS IN CONSENSUS. THE BEST RESULTS ARE REPORTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE RELATED STANDARD DEVIATION IS REPORTED IN PARENTHESES

	NWPU		ELU	
Divergence	OA (%)	mF1 (%)	OA (%)	mF1 (%)
MSE	96.14 (0.05)	96.14 (0.05)	**38.41 (0.39)**	34.54 (0.39)
KLD	96.05 (0.10)	96.05 (0.10)	38.08 (0.14)	34.29 (0.17)
Sym KLD	96.21 (0.17)	96.22 (0.17)	38.40 (0.30)	34.51 (0.22)
TV	96.25 (0.16)	96.24 (0.15)	38.21 (0.15)	34.29 (0.16)
JSD	**96.55 (0.08)**	**96.56 (0.07)**	38.38 (0.31)	**34.56 (0.32)**

used in the self-consistent loss and 2) the fusion operator used to combine direct and projected predictions. To assess the sensitivity of the method to these choices, we conduct two complementary ablation analyses.

1) Sensitivity to the Divergence Measure: The proposed loss is based on the JSD divergence, as the direct and projected predictions are jointly learned distributions, with no fixed reference distribution. We compare it with: 1) KLD; 2) symmetric KLD; 3) mean squared error (mse); and 4) total variation (TV) distance. Table III shows that JSD achieves the best results on NWPU-RESISC45, with 96.55% OA and 96.56% m$F1$. On ELU, the gap tends to shrink. Indeed, mse obtains the highest OA (38.41%), whereas JSD obtains the highest m$F1$ (34.56%), both with a small margin. Overall, the formulation is robust to the selected divergence. We retain JSD because it is symmetric, bounded, and remains among the best configurations on both datasets.

2) Sensitivity to the Fusion Operator: The proposed consensus uses a normalized geometric mean, which favors

TABLE IV

SENSITIVITY TO THE FUSION OPERATOR IN CONSENSUS. THE BEST RESULTS ARE REPORTED IN BOLD. THE RELATED STANDARD DEVIATION IS REPORTED IN PARENTHESES

	NWPU		ELU	
Fusion Operator	OA (%)	mF1 (%)	OA (%)	mF1 (%)
Arithmetic Mean	96.05 (0.10)	96.06 (0.08)	36.31 (0.53)	32.06 (0.75)
Geometric Mean	**96.55 (0.08)**	**96.56 (0.07)**	**38.38 (0.31)**	**34.56 (0.32)**

classes supported by all hierarchical levels. We compare it with arithmetic averaging while keeping all the other components unchanged. As shown in Table IV, geometric fusion outperforms arithmetic fusion on both datasets. On NWPU-RESISC45, it improves OA from 96.05% to 96.55%. On ELU, it improves OA from 36.31% to 38.38% and m$F1$ from 32.06% to 34.56%. Thus, the fusion operator has a larger effect than the specific divergence measure in the tested scenarios, supporting the usage of agreement-sensitive geometric fusion.

D. Sensitivity Analysis to Granularity Weights

To evaluate the robustness of the proposed hierarchical architecture, we conducted a sensitivity analysis on the VHR dataset to assess how model performance varies with the selection of different granularity weights λ^h. The search space was constrained such that the weights sum to one ($\sum_{h=1}^{H} \lambda^h = 1$), with each λ^h sampled using a step size of 0.1. Since the architecture adopts three hierarchical levels, the resulting parameter space can be visualized as a probability simplex that illustrates the various weight combinations across the different granularities. Fig. 7 illustrates a heatmap reporting the performance across all evaluated configurations of the three layer weights. As demonstrated in the figure, the optimal

TABLE V
ABLATION STUDY ON THE ELU AND NWPU-RESISC45 DATASETS. THE BEST RESULTS ARE REPORTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Dataset	Level	Metric	Baseline	+MH	+MH+$\hat{\mathbf{p}}_{HC}$	+MH+$\mathcal{L}_{HC}$	SAHC-fixed	SAHC-direct	**SAHC-proj**
NWPU	Fine	OA	96.12 (0.12)	96.26 (0.24)	96.30 (0.28)	96.39 (0.20)	96.12 (0.05)	**96.59 (0.10)**	96.55 (0.08)
		mF1	96.11 (0.11)	96.26 (0.24)	96.30 (0.28)	96.39 (0.21)	96.13 (0.04)	**96.59 (0.09)**	96.56 (0.07)
	Intermediate	OA	97.29 (0.11)	97.32 (0.14)	97.30 (0.16)	97.42 (0.20)	97.28 (0.12)	97.47 (0.08)	**97.48 (0.07)**
		mF1	96.59 (0.10)	96.25 (0.21)	96.48 (0.21)	96.65 (0.30)	96.49 (0.16)	96.72 (0.13)	**96.75 (0.16)**
	Coarse	OA	97.43 (0.08)	97.65 (0.33)	97.79 (0.16)	97.79 (0.26)	97.80 (0.10)	97.42 (0.69)	**97.90 (0.07)**
		mF1	97.07 (0.10)	97.35 (0.25)	97.43 (0.17)	97.42 (0.28)	97.45 (0.10)	97.54 (0.13)	**97.58 (0.09)**
ELU	Fine	OA	36.84 (0.37)	38.03 (0.50)	37.73 (0.41)	38.29 (0.18)	38.15 (0.25)	**38.39 (0.36)**	38.38 (0.31)
		mF1	32.57 (0.23)	33.89 (0.41)	33.57 (0.32)	33.96 (0.15)	34.35 (0.27)	**34.67 (0.34)**	34.56 (0.32)
	Intermediate	OA	47.84 (0.46)	48.64 (0.51)	49.02 (0.56)	48.93 (0.31)	49.79 (0.33)	49.13 (0.53)	**50.20 (0.13)**
		mF1	40.64 (0.38)	41.31 (0.41)	41.59 (0.54)	41.27 (0.33)	42.61 (0.47)	42.22 (0.23)	**42.97 (0.37)**
	Coarse	OA	57.64 (0.07)	58.63 (0.48)	59.29 (0.34)	58.92 (0.52)	59.98 (0.21)	58.97 (0.50)	**60.21 (0.21)**
		mF1	52.89 (0.14)	54.14 (0.36)	54.66 (0.21)	54.46 (0.30)	55.39 (0.06)	54.56 (0.55)	**55.54 (0.15)**
	Macro	OA	77.55 (0.04)	77.58 (0.08)	77.89 (0.18)	77.91 (0.19)	**78.46 (0.12)**	77.80 (0.16)	78.33 (0.17)
		mF1	68.35 (0.18)	68.29 (0.08)	68.72 (0.17)	68.10 (0.27)	**69.10 (0.43)**	68.33 (0.22)	69.06 (0.16)

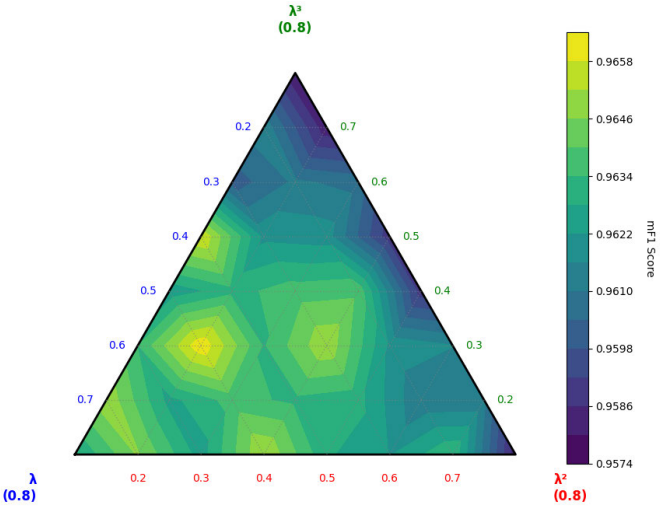


Fig. 7. Sensitivity analysis of the hierarchical weights λ^h on the VHR dataset. The performance surface is visualized as a probability simplex, where each point represents a weight combination for the three hierarchical levels sampled at a step size of 0.1. The simplex heatmap shows the optimal configuration as $\lambda^1 = 0.5$, $\lambda^2 = 0.2$, and $\lambda^3 = 0.3$.

combination is achieved at $\lambda^1 = 0.5$, $\lambda^2 = 0.2$, and $\lambda^3 = 0.3$. This optimal configuration justifies the hypothesis that priority should be allocated to both the fine-grained classification task and the coarsest class. Nevertheless, the analysis also reveals the presence of several local maxima across the performance surface, which are generally skewed toward configurations in favor of the finest-level weight. The best validation performance was obtained by assigning the largest weight to the finest level, a lower weight to the intermediate level, and a moderate weight to the coarse-grained level. The results indicate that the hierarchical levels should emphasize the most discriminative fine-grained objective, while increasing the contribution of the coarsest semantic ones. Intermediate levels are instead down-weighted, providing a better balance between fine-grained discrimination and global semantic regularization.

E. Ablation Study

The performance contribution of each component in the proposed hierarchical approach is analyzed in the ablation study on both datasets presented in Table V. In this study, we evaluate the ResNet50 backbone on NWPU-RESISC45 and the Swin Transformer backbone on ELU. The considered configurations include the following cases.

1) *Baseline:* The standard approach with no multilevel classifiers.
2) *Multihead (MH):* The baseline approach modified with additional classification heads for the coarse-grained and intermediate classes.
3) *Multilevel With HC Inference* $\hat{\mathbf{p}}_{HC}$*:* The multilevel approach with consensus applied solely during inference (considering frozen *user-defined* hierarchies).
4) *Multilevel With* $\mathcal{L}_{HC}$ *Training:* The multilevel approach trained with the self-consistent loss, excluding consensus-based classification during inference.
5) *SAHC With Fixed Projectors (SAHC-Fixed):* A taxonomy-guided SAHC in which the projectors are obtained by directionally normalizing the *user-defined* hierarchy indicator matrices and are kept frozen throughout training.
6) *SAHC With Direct-Head Inference (SAHC-Direct):* The adaptive-projector SAHC configuration evaluated using the direct hierarchy-specific heads at inference.
7) *SAHC With Adaptive Projectors (SAHC-Proj):* The adaptive-projector SAHC configuration, which jointly optimizes taxonomy-initialized hierarchy projections together with the remaining SAHC components, consensus inference, and the supervised consensus loss ($\mathcal{L}_{CCE}^{HC}$) at multiple levels.

Table V reports the contribution of the different modules composing SAHC. Adding MH supervision alone improves

the fine-level performance on both datasets, with a modest gain on NWPU-RESISC45 and a larger improvement on ELU. In particular, fine-level m$F1$ on ELU increases from 32.57% to 33.89%, indicating that multilevel supervision is beneficial in the deeper and more complex hierarchies. The consensus-training objectives further improve the direct-head model. The Baseline + MH + $\hat{\mathbf{p}}_{HC}$ case shows the effect of fixed taxonomy-based aggregation alone. On NWPU-RESISC45, its effect is limited. On ELU, fixed consensus slightly reduces fine-level performance relative to the MH model, while improving the intermediate, coarse, and macro-level results. This confirms that fixed aggregation can favor cross-level agreement. Compared with Baseline + MH + $\mathcal{L}_{HC}$ case, SAHC-direct increases fine-level m$F1$ by 0.20% on NWPU-RESISC45 and by 0.71% on ELU. Applying consensus inference after complete training leaves fine-level performance nearly unchanged, while improving the intermediate, coarse, and macro-level ELU results. Thus, consensus inference mainly benefits coherent multilevel outputs and coarser-level performance rather than the fine-level score alone. The preferred inference output is application-dependent. Indeed, $\hat{\mathbf{p}}^{h}_{\mathrm{HC}}$ is preferable when coherent predictions are required across multiple hierarchical levels or when high performances in coarser label predictions are favored, whereas $\hat{\mathbf{p}}^{h}$ can be retained when only the finest-level metric is the operational objective.

Finally, the comparison between SAHC-fixed and the adaptive-projector SAHC configuration (SAHC-proj) isolates the calibration provided by the trainable hierarchy projectors. Adaptive projectors improve fine-level OA and m$F1$ on NWPU-RESISC45 by approximately 0.43%. On ELU, they provide smaller gains at the fine, intermediate, and coarse levels, while macro-level performance remains comparable. Overall, the ablation indicates that the main benefit of SAHC arises from multilevel supervision and self-consistent training. Consensus inference further supports coherent predictions across hierarchical levels, whereas adaptive projectors provide a modest, level-dependent calibration beyond frozen taxonomy-derived projectors.

F. Sensitivity to Projectors Initialization

The hierarchy parameters in (6) are initialized in the unconstrained logit space from the binary indicator matrices. We evaluate different values of Δ, which controls the logit gap between valid and invalid hierarchical connections. The case $\Delta = 0$ corresponds to a uniform initialization, whereas larger values of Δ impose a stronger taxonomy prior while preserving the trainability of all projectors.

As reported in Table VI, the effect of Δ is limited and varies across datasets and hierarchical levels. On NWPU-RESISC45, larger values provide slightly better average results, whereas on ELU smaller values are marginally preferable at the finest level and larger values are marginally preferable at some coarser levels. This trend may suggest that stronger priors are more effective where semantic aggregation is more stable, while finer grained classes may benefit from a more flexible calibration of cross-level relations. However, these differences are negligible, as suggested by the observed standard deviations over three runs. Overall, the results indicate that SAHC is only moderately sensitive to Δ, and we retain $\Delta = 5$ as the common configuration adopted in the main experiments on both datasets.

TABLE VI
SENSITIVITY TO THE PROJECTOR'S INITIALIZATION. THE BEST RESULTS ARE REPORTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE RELATED STANDARD DEVIATION IS REPORTED IN PARENTHESES

Dataset	Level	Metric	$\Delta = 0$	$\Delta = 1$	$\Delta = 3$	$\Delta = 5$
NWPU	Fine	OA	96.24 (0.05)	96.17 (0.04)	96.34 (0.11)	**96.55 (0.08)**
		mF1	96.24 (0.05)	96.18 (0.04)	96.35 (0.10)	**96.56 (0.07)**
	Inter	OA	97.33 (0.02)	97.21 (0.06)	97.44 (0.05)	**97.48 (0.07)**
		mF1	96.55 (0.02)	96.49 (0.07)	96.70 (0.07)	**96.75 (0.16)**
	Coarse	OA	97.80 (0.02)	97.67 (0.11)	97.85 (0.15)	**97.90 (0.07)**
		mF1	97.45 (0.02)	97.35 (0.15)	97.51 (0.16)	**97.58 (0.09)**
ELU	Fine	OA	**38.76 (0.17)**	38.54 (0.36)	38.48 (0.31)	38.38 (0.31)
		mF1	**34.96 (0.15)**	34.64 (0.43)	34.56 (0.28)	34.56 (0.32)
	Inter	OA	49.72 (0.28)	50.02 (0.41)	50.11 (0.31)	**50.20 (0.13)**
		mF1	42.61 (0.21)	42.81 (0.56)	42.82 (0.47)	**42.97 (0.37)**
	Coarse	OA	59.76 (0.06)	60.08 (0.19)	60.20 (0.23)	**60.21 (0.21)**
		mF1	55.47 (0.12)	**55.59 (0.17)**	55.55 (0.25)	55.54 (0.15)
	Macro	OA	78.39 (0.13)	78.28 (0.11)	**78.43 (0.14)**	78.33 (0.17)
		mF1	69.04 (0.29)	69.02 (0.22)	**69.11 (0.26)**	69.06 (0.16)

To further assess the stability of the learned hierarchy matrices, we compared the final projector matrices after training with different random seeds. With the default strong-prior initialization (i.e., $\Delta = 5$ and $s = 0.01$), the average cosine similarity and Pearson correlation between learned matrices were both 99.81%. When increasing the initialization noise to $s = 0.2$, the corresponding similarities remained high, namely, 97.52% and 97.45%, respectively. These results indicate that the learned semantic relationships are stable across training runs and are not artifacts of a particular random initialization.

G. Hierarchy Projection Matrices Analysis

To characterize the cross-level relations encoded by the adaptive projectors, we analyze the estimated joint probability matrix $\mathbf{J}^{h_2,h_1}$ representing adaptive relationships between the intermediate- and fine-grained levels for the ELU dataset. The corresponding visualization on the NWPU dataset is omitted, since the learned projectors do not reveal qualitatively new off-taxonomy structures. Fig. 8 presents a heatmap of this joint probability matrix between intermediate- and fine-level classes, illustrating the hierarchical structure captured by the proposed SAHC methodology on the ELU dataset. The visualization confirms that the adaptive projectors remain anchored to the *user-defined* semantic aggregations between the levels. Specifically, the cyan rectangles in the heatmap correspond to the *user-defined* expected mappings, from fine-grained classes to intermediate classes. The higher weights within these regions confirm that the *data-driven* estimated mappings largely preserve the intended *user-defined* taxonomy. In addition to these rigid hierarchies, the network also captures secondary cross-level relationships. Intuitively, one can expect that these secondary cross-level relationships correspond to relationships related to coarser hierarchical levels. To help visualize this, Fig. 8 shows with blue rectangles

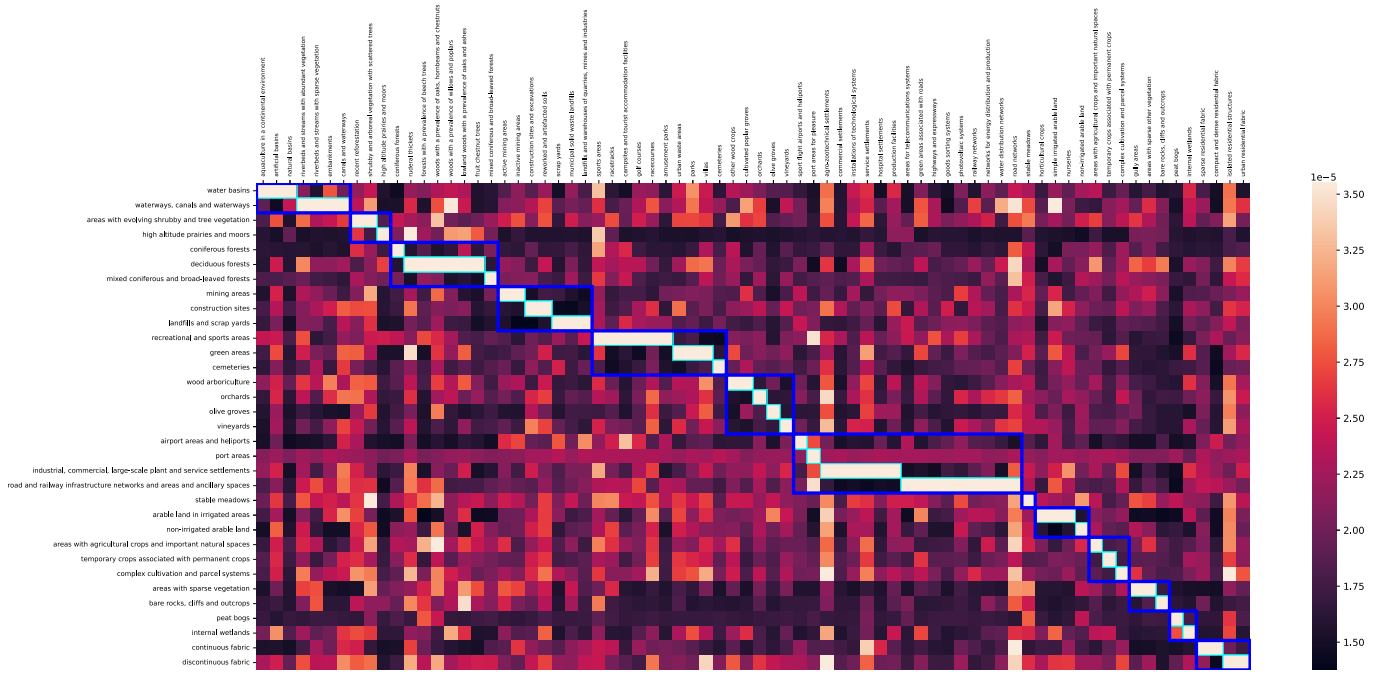


Fig. 8. Heatmap of the estimated joint probability matrix $\mathbf{J}^{h_2,h_1}$ considering the Swin Transformer backbone on the ELU dataset from intermediate classes to fine-grained classes. The cyan rectangles identify the expected hierarchical aggregation of the fine-grained classes at the intermediate level.

how the classes at the fine level aggregate at the coarse level. However, because separate joint matrices are learned for different hierarchical transitions in the proposed architecture, the model is able to assign secondary relationship weights that largely ignore the initial *user-defined* hierarchy. For instance, the adaptive weights indicate active associations between "areas with agricultural crops and important natural spaces" and "woods with a prevalence of oaks, hornbeams, and chestnuts," as well as between "continuous fabric" and "road networks." These weights provide a qualitative example of recurrent cross-level ambiguity in the dataset, highlighting the importance of a *data-driven* calibration. Nonetheless, some associations with relatively high weights show semantically wrong relationships, such as "high altitude prairies and moors" and "sport areas" or "airports" and "campsites," which could be related to both spectral similarity and semantic ambiguity.

Finally, we examined the dominant cross-level weights within the adaptive hierarchy projectors to interpret the task-dependent associations used by the projectors. For the adjacent forward projectors, we ranked the learned weights corresponding to the hierarchical transitions and extracted the five most significant relations, reported in Table VII. At finer levels, several dominant weights correspond to plausible spatial co-occurrences, such as "road networks" with "continuous fabric" or "waterways and canals" with "simple irrigated arable land." Toward the coarsest levels, pairs such as "urbanized areas" and "agricultural territories" are more likely attributable to spectral similarities, comparable structures, or classifier confusions rather than newly discovered semantic hierarchy links.

H. Discussion

The analysis of the results reveals that the proposed methodology is able to exploit both complex and simple hierarchies. Indeed, SAHC achieves the most accurate overall results, obtaining the best fine-, intermediate-, and coarse-level performance on ELU and the best results across all levels on NWPU-RESISC45, while remaining competitive at the ELU macro level. Furthermore, the self-consistency loss effectively constrained the hierarchical levels to converge toward a unified prediction, promoting cross-level agreement in the final classification maps. The experiments demonstrate that SAHC provides the best overall tradeoff between fine-grained discrimination and coherent multilevel prediction. The ablation further shows that the learned self-consistency is the principal contributor to fine-level performance, whereas consensus inference primarily improves intermediate, coarse, and macro-level predictions. Therefore, direct-head predictions may be preferable when only the finest-level discriminative metric is operationally relevant, while consensus outputs are more appropriate when the downstream application requires coherent predictions across the hierarchy or better performance at coarser levels.

The proposed SAHC formulation is lightweight for the three- and four- level taxonomies considered in this study. Its dense cross-level parameterization nevertheless scales quadratically with hierarchy depth and with the products of the involved label cardinalities. Taxonomies substantially deeper than those evaluated here may therefore benefit from adjacent-level, sparse, or selectively connected consensus variants.

A limitation of the proposed method is its reliance on a predefined hierarchy to initialize and train the HC. While the hierarchy matrices are trainable and allow *data-driven* calibration of the projection weights, the present study does not explicitly evaluate robustness to intentionally corrupted, incomplete, or partially inaccurate hierarchy definitions. If the predefined hierarchy poorly reflects the semantic relationships present in the data, the initial hierarchy prior and the

TABLE VII

TOP-5 CROSS-LEVEL SEMANTIC TRANSITIONS EXTRACTED FROM THE LEARNED FORWARD HIERARCHY PROJECTORS. THE TABLE REPORTS THE ASSOCIATIONS WITH THE HIGHEST TRANSITION WEIGHTS MAPPED FROM THE COARSER HIERARCHY LEVEL h TO THE FINER LEVEL $\hat{h}$

$\mathbf{J}^{h,\hat{h}}$	h	$\hat{h}$
$\mathbf{J}^{h_2,h_1}$	continuous fabric	road networks
	discontinuous fabric	agro-zootechnical settlements
	high altitude prairies and moors	ruderal thickets
	areas with agricultural crops and important natural spaces	woods with a prevalence of oaks, hornbeams and chestnuts
	waterways, and canals	simple irrigated arable land
$\mathbf{J}^{h_3,h_2}$	urbanized areas	road and railway infrastructure networks and areas and ancillary spaces
	urbanized areas	industrial, commercial, large-scale plant and service settlements
	heterogeneous agricultural areas	arable land in irrigated areas
	maritime wetlands	arable land in irrigated areas
	non-irrigated arable land	stable meadows
$\mathbf{J}^{h_4,h_3}$	artificially modeled territories	arable land
	artificially modeled territories	heterogeneous agricultural areas
	agricultural territories	environments with evolving shrubby and/or herbaceous vegetation
	wooded territories and semi-natural environments	heterogeneous agricultural areas
	urbanized areas	agricultural territories

supervised losses at the coarser levels may bias the learned projection matrices toward suboptimal interlevel compatibility. Furthermore, this constraint bounds transferability to geographically distinct regions or different LC distributions with misaligned or structurally heterogeneous taxonomies.

VI. CONCLUSION

In scene classification and semantic segmentation RS tasks, the hierarchical relationship among the target classes is often under-exploited, resulting in the exclusion of valuable semantic information from the decision-making process. To address this limitation, this article proposes the SAHC framework, which combines hierarchy-specific classification branches, cross-level projections, geometric consensus, self-consistent training, supervised consensus learning, and optional consensus inference. Taxonomy-derived projectors provide the core of the method, while trainable hierarchy log-joint matrices offer adaptive tuning for global cross-level relationships around the *user-defined* semantic prior. The proposed methodology is lightweight and scalable, making it adaptable to different architectures and datasets organized in a hierarchical manner. The experimental results demonstrated the effectiveness of the proposed approach on two different RS datasets, related to the classification of VHR and MS images. Indeed, the proposed approach is particularly relevant in VHR image and MS SITS classification, where the semantic structure of the data and available datasets support an explicit hierarchical representation of the task.

Future work will investigate robustness under hierarchy perturbations, assessing the ability of the adaptive projectors to compensate for incomplete or partially inaccurate taxonomy definitions. Moreover, adaptation across wide geographic areas will be explored under misaligned or structurally heterogeneous taxonomies, evaluating the transferability and recalibration of the learned cross-level relationships.

REFERENCES

[1] C. Small, "Grand challenges in remote sensing image analysis and classification," *Frontiers Remote Sens.*, vol. 1, Apr. 2021.

[2] Y. Wang, Y. Sun, X. Cao, Y. Wang, W. Zhang, and X. Cheng, "A review of regional and global scale land use/land cover (LULC) mapping products generated from satellite remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 206, pp. 311–334, Dec. 2023.

[3] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.

[4] E. Sertel, R. Topaloğlu, B. Şallı, I. Y. Algan, and G. Aksu, "Comparison of landscape metrics for three different level land cover/land use maps," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, p. 408, Oct. 2018.

[5] A. Pérez-Hoyos, F. J. García-Haro, and N. Valcárcel, "Incorporating sub-dominant classes in the accuracy assessment of large-area land cover products: Application to GlobCover, MODISLC, GLC2000 and CORINE in Spain," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 187–205, Jan. 2014.

[6] J. Chen, J. Huang, P. Wang, F. Xiong, Y. Qian, and L. Xiao, "Adaptive label granularity selection for remote sensing targets: Balancing accuracy and specificity," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4706817.

[7] J. Chen, F. Xiong, Y. Qian, and L. Xiao, "Hierarchical contrastive learning for multigranularity ship classification with learnable class queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5622916.

[8] A. Bilal, A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Do convolutional neural networks learn class hierarchy?" *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 152–162, Jan. 2018.

[9] R. Cerri, R. C. Barros, and A. C. P. L. F. de Carvalho, "Hierarchical multi-label classification using local neural networks," *J. Comput. Syst. Sci.*, vol. 80, no. 1, pp. 39–56, Feb. 2014.

[10] J. Fan et al., "HD-MTL: Hierarchical deep multi-task learning for large-scale visual recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1923–1938, Apr. 2017.

[11] E. Giunchiglia and T. Lukasiewicz, "Coherent hierarchical multi-label classification networks," in *Proc. NIPS*, vol. 33, 2020, pp. 9662–9673.

[12] L. Bertinetto, R. Mueller, K. Tertikas, S. Samangooei, and N. A. Lord, "Making better mistakes: Leveraging class hierarchies with deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12503–12512.

[13] A. Garg, D. Sani, and S. Anand, "Learning hierarchy aware features for reducing mistake severity," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 252–267.

[14] J. Deng et al., "Large-scale object classification using label relation graphs," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 48–64.

[15] J. Chen, P. Wang, J. Liu, and Y. Qian, "Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4848–4857.

[16] R. Zhou, J. Wei, Q. Zhang, R. Qi, X. Yang, and C. Li, "Multi-granularity archaeological dating of Chinese bronze dings based on a knowledge-guided relation graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 3103–3113.

[17] D. Cai et al., "HiCervix: An extensive hierarchical dataset and benchmark for cervical cytology classification," *IEEE Trans. Med. Imag.*, vol. 43, no. 12, pp. 4344–4355, Dec. 2024.

[18] S. Illarionova, A. Trekin, V. Ignatiev, and I. Oseledets, "Neural-based hierarchical approach for detailed dominant forest species classification by multispectral satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1810–1820, 2021.
[19] D. Ç. Demirkan, A. Koz, and H. Ş. Düzgün, "Hierarchical classification of Sentinel 2-a images for land use and land cover mapping and its use for the CORINE system," *J. Appl. Remote Sens.*, vol. 14, no. 2, Jun. 2020, Art. no. 026524.
[20] A. Waśniewski, A. Hościło, and M. Chmielewska, "Can a hierarchical classification of Sentinel-2 data improve land cover mapping?" *Remote Sens.*, vol. 14, no. 4, p. 989, Feb. 2022.
[21] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your 'Flamingo' is my 'bird': Fine-grained, or not," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11476–11485.
[22] B. Settles, *Active Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning)*, R. J. Brachman, W. W. Cohen, and T. Dietterich, Eds., Cham, Switzerland: Springer, Aug. 2012.
[23] G. Weikmann, G. Perantoni, and L. Bruzzone, "A class-driven hierarchical ResNet for classification of multispectral remote sensing images," *Proc. SPIE*, vol. 12733, Sep. 2023, Art. no. 127330D.
[24] M. J. Rufo, J. Martín, and C. J. Pérez, "Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach," *Bayesian Anal.*, vol. 7, no. 2, pp. 411–438, Jun. 2012.
[25] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[26] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.
[27] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "Training general representations for remote sensing using in-domain knowledge," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Waikoloa, HI, USA, Sep. 2020, pp. 6730–6733.
[28] European Union, Copernicus Land Monitoring Service (CLMS). (2024). Sentinel-2 Global Mosaic (S2GM) Service. *European Environment Agency (EEA)*. Accessed: Feb. 27, 2026. [Online]. Available: https://s2gm.land.copernicus.eu/
[29] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[30] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-time spatial temporal transformer for space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17441–17451.
[31] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. 34th Conf. Uncertainty Artif. Intell.*, vol. 2, Monterey, CA, USA, 2018, pp. 876–885.



Giulio Weikmann (Member, IEEE) received the M.S. degree (summa cum laude) in information and communication engineering and the Ph.D. degree (cum laude) in information engineering and computer science from the University of Trento, Trento, Italy, in 2020 and 2024, respectively.

He is currently an Assistant Professor with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. He has participated in several national and international research projects and coordinated his research team's contributions to European Commission H2020 Projects ExtremeEarth and AI4Copernicus and Italian Space Agency (ASI) Projects PrismaLearn and Space It Up! His research interests include deep learning methods for the automatic processing and classification of remotely sensed optical data, with particular emphasis on multitemporal data and time-series characterization and analysis.

Dr. Weikmann was a recipient of the 17th Eugenio Zilioli Award for the Best Master's Thesis from the Associazione Italiana Telerilevamento and the Consiglio Nazionale delle Ricerche and the 2023 Best Student Paper Award from the SPIE Image and Signal Processing for Remote Sensing XXIX Conference. He has served as a reviewer for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.



Gianmarco Perantoni (Member, IEEE) received the M.S. degree (summa cum laude) in information and communication engineering and the Ph.D. degree (cum laude) in information engineering and computer science from the University of Trento, Trento, Italy, in 2020 and 2025, respectively.

He is currently a Post-Doctoral Researcher and a Teaching Assistant with the Remote Sensing Laboratory (https://rslab.disi.unitn.it/), Department of Information Engineering and Computer Science, University of Trento. His current work supports the High Resolution Land Cover Project in the framework of the Climate Change Initiative of European Space Agency (ESA). His research interests include deep learning for the automatic processing and classification of remote sensed optical data, with a specific focus on land cover mapping using weak supervision.

Dr. Perantoni was a recipient of the 2021 IEEE GRS29-CNI-SI Prize for the three best Italian master's theses in geoscience and remote sensing. He serves as a reviewer for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the IEEE International Geoscience and Remote Sensing Symposium (IGARSS).



Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree (summa cum laude) in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory (https://rslab.disi.unitn.it/), Department of Information Engineering and Computer Science, University of Trento. He is a Principal Investigator of many research projects. Among the others, he is currently a Principal Investigator of the Radar for Icy Moon Exploration (RIME) and the Subsurface Radar Sounder (SRS) instruments in the framework of the JUICE and the Envision missions, respectively, of European Space Agency (ESA) and is the Science Lead of the High Resolution Land Cover Project in the framework of the Climate Change Initiative of ESA. He is the author (or a co-author) of more than 450 scientific publications in refereed international journals, more than 410 papers in conference proceedings, and 25 book chapters. He is the editor/co-editor of 18 books/conference proceedings and two scientific books. His papers are highly cited, as proven from the total number of citations (more than 65 000) and the value of the H-index (119) (source: Google Scholar). He was invited as a keynote speaker at more than 40 international conferences and workshops. His research interests are in the areas of remote sensing, radar and synthetic aperture radar (SAR), signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects related to Earth observation and planetary exploration.

Dr. Bruzzone was a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) from 2009 to 2023, where he has been the Vice-President for Professional Activities since 2019. He ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, WA, USA, July 1998. Since 1998, he has been a recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, the 2019 WHISPER Outstanding Paper Award, and the 2022 Letter Prize Paper Award for the best paper published in IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was a Guest Co-Editor of many special issues of international journals. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the Founder of *IEEE Geoscience and Remote Sensing Magazine*, for which he was the Editor-in-Chief from 2013 to 2017. He was a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society from 2012 to 2016.

Open Access funding provided by 'Universita degli Studi di Trento' within the CRUI CARE Agreement