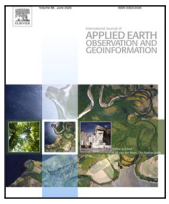
International Journal of Applied Earth Observation and Geoinformation 139 (2025) 104471

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Spatio-temporal-text fusion for hierarchical multi-label crop classification based on time-series remote sensing imagery

Xiyao Li [a], Jiayi Li [a,*], Jie Jiang [b,*], Xiaofeng Pan [c], Xin Huang [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, PR China
[b] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and architecture, Beijing, 100044, PR China
[c] Shezhen Ecological and Environmental Monitoring Center of Guangdong Province, Shenzhen, 518049, PR China

ARTICLE INFO

Keywords:
Deep learning
Crop classification
Hierarchical classification
Satellite image time series
Class semantic

ABSTRACT

Recent advances in deep learning have enhanced crop classification, yet current research still underutilizes the hierarchical information of crop types, limiting classification accuracy. As categories are subdivided, the sample imbalance intensifies, posing a challenge to fine classification of crops. To address this, we propose the Class **Sem**antic Guided **Hi**erarchical Segmentation Framework (SemHi framework) for satellite image time series (SITS) crop classification. This framework effectively leverages the hierarchical information in the class system and outputs classification results at each level in an end-to-end manner. The SemHi framework comprises four modules: (1) The backbone that learns the spatio-temporal features; (2) The label embedding module, a core component that guides the feature learning through the fusion of hierarchical structure and textual representations; (3) The prototype distance measurement module to enhance class separation and reduce within-class variation; (4) The hierarchical logic regularization module, which enables multi-granularity crop predictions and strengthens the hierarchical logic between them. The validation results on the public crop classification dataset show that the SemHi framework with multi-dimensional fusion of spatio-temporal-text information significantly improves the performance of all state-of-the-art (SOTA) networks on fine-grained classes, with an overall accuracy improvement of 0.48% to 13.86%. Furthermore, experiments on a remote sensing classification dataset demonstrate the framework's generality and potential for broader applications in remote sensing tasks.

1. Introduction

The world's growing population, rapid urban expansion, and climate change present challenges to agricultural and uncultivated lands. Crop mapping is helpful in making agricultural decisions to promote the harmonious development of local agricultural economies and ecologies (Futerman et al., 2023; Li et al., 2024).

Remote sensing has unique advantages such as fast revisit, low cost and the consistency and comparability of the generated crop distribution maps, which has become a popular and effective means of crop classification and mapping (Bueno et al., 2023). From a methodological perspective, remote sensing-based crop classification includes traditional machine learning and deep learning (DL) approaches. Traditional models, such as Random Forest (RF) and Support Vector Machine (SVM), rely on handcrafted spectral indices (e.g., NDVI, NDWI) to extract features from time-series imagery (Hu et al., 2021; Hao et al., 2015; Shang et al., 2015), but face challenges in complex farmland environments, diverse study areas and crop types due to the lack of spatial information. In contrast, DL models automatically extract multi-scale and multi-modal features, reducing reliance on handcrafted features while achieving strong generalization capabilities on large-scale remote sensing datasets (Zheng et al., 2024). Currently, deep learning methods in crop classification mainly fall into two categories: methods that combine temporal and spectral characteristics, and methods that comprehensively consider spatial–temporal-spectral characteristics. The first category utilizes the temporal and spectral features of single pixels to map them into the semantic space of crop classes. The second category uses satellite image time series (SITS) to learn features, further integrating spatial information to achieve better crop classification. Representative state-of-the-art (SOTA) works include convolutional neural networks, such as UNet3D and UNet3Df (Tarasiou et al., 2022), convolutional recurrent neural networks, such as ConvSTAR (Turkoglu et al., 2021) and ConvGRU (Rußwurm and Körner, 2018), ViT-based networks (Dosovitskiy et al., 2020), such as Temporo-Spatial Vision Transformer (TSViT) (Tarasiou et al., 2023) and Swin UNEt Transformer(SwinUNETR) (Tang et al., 2022). Compared to the first category, the second utilizes the fusion of spatial, temporal, and spectral features to

* Corresponding authors.
E-mail addresses: zjjerica@whu.edu.cn (J. Li), jiangjie@bucea.edu.cn (J. Jiang).

https://doi.org/10.1016/j.jag.2025.104471
Received 18 November 2024; Received in revised form 15 February 2025; Accepted 6 March 2025
Available online 22 March 2025

capture a more comprehensive understanding of crop growth dynamics. Overall, although deep learning is widely used in crop classification, these methods fall short in incorporating the fusion of textual semantics and hierarchical structures within the crop classification system. This limits their capacity to capture nuanced crop relationships and prevents them from achieving a top-down understanding of the semantics and hierarchy of crop species.

Hierarchical Multi-Label Classification (HMC) is an effective learning paradigm that embeds priori knowledge of class hierarchy and can interpret multi-level semantics in an integrated manner. Recently, a number of studies (Wang et al., 2022; Sulla-Menashe et al., 2019; Kang et al., 2024) have shown that the utilization of hierarchical information, namely Hierarchical Multi-Label Image Classification (HMIC), is of great help in improving image classification. According to the utilization of hierarchical information, HMIC methods based on deep learning usually adopt two approaches, designing new network structures or new loss functions. The first approach attempts to directly embed class hierarchy information in the network structure, usually using multiple branches to output all predictions from coarse to fine (Li et al., 2020a). However, this approach is relatively complex, causing the training process to be less flexible, and classification errors at high levels can propagate to the lower levels (Sinha et al., 2018). The other approach embeds hierarchical constraints of classes into the optimization objective by constructing a specific loss function (Chen and Qian, 2022), in order to obtain prediction results at different levels. These methods avoid the extensive experimentation and tuning required to build complex multi-branch networks. Based on this, we adopt the second approach, i.e., optimizing the global training logic by designing multiple loss functions to ensure that the prediction results match the class hierarchy.

Existing HMIC studies in remote sensing heavily rely on rule-based designs and prior knowledge, primarily using traditional machine learning methods such as support vector machines (SVM) and decision trees, etc (Jiao et al., 2019). In contrast, HMIC approaches with deep learning are limited due to the high computational costs of segmentation tasks and the complexity of designing pixel-level hierarchical loss functions. Additionally, existing remote sensing image interpretation often embeds semantic hierarchy from the perspective of images, lacking exploration of semantic hierarchy structures in class systems (i.e. natural language). Moreover, due to the limitation of sample distribution along the hierarchy, the number of samples at the fine level is much less than that at the coarse level (Chen and Qian, 2022). Given the challenges in fine-grained crop classification, especially with sample imbalance and complex class hierarchies, we propose a framework that leverages spatio-temporal-text fusion. This fusion combines spatial patterns, temporal dynamics, and textual knowledge of crop classes, providing a more robust approach to capturing detailed semantic relationships. We also introduce metric-based prototype learning (Goel et al., 2019) to correct errors caused by large intra-class variance and small inter-class differences in features.

In summary, conventional DL methods overlook the hierarchical structure of classification systems, while DL methods for HMIC are not well-suited for segmentation tasks and do not incorporate textual semantics. Therefore, we propose an innovative framework integrating spatio-temporal-text information with feature enhancement and multi-level output. Six SOTA networks, including SwinUNETR, TSViT, ConvGRU, ConvSTAR, UNet3Df and UNet3D, are used as backbone models. The main contributions of this work are as follows:

(1) The Class **Sem**antic Guided **Hi**erarchical Segmentation Framework (SemHi framework) is proposed, integrating and utilizing crops' spatio-temporal information and class hierarchy. This framework adapts to various semantic segmentation networks, providing crop predictions at all granularity levels for each pixel in accordance with their logical relationships.

(2) Considering the hierarchical characteristics of crop classes, we propose the label embedding module. This module encodes the hierarchical structure and text of crop classes to obtain hierarchy-aware label representation. Then this representation is brought closer to the deep features derived from encoder, to facilitate the class semantic and structure guidance for the network.

(3) To improve the separability of crops in fine-grained classes, we propose the prototype distance measurement module to measure the distance between prototypes of deep features extracted by network decoder and make features more separable after each iteration through loss functions.

2. Methodology

We aim to adapt different networks suitable for HMIC tasks and use hierarchical structures to generate features and predictions that conform to logical relationships, thereby improving the segmentation performance. Based on this goal, we propose the SemHi framework for crop classification. In this section, a detailed introduction to the composition of proposed framework is provided (Fig. 1).

The proposed SemHi framework consists of the semantic segmentation backbone network (left column of Fig. 1) for fusing spatial and temporal features, the label embedding module (A), the prototype distance measurement module (B), and the hierarchical logic regularization module (C). During training, the input includes images, class hierarchy, and labels (I, II, III in Fig. 1), but in inference, only images and class hierarchy are needed to output multi-level classification results.

2.1. Backbone

The SemHi framework can adapt to various DL semantic segmentation models. Given Transformer's adaptability to large-scale data, we select SwinUNETR as the backbone. As shown in Fig. 2, we retain three of the four original SwinUNETR stages due to the small input size, outputting the encoder deepest feature S, decoder deepest feature F, and the classification probability map P. The SwinUNETR encoder utilizes (S)W-MSA (Liu et al., 2021) to compute time-space-spectral self-attention, capturing long-range dependencies. Given an input image of $C \times T \times H \times W$, where C, T, H and W represent the number of bands, time sequence, height and width of image, respectively, the encoder extracts hierarchical features with the resolution of $C_m \times T/2^m \times H/2^m \times W/2^m$, where $m \in \{1,2,3\}$. C_1, C_2 and C_3 are 96, 192, 384, respectively. These include high-resolution shallow features and low-resolution deep features, enhancing hierarchical crop representation and classification performance (Zhao et al., 2017). The SwinUNETR decoder, based on the U-Net skip connection structure (Ronneberger et al., 2015), fuses shallow features via deconvolution, reconstructing F while preserving the original input spatial size $C' \times T \times H \times W$, where $C' = C_1 + C$.

2.2. Label embedding module

In the HMIC tasks, it is necessary to abstract and integrate concepts of different levels into the network, i.e. hierarchical encoding (Ma et al., 2022). Hierarchical encoding typically organizes data samples (such as words or images) into a high-dimensional space, where feature distances represent their semantic similarity (Nickel and Kiela, 2017). Usually, natural language has a high degree of summarization and conciseness, with the most intuitive semantic information (Kowsari et al., 2017). For instance, in a system consisting of coarse-grained classes like *grains* and *fruits*, as well as fine-grained classes like *rice* and *apple*, people can realize that the correlation between *grains* and *rice* is stronger, while the correlation between *grains* and *fruits* or *apple* is weaker. From this, we refer that encoding the class name and hierarchy structure could guide crop classification. Therefore, we embed the class text information and hierarchical structure information on the labels into the process of feature extraction by the backbone network to assist in classification. Our label embedding module is shown in Fig. 3.

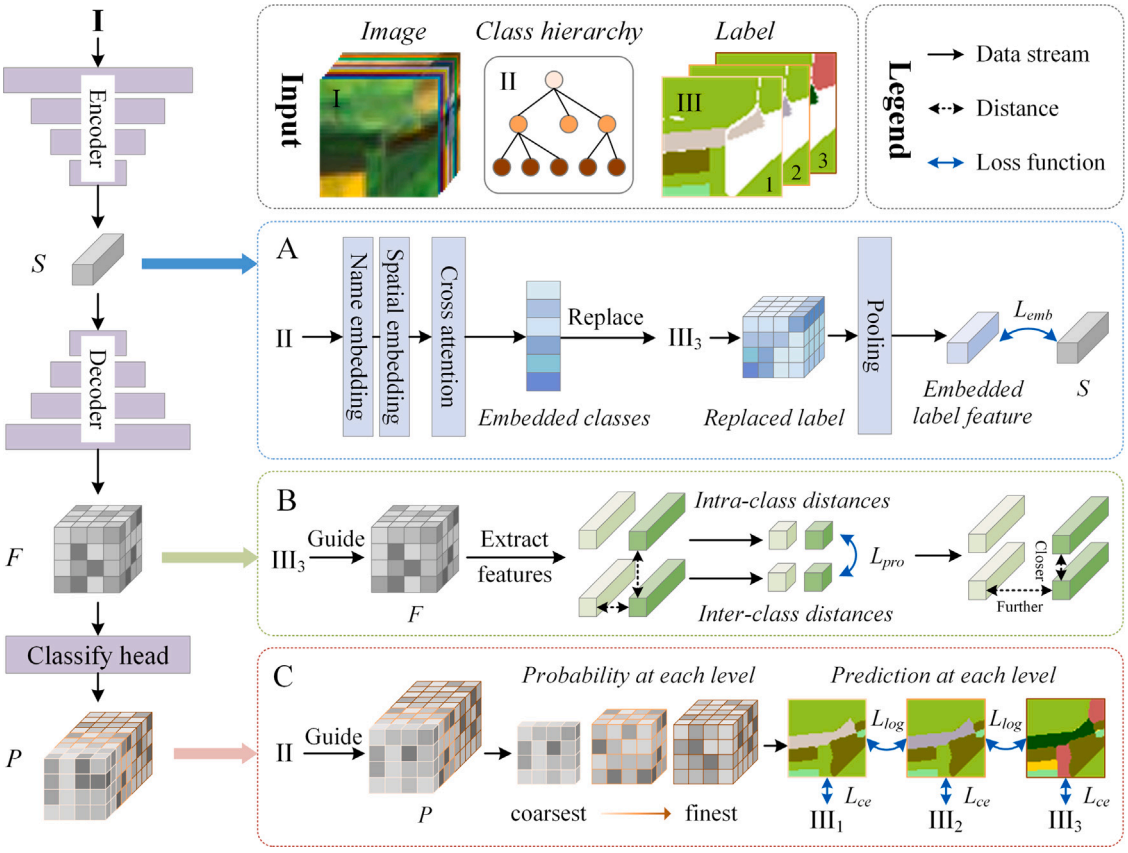


Fig. 1. Structure of the proposed SemHi framework. The image (I) has three outputs after passing through the left-side backbone which are encoder deepest feature *S*, decoder deepest feature *F* and classification probability map *P*. *S*, *F*, *P* corresponds to the label embedding module (A), prototype distance measurement module (B), and hierarchical logic regularization module (C). The number 1–3 on label (III) indicates the class hierarchy from coarse to fine, corresponding to the light to dark colors in the class hierarchy (II). Blocks of the same color in B represent similar feature and different colors represent different types of feature. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

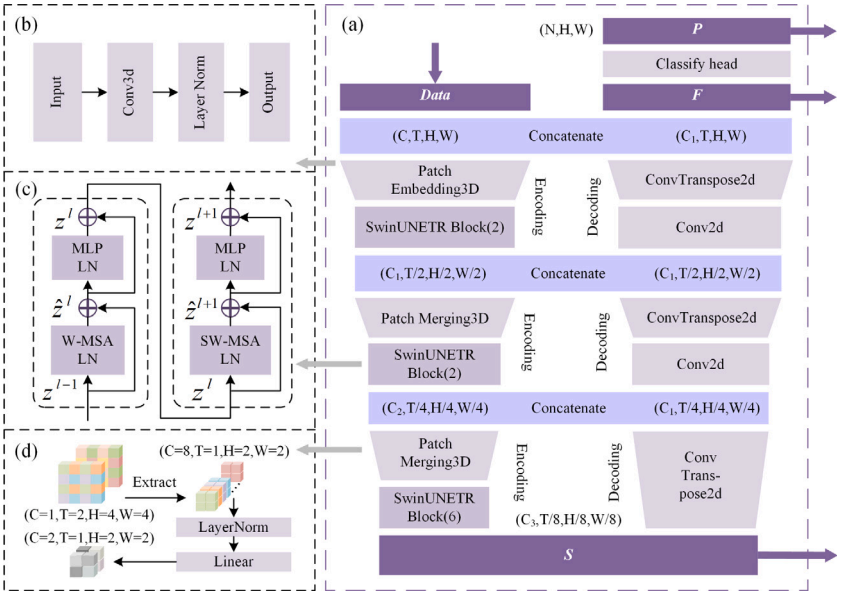


Fig. 2. SwinUNETR Backbone. (a) The structure of SwinUNETR. (b) 3D Patch Embedding which is used to divide the input image into small patches and encode them. (c) SwinUNETR block mainly composed of Layer Norm (LN), (Shifted) Window Multi-head Self-Attention ((S)W-MSA), and Multilayer Perceptron (MLP). (d) 3D Patch Merging, used to reduce the size of feature maps. The purple arrows represent inputs and outputs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The inputs of the label embedding module are the class hierarchy and fine-grained label. For the class hierarchy represented by a tree, we encode it by name embedding and spatial embedding. Firstly, we use a BERT encoder (Devlin et al., 2019) to encode the names of all

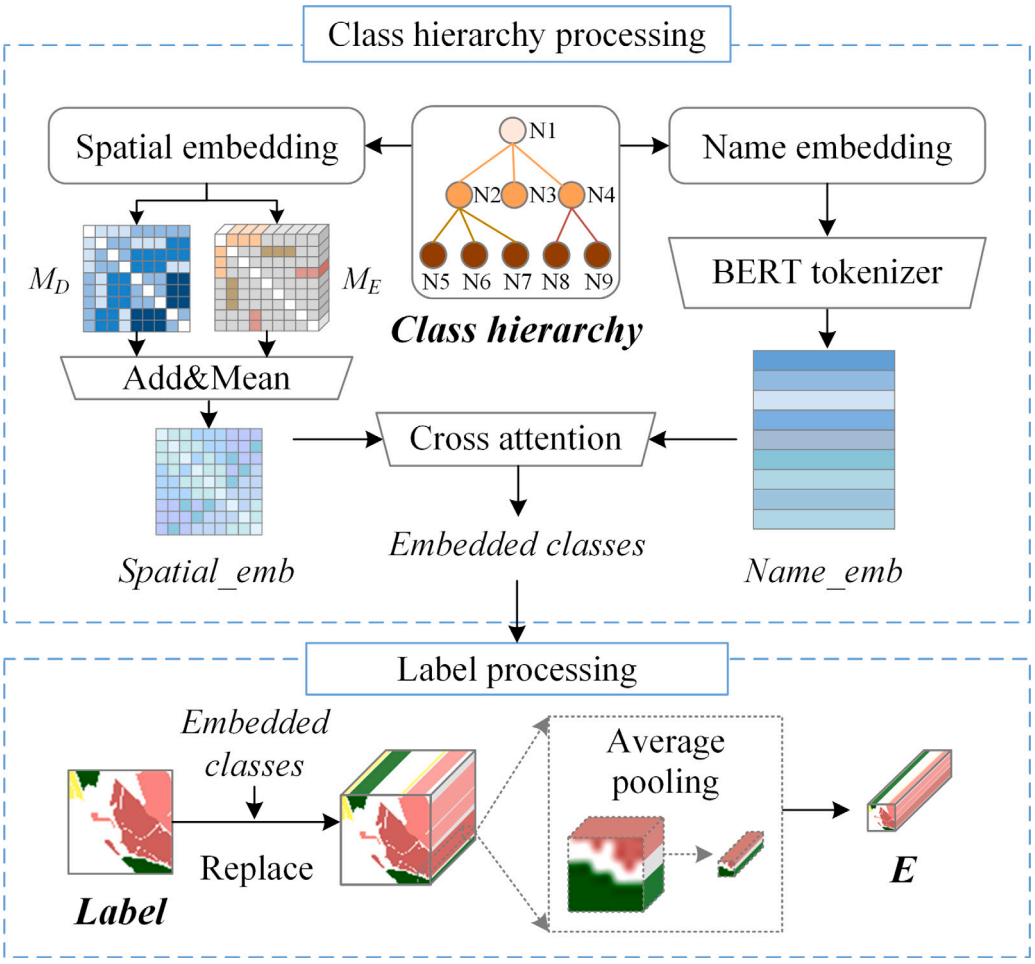


Fig. 3. The structure of the label embedding module, including class hierarchy processing and label processing. First, class hierarchy is processed by encoding the hierarchy structure and class name separately to obtain *spatial_emb* and *name_emb*. Then, they are fused through a cross-attention layer to obtain the embedded classes. Next, fine-grained label is processed by replacing each pixel value with embedded classes. Finally, the replaced 3D label passes through an average pooling layer to obtain the embedded label feature as the output of this module. Bold represents input or output.

class nodes (e.g. N1 to N9) and obtain the embedded name information ($name\_emb \in \mathrm{R}^{n\times d}$) as the node feature. n represents the number of all nodes in the class hierarchy, and d represents the length of the word vector output by BERT. Secondly, we encode the hierarchy structure to get the embedded spatial information ($spatial\_emb \in \mathrm{R}^{n\times n}$) which is composed of a distance matrix ($M_D \in \mathrm{R}^{n\times n}$) and an edge matrix ($M_E \in {}^{a\times n\times n}$). a represents the maximum number of different nodes on branches of two classes, and also represents the maximum length of class name. In Fig. 3 the darker the blue color on M_D, the farther the distance between the two nodes is, and the gray color on M_E indicates that there is no connection between two nodes. We then fuse the *name_emb* and *spatial_emb* through a cross-attention layer to obtain the embedded classes Fig. 4.

The above embedding processes can be defined as:

$$C_e = \mathrm{CrossAttention}(spatial\_emb, name\_emb) \tag{1}$$

$$spatial\_emb = M_D + (\sum_{i=1}^{a} M_E(i))/a \tag{2}$$

$$name\_emb = \sum_{i=1}^{a} \mathrm{Concat}(\mathrm{BERT}(node\_text)) \tag{3}$$

Where $C_e \in \mathrm{R}^{n\times d}$ represents the embedded classes which encodes all class nodes into features containing class name and hierarchical structure information. Subsequently, the vectors of all class names from BERT encoder are concatenated. M_D stores the distance between any two nodes and M_E records the edge information passed from one node to another. Taking N4 and N6 nodes in Fig. 3 as an example, the distance from N4 to its parent node N1, from N1 to N2, and from N2 to N6 is 1. Therefore, the distance between N4 and N6 nodes is 3 and the corresponding vector $m^d_{4,6} = 3$ in M_D. The edge information passing from N4 to N6 is N4 to N1, N1 to N2, and N2 to N6, and hence, the corresponding vector $m^e_{4,6} = [4, 1, 2, 6, 0, 0, 0, 0, 0, 0]$ in M_E, with a length of a. M_D and M_E are both symmetric matrices.

In label processing, the pixel values of the fine-grained label of size H × W are replaced by the vector of the corresponding class in C_e. As a result, the two-dimensional label is embedded into three-dimensional representations ($\mathrm{R}^{\mathrm{H}\times\mathrm{W}} \rightarrow \mathrm{R}^{d\times\mathrm{H}\times\mathrm{W}}$). Afterwards, it is sampled through an average pooling layer, with the same down-sampling factor of the encoder deepest feature S to obtain the embedded label feature as the output of the module, represented by $E \in \mathrm{R}^{d\times\mathrm{H}'\times\mathrm{W}'}$, with H′ × W′ as the spatial size of S:

$$\mathrm{H}' = \mathrm{H}/2^{\mathrm{m}}, \mathrm{W}' = \mathrm{W}/2^{\mathrm{m}} \tag{4}$$

Where m represents the down-sampling factor of the encoder.

The rationale for selecting the BERT encoder is discussed in Section 4.3.

2.3. Prototype distance measurement module

Fine-grained classes have significantly fewer samples than coarse-grained ones, increasing sample imbalance and weakening model learning. Accordingly, we propose the prototype distance measurement

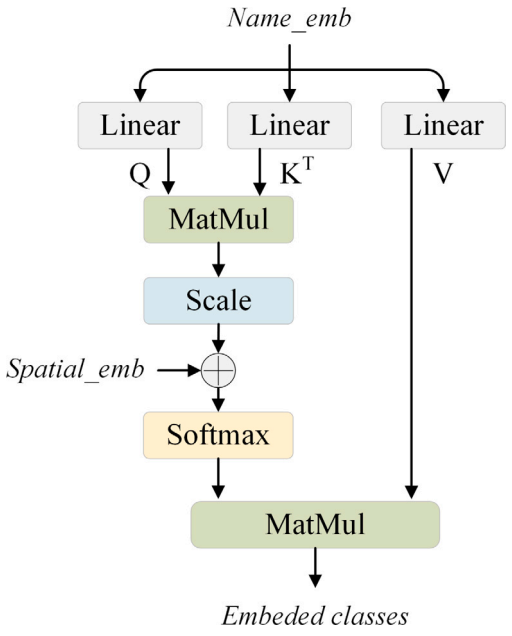


Fig. 4. Cross attention calculation process of *name_emb* and *spatial_emb*. This module mainly consists of linear layers, matrix multiplication operations (MatMul), a scaling step (divided by $\sqrt{d}$) and a softmax activation function. Q, K, V are the outputs of three linear layers, representing the Query, Key, and Value in the Attention mechanism.

module to enhance feature separability. For the deep features of decoder, pixel feature vectors with the same label can be regard as positive samples, while those with different labels are negative samples. After feature extraction and aggregation through the backbone, positive samples should be as similar as possible, while negative samples should be further apart in the representation space. Assuming that there are k samples in a batch of data, with corresponding decoder deepest features and fine-grained labels are $F_i \in \mathrm{R}^{\mathrm{C}'\times\mathrm{H}\times\mathrm{W}}(i = 1,\dots,k)$, $Y_i \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}}(i = 1,\dots,k)$, where C′ represents the number of channels. The mean feature vector with a label of j ($j = 1,\dots,N$) on the decoder deepest feature F_i is presented by $\hat{f}_i^j \in \mathrm{R}^C$ ($i = 1,\dots,k$). Then, the set $\{\hat{f}_1^j,\dots,\hat{f}_k^j\}$ is further averaged to obtain the feature prototype of class j, denoted by $\hat{f}^j$.

We consider the mean feature vector of the same class in F as positive samples and the prototype of different classes in F as negative samples. For example, the set of positive samples for class j is $\{\hat{f}_1^j,\dots,\hat{f}_k^j\}$, and the set of negative samples is $\{\hat{f}^1,\dots,\hat{f}^N\}$, which includes the prototype of feature vectors for all classes except for $\hat{f}^j$. After obtaining the sets of positive and negative samples, the intra-class distance and inter-class distance of each class can be calculated. We use Euclidean Distance to measure the similarity between feature vectors:

$$d_{intra}^j = \frac{1}{k-1}\sum_{i=1}^{k-1}\sqrt{(\hat{f}_i^j - \hat{f}_{i+1}^j)^2} \tag{5}$$

$$d_{inter}^j = \frac{1}{N-1}\sum_{j'=1,j'\neq j}^{N}\sqrt{(\hat{f}^{j'} - \hat{f}^j)^2} \tag{6}$$

Where d_{intra}^j represents the intra-class distance of class j and d_{inter}^j represents the inter-class distance between class j and other classes.

After obtaining the intra-class distance and inter-class distance of each class, the constraints of the loss function can be further used to cluster similar features and disperse different features, thereby improving the separability of deep features in the decoder. The overall structure of the prototype distance measurement module is shown in Fig. 5, which illustrates the above processes using two samples and four classes as example.

2.4. Optimization objectives

The SemHi Framework has four main optimization objectives.

(1) Guide the encoder deepest feature S to fuse class hierarchy and text information through the label embedding module, achieving the embedding of hierarchical concepts.

(2) Improve the separability of the decoder deepest feature F through the prototype distance measurement module.

(3) Output classification results for each level in an end-to-end manner.

(4) Ensure that the classification results of each level conform to the hierarchical logical relationship.

For the first objective, the mean square error is used to guide S to align with the embedded label feature E. $S, E \in d\times\mathrm{H}'\times\mathrm{W}'$, where d represents the length of the word vector output by BERT and H′, W′ represent the height and width, respectively. The vector s_i and e_i in S and E ($s_i, e_i \in \mathrm{R}^d$) can be calculated as follows:

$$L_{emb} = \frac{1}{k}\sum_{i=1}^{k}(s_i - e_i)^2 \tag{7}$$

Where L_{emb} represents embedding loss, k represents the number of samples. The optimization objective is $f : s_i \rightarrow e_i$, which guides the two features to be as similar as possible.

For the second objective, we obtain the intra-class distance $d_{intra} = \{d_{intra}^1,\dots,d_{intra}^N\}$ and inter-class distance $d_{inter} = \{d_{inter}^1,\dots,d_{inter}^N\}$ of N classes through the prototype distance measurement module in the expectation of reducing the d_{intra} and increasing the d_{inter}. Therefore, we construct the prototype loss L_{pro}:

$$L_{pro} = \frac{1}{N}\sum_{j=1}^{N}(d_{intra}^j / d_{inter}^j) \tag{8}$$

For the third objective, the probability of all class nodes are obtained after passing through the backbone network. Therefore, the probability map of each level is activated by simply masking other nodes from different levels. Assuming there are L levels in total, for level l ($l < L$), the probability map P_l is $[P_{N_{l-1}},\dots,P_{N_l}]$, where $[N_{l-1}, N_l]$ represents the class numbers of level l from N_{l-1} to N_l. After activating the classification probability map for each level, the cross entropy loss is calculated:

$$L_{ce}^l = \sum_{i=1}^{k}\sum_{j=N_{l-1}}^{N_l}\alpha_j Y_i^j \log(P_i^j) \tag{9}$$

$$L_{ce} = \frac{1}{L}\sum_{l=1}^{L}L_{ce}^l \tag{10}$$

Where Y represents the label, P represents the classification probability, k is the number of samples and α_j is the weighting of class j, ensuring that each class contributes the same share in the total loss. L_{ce} represents the overall cross entropy loss.

Afterwards, the final classification result for each level can be obtained:

$$\hat{Y}_l = \arg\max[P_{N_{l-1}},\dots,P_{N_l}] \tag{11}$$

Where $\hat{Y}_l$ indicates the classification result at level l, and argmax expresses the operation of taking the index value with the highest probability.

For the last objective, hierarchical logical regularization is conducted to check the logical consistency of the predictions between levels. Specifically, assuming L=3, the predictions should satisfy $\hat{Y}_3 \subseteq \hat{Y}_2$, $\hat{Y}_2 \subseteq \hat{Y}_1$, $\hat{Y}_3 \subseteq \hat{Y}_1$. Let $s = [0, 1, 2, 3]$ be the number of items that meet the requirements and its corresponding loss value is $v = [3.1, 2.1, 1.1, 0]$. The hyper parameter v is empirically determined. Due to variations in the numerical ranges of loss terms, we adjust v so that the initial value of logical loss is of a similar order of magnitude as the other loss

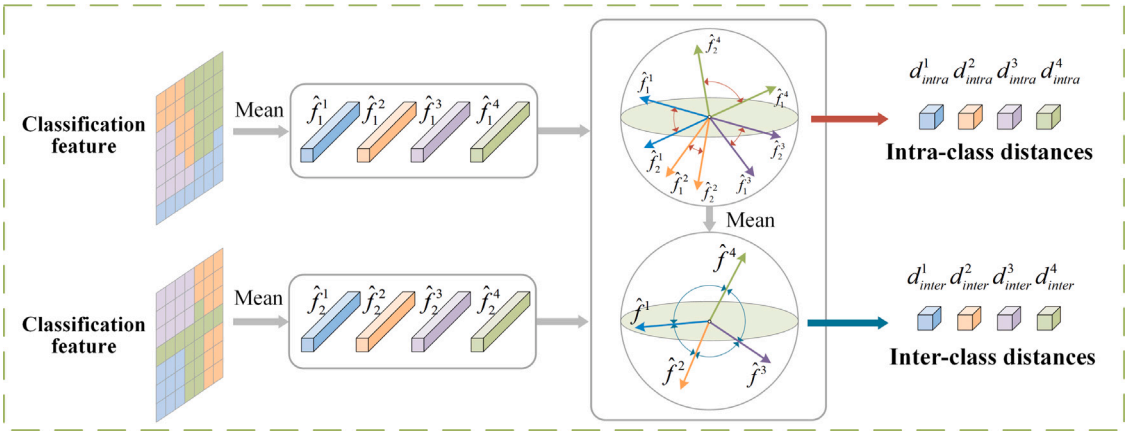


Fig. 5. The structure of the prototype distance measurement module. Different colors represent different classes. Circles represent the feature space, with single arrows indicating feature vectors, red double arrows for intra-class distances, and blue double arrows for inter-class distances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

terms. This keeps loss components at a consistent scale, and stabilizes the optimization process. The logical loss is constructed as follows:

$$L_{log} = \frac{1}{\mathrm{H} \times \mathrm{W}} \sum v_{ij} \tag{12}$$

Where v_{ij} represents the loss value of pixel (i, j).

Finally, our overall loss can be expressed as:

$$Loss = \lambda_1 L_{emb} + \lambda_2 L_{pro} + \lambda_3 L_{ce} + \lambda_4 L_{log} \tag{13}$$

Where λ represents the weight of each type of loss.

3. Datasets

In this study, we employ both scene classification and pixel-level segmentation datasets to evaluate the framework's versatility and applicability across different task levels. The scene classification dataset is used to assess the framework's performance on scene-level classification tasks, thereby validating its adaptability across varying task granularities. The two crop segmentation datasets aid in evaluating the framework's fine-grained classification capabilities in segmentation tasks under a multi-level fusion of spatial, temporal, and hierarchical information. By selecting datasets spanning multiple tasks, this study demonstrates the model's task adaptability and structural generalization capacity. Additionally, we evaluate text encoder performance within the label embedding module using the scene classification dataset, whose lower resolution and simpler label structure offer a preliminary validation of the selection of text encoder. This design provides insights into the framework's architecture under reduced data and computational demands.

3.1. Crop segmentation datasets

The Sen4AgriNet dataset (Sykas et al., 2021) comprises Sentinel-2 patches from Spain and France. After sorting and integrating, we obtain 33 class nodes, including 7 primary nodes, 20 secondary nodes, and 25 tertiary nodes, following the principle that each level of classification includes all nodes of that level and leaf nodes of coarser levels. Fig. 6 shows the label tree for all classes used in this dataset. Sen4AgriNet includes a total of 225,000 SITS samples with a size of 366 × 366, each containing 30–50 time acquisitions and 13 spectral bands. We crop samples into 61 × 61 and resample the time series to 12 (taking the median of monthly observations), and then divide the dataset into training, test and validation sets in a 7:2:1 ratio. Fig. 7 shows the true color composite images and their tertiary labels of this dataset.

The ZueriCrop dataset (Turkoglu et al., 2021) contains Sentinel-2 images from the Zurich and Thurgau regions of Switzerland in 2019, including 58 crop types, with 5 primary class nodes, 14 secondary class nodes, and 48 tertiary class nodes. Fig. 8 shows the label tree of the ZueriCrop dataset. This dataset includes a total of 116,000 SITS samples with a size of 24 × 24, each containing 71 time acquisitions and 9 spectral bands. This dataset is also divided into the training, test, and validation sets in a 7:2:1 ratio. Fig. 9 shows the true color composite images and their tertiary labels of this dataset.

These two datasets have unique properties, which make them ideal for evaluating the SemHi framework. First, they cover crop-growing regions across three countries, testing the model's generalization across geographic variations. Second, the datasets differ in size – Sen4AgriNet is larger, while ZueriCrop is smaller – allowing for robustness evaluation across different scales. Finally, both datasets feature complex hierarchical class structures, which are essential for our framework and increase the challenge of fine-grained classification.

3.2. RS scene classification dataset

The SemHi framework can be applied to various DL networks, demonstrating its potential as a generic framework. To test its generality, we also conducted experiments on a remote sensing image classification dataset RSI-CB (Li et al., 2020b), which consists of multiple remote sensing scenes. This dataset is also used for comparative experiments of text encoders. The RSI-CB dataset contains a total of 42 scene types, including 7 primary class nodes and 35 secondary class nodes. Each fine-grained class contains 198 to 1331 remote sensing images. RSI-CB consists of multiple data sources, with 3 bands and a size of 256 × 256, and resolutions ranging from 0.3–3 m. Each image has a corresponding coarse-grained and fine-grained label. This dataset is also divided into the training, test, and validation sets in a 7:2:1 ratio. Fig. 10 shows the hierarchical structure of the RSI-CB and the proportion of each fine-grained class.

4. Experiments and results

4.1. Implementation details

The SemHi framework is implemented using PyTorch with Adaptive Moment Estimation (Adam) algorithm as the optimizer and a batch size

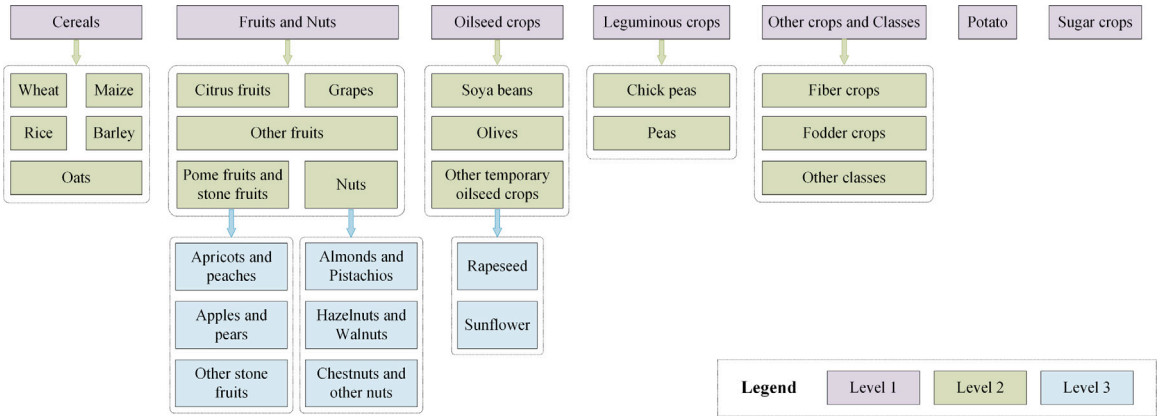


Fig. 6. Hierarchical classification architecture of the Sen4AgriNet dataset.

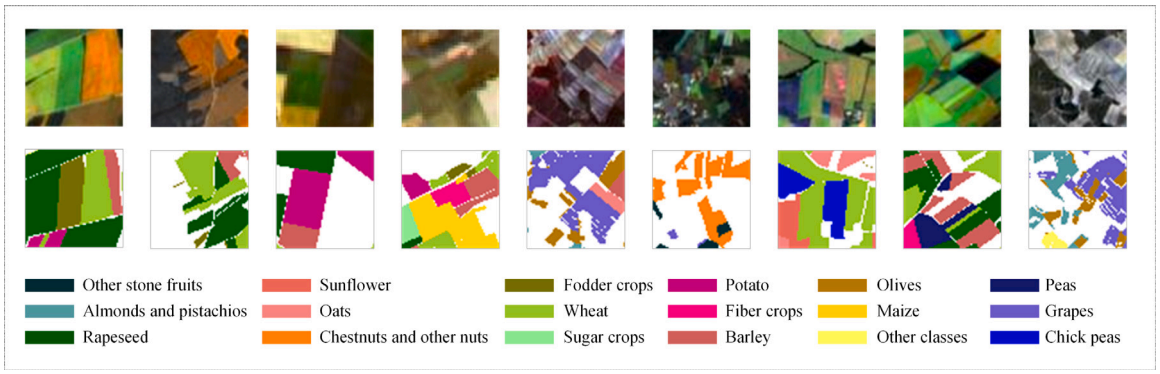


Fig. 7. True color composite images and tertiary labels in the Sen4AgriNet dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

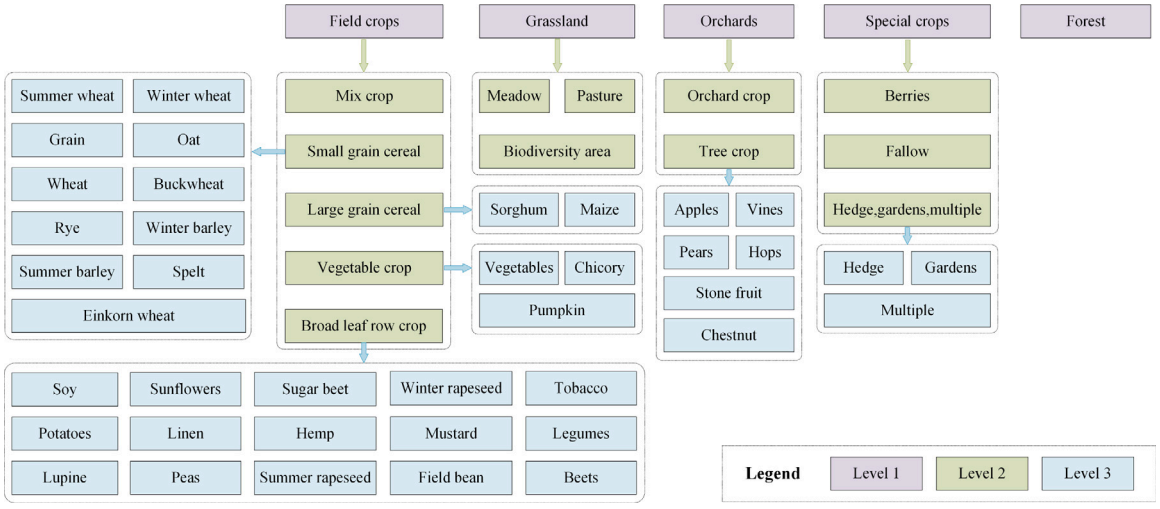


Fig. 8. Hierarchical classification architecture of the ZueriCrop dataset.

of 8. Initial learning rate is set to 0.0001 at the beginning of training, and then use cosine annealing to adjust the learning rate and regularize it with a weight decay of 10^{-6}. The weights of multiple optimization objectives are set as $\lambda_1 = 2, \lambda_2 = 4, \lambda_3 = 1, \lambda_4 = 3$. To simplify weight tuning, we adopt an empirical weight adjustment approach (Eigen and Fergus, 2015). Using one loss term as a reference, we scale others to a similar magnitude for initial balance in optimization. During training, loss values adjust dynamically, preventing any single loss from dominating optimization, leading to more stable training and faster convergence. Considering the computation cost, training epoch is set to 20 for Sen4AgriNet, 60 for ZueriCrop and 30 for RSI-CB.

4.2. Evaluation metrics

For crop segmentation, performance is evaluated on a per-pixel basis using four different metrics, F1-score, overall accuracy (OA), Kappa coefficient, and mean intersection over union (MIoU). For scene classification, only F1-score, OA, and Kappa are used. F1-score is the

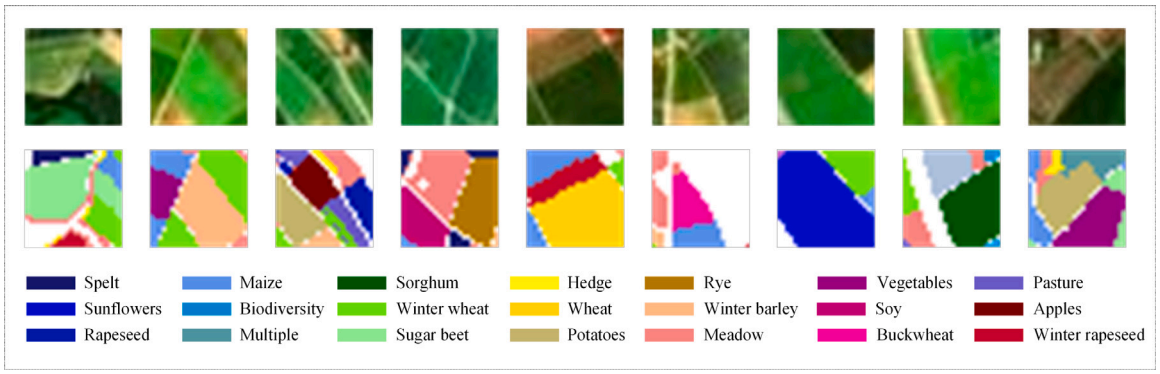


Fig. 9. True color composite images and tertiary labels in the ZueriCrop dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

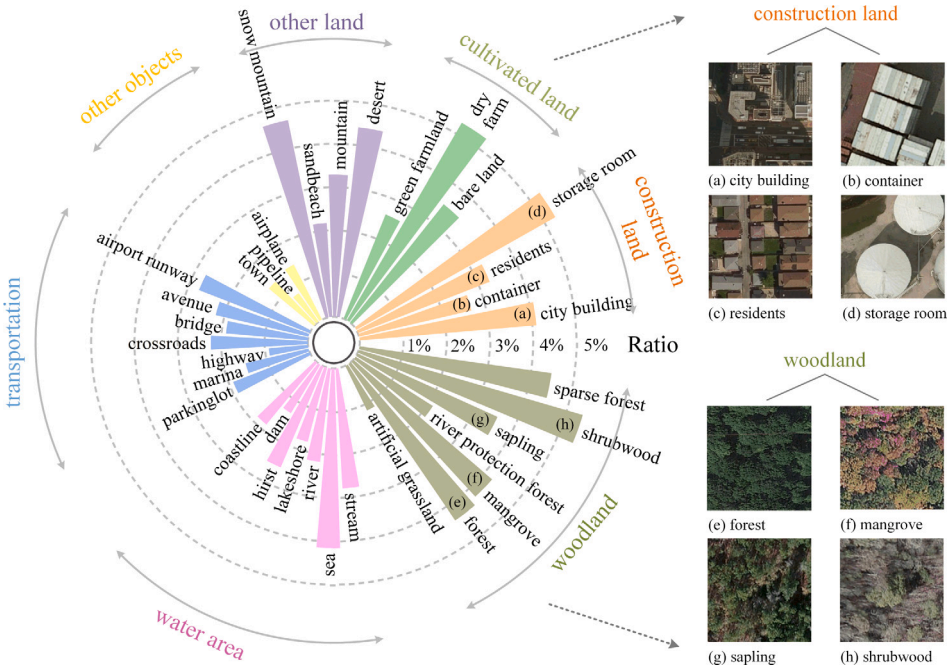


Fig. 10. Hierarchical structure of the RSI-CB and the proportion of each fine-grained class. The right side displays samples of two coarse-grained classes and some of their fine-grained classes.

harmonic mean of precision and recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

Where TP represents true positives, FP is false positives, FN is false negatives, TN is true negatives.

OA is the proportion of correctly classified pixels over the total pixels:

$$OA = \frac{TP}{TP + FP + TN + FN} \tag{17}$$

Kappa measures the agreement between predicted and ground truth while accounting for random chance:

$$kappa = \frac{OA - p_e}{1 - p_e} \tag{18}$$

Where p_e is the expected agreement by chance.

MIoU is the average IoU across all classes, defined as:

$$MIoU = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{19}$$

where C is the total number of classes.

In addition to performance evaluation, we compare computational complexity using floating point operations (FLOPs), parameter count, and per-epoch training time to compare computational complexity. FLOPs measure the total number of arithmetic operations (multiplications and additions) performed during inference or training. Parameter count refers to the total number of trainable weights in the model, affecting memory usage. Per-epoch training time provides an intuitive comparison of computational cost in real-world scenarios.

4.3. Selection of text encoder

Text encoder plays a crucial role in transforming class names from natural language into a computer-recognizable programming language in the label embedding module. Six candidate text encoders are selected for comparison, including Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019), Decoding-enhanced BERT

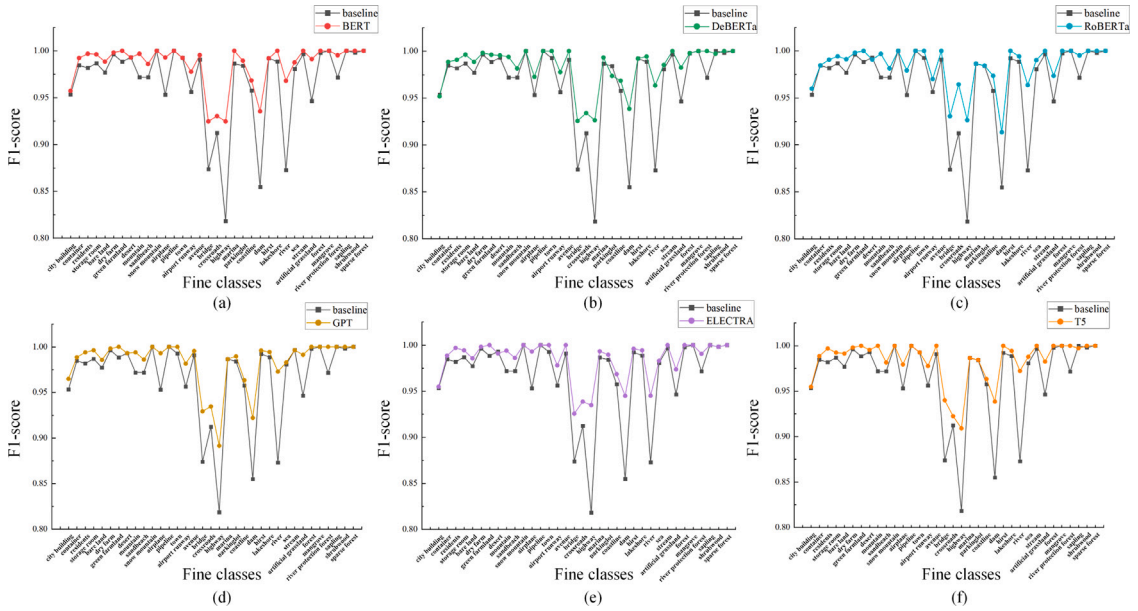


Fig. 11. Comparison of accuracy between six text encoders and baseline networks.

Table 1
Accuracy of the SwinT-F with different text encoder on the RSI-CB test set. The best results are highlighted in bold.

Model	OA (%)	Kappa (%)
Baseline	97.64	97.56
-BERT	**98.87**	**98.83**
-DeBERTa	98.71	98.66
-RoBERTa	98.83	98.79
-GPT	98.81	98.77
-ELECTRA	98.75	98.71
-T5	98.81	98.76

with disentangled attention (DeBERTa) (He et al., 2020), Robustly Optimized BERT (RoBERTa) (Liu et al., 2019), Generative Pre-Trained Transformer (GPT) (Radford et al., 2019), ELECTRA (Clark et al., 2020), and Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020). Considering the time cost, the experiments are conducted on the RSI-CB dataset which is designed for classification tasks. Swin Transformer (SwinT) is selected as the baseline. Then, the SemHi framework is applied to obtain SwinT-F, after which the text encoder is sequentially replaced. Table 1 presents the accuracy results of the SemHi framework on fine-grained classes when different text encoders are applied.

From the Table 1, it can be observed that all text encoders can improve the classification accuracy of the baseline and BERT achieves the best performance. BERT adopts a bidirectional Transformer to model bidirectional context effectively. Additionally, it is pre-trained using the Masked Language Model (MLM) by predicting masked words. This enables BERT to learn rich contextualized representations, making it effective for words, phrases, and full sentences. RoBERTa and DeBERTa improve BERT with optimized attention and training strategies but show limited benefits in this task, as crop class names are short and do not require complex dependency modeling. GPT's unidirectional design may hinder full contextual understanding. ELECTRA, using Replaced Token Detection (RTD) rather than MLM, relies more on complete contextual, reducing its effectiveness in learning deep contextual representations. T5, though highly generalizable, is designed for natural language generation (NLG), whereas our task focuses on text embedding.

We also visualized the f1-scores of all fine-grained classes for the six text encoders. Fig. 11 clearly demonstrates that, compared to the baseline, SwinT-F consistently achieves superior fine-grained classification accuracy regardless of the text encoder applied, with BERT showing more notable improvements. Therefore, BERT is selected as the text encoder. Then the SemHi framework is evaluated in the subsequent crop segmentation experiments.

4.4. Crop segmentation experiments

Six SOTA networks, including SwinUNETR, TSViT, ConvGRU, ConvSTAR, UNet3Df and UNet3D, is adopted and compared as the backbone of framework (suffix -F in Table 2).

For the Sen4AgriNet dataset (Table 2), it is observed that all models under the SemHi framework perform better than their original versions. Our proposed framework shows the most significant improvement of accuracy at level 3. Even for ConvSTAR-F which has lower accuracy, these three metrics are about 5% better than ConvSTAR.

Fig. 12 shows the prediction results at level 3 class for a set of samples. From Fig. 12(a)–(d), it can be seen that the wrong part within the black circle is significantly less than that within the red circle, which means the networks under our SemHi framework have significantly fewer erroneous predictions than the SOTA versions. In addition, our SemHi framework predictions have less noise and more complete crop parcels, indicating it can obtain clearer crop boundary features, and suppress confusion of edge classes. We also compare the network performance and computational complexity under the SemHi framework (suffix -F) with different networks on the ZueriCrop dataset (Table 3). Consistent with the results of the Sen4AgriNet, all networks perform better under our SemHi framework. In addition, as shown in Table 3, the FLOPs and parameters of the baselines and frameworks remains nearly identical, but training time differs. Under the same hardware conditions, the training time per epoch for the proposed framework is 1.25 to 1.82 times that of the baselines. Fig. 13 shows the prediction results of ZueriCrop test samples.

Furthermore, for the two datasets, the networks under the SemHi framework with the best performance are different, namely SwinUNETR-F and UNet3D-F. This may be attributed to the large amount of data in the Sen4AgriNet, which is about 70 times that of ZueriCrop, while SwinUNETR-F based on self-attention mechanism can show advantages over convolution in processing larger amounts of data (Dosovitskiy et al., 2020). Therefore, SwinUNETR-F performs better on the Sen4AgriNet.

Table 2
Accuracy of the proposed SemHi framework on the Sen4AgriNet test set. The best results are highlighted in bold.

Model	Level 1 (%)			Level 2 (%)			Level 3 (%)		
	OA	Kappa	MIoU	OA	Kappa	MIoU	OA	Kappa	MIoU
SwinUNETR	92.18	87.79	79.17	86.61	84.42	67.16	83.33	80.83	60.73
TSViT	89.65	84.20	74.72	84.24	81.71	63.80	80.75	77.96	57.77
UNet3D	91.12	86.03	77.47	85.74	83.57	68.19	85.56	83.39	66.56
UNet3Df	92.32	88.09	80.72	86.14	83.78	68.28	85.92	83.75	65.14
ConvGRU	88.43	82.63	71.07	82.56	79.10	59.73	71.58	67.54	46.17
ConvSTAR	89.88	84.38	73.53	82.07	79.26	61.17	76.91	73.58	52.86
SwinUNETR-F	92.64	88.53	80.66	**86.72**	**84.60**	**69.10**	**86.63**	**84.61**	**67.98**
TSViT-F	89.74	84.23	74.89	84.85	82.40	64.08	84.78	82.46	63.59
UNet3D-F	91.21	86.46	77.82	86.36	84.19	68.27	86.08	83.98	66.18
UNet3Df-F	**93.01**	**89.03**	**81.23**	86.52	84.38	68.72	86.40	84.35	67.28
ConvGRU-F	88.90	83.04	72.11	82.64	79.88	60.27	82.46	79.82	59.40
ConvSTAR-F	88.76	82.82	70.79	82.30	79.52	60.61	81.99	79.32	58.80

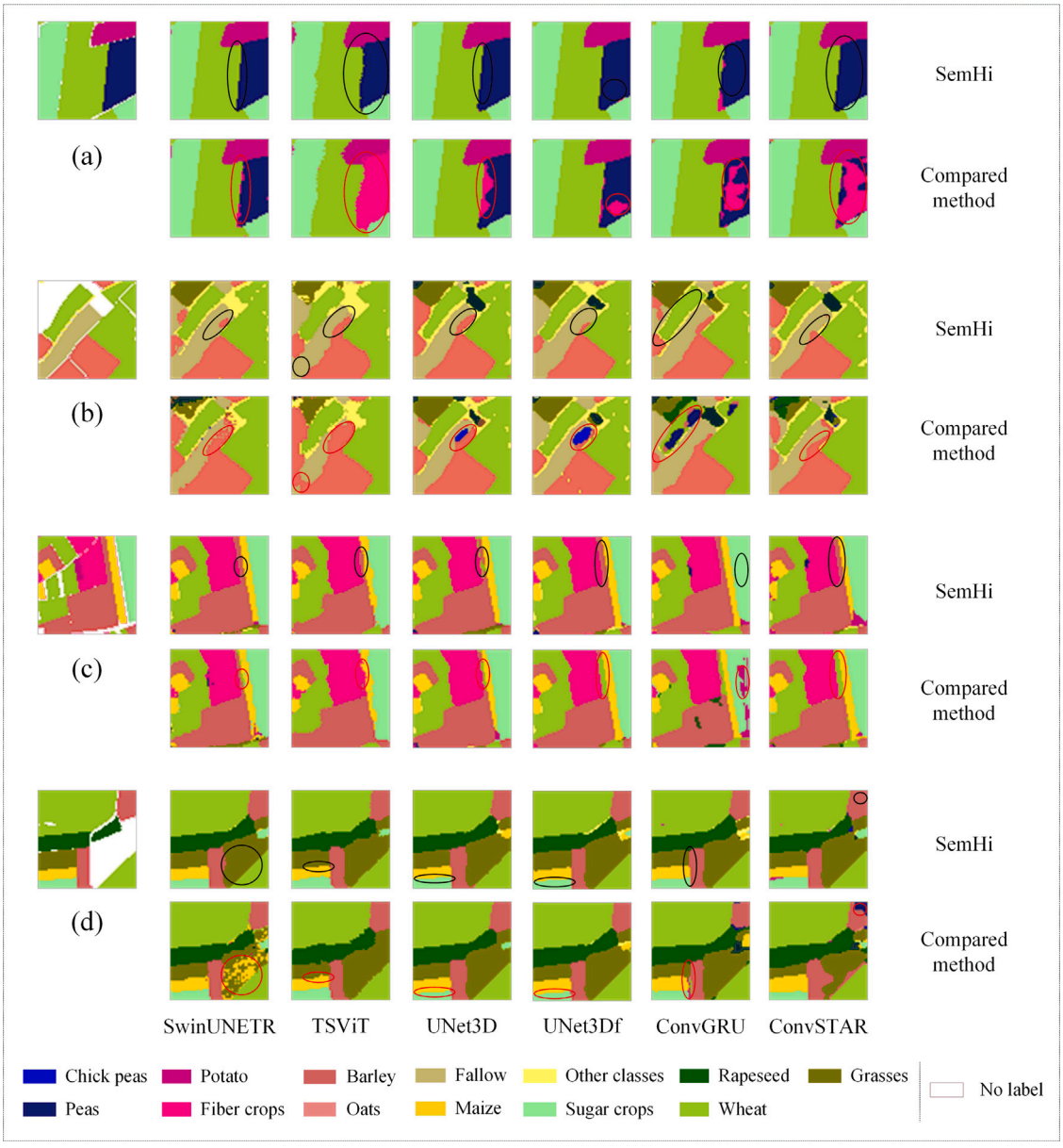


Fig. 12. The prediction results of four test images in Sen4AgriNet. For the samples (a)–(d), the leftmost column displays the labels of level 3 classes, each column on the right represents different networks. The red circles represent the significant prediction errors of the compared method, and the black circles represent the errors from the SemHi framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Accuracy, FLOPs, parameters, and per-epoch training time of the proposed SemHi framework on the ZueriCrop test set. The best accuracy results are highlighted in bold.

Model	Level 1 (%)			Level 2 (%)			Level 3 (%)			FLOPs (G)	Para. (M)	Time (min)
	OA	Kappa	MIoU	OA	Kappa	MIoU	OA	Kappa	MIoU			
SwinUNETR	90.86	82.92	52.16	79.38	72.03	40.34	64.73	55.86	24.46	359.83	6.74	11.10
TSViT	88.08	78.19	46.23	74.90	67.41	34.19	58.37	48.80	15.21	417.01	2.17	7.93
UNet3D	94.02	87.23	60.21	82.35	75.97	45.49	72.96	65.59	31.84	440.34	6.38	7.33
UNet3Df	93.84	88.07	64.35	81.32	76.15	48.74	72.78	65.44	30.77	440.34	6.38	7.58
ConvGRU	89.49	80.87	49.36	76.17	69.54	38.10	64.67	54.92	15.63	176.86	0.50	18.84
ConvSTAR	88.86	79.85	51.59	74.34	67.18	40.01	70.57	63.72	30.03	216.85	0.63	20.10
SwinUNETR-F	91.82	84.53	58.09	80.27	73.44	43.24	77.60	70.67	**39.09**	↑0.001	↑0.004	13.91
TSViT-F	90.26	81.82	50.30	75.02	67.46	38.18	72.23	64.80	28.56	↑0.002	↑0.004	14.32
UNet3D-F	**94.34**	**88.95**	60.77	82.79	76.33	45.54	**80.46**	**73.88**	32.45	↑0.003	↑0.002	13.33
UNet3Df-F	94.10	88.54	**65.64**	**82.90**	**76.64**	**49.27**	79.28	72.52	32.38	↑0.003	↑0.002	13.78
ConvGRU-F	89.77	81.28	50.88	77.62	69.69	38.36	75.08	67.00	26.41	↑0.003	↑0.004	28.62
ConvSTAR-F	89.41	80.71	52.35	75.90	68.75	40.38	73.02	65.93	30.81	↑0.003	↑0.003	29.25

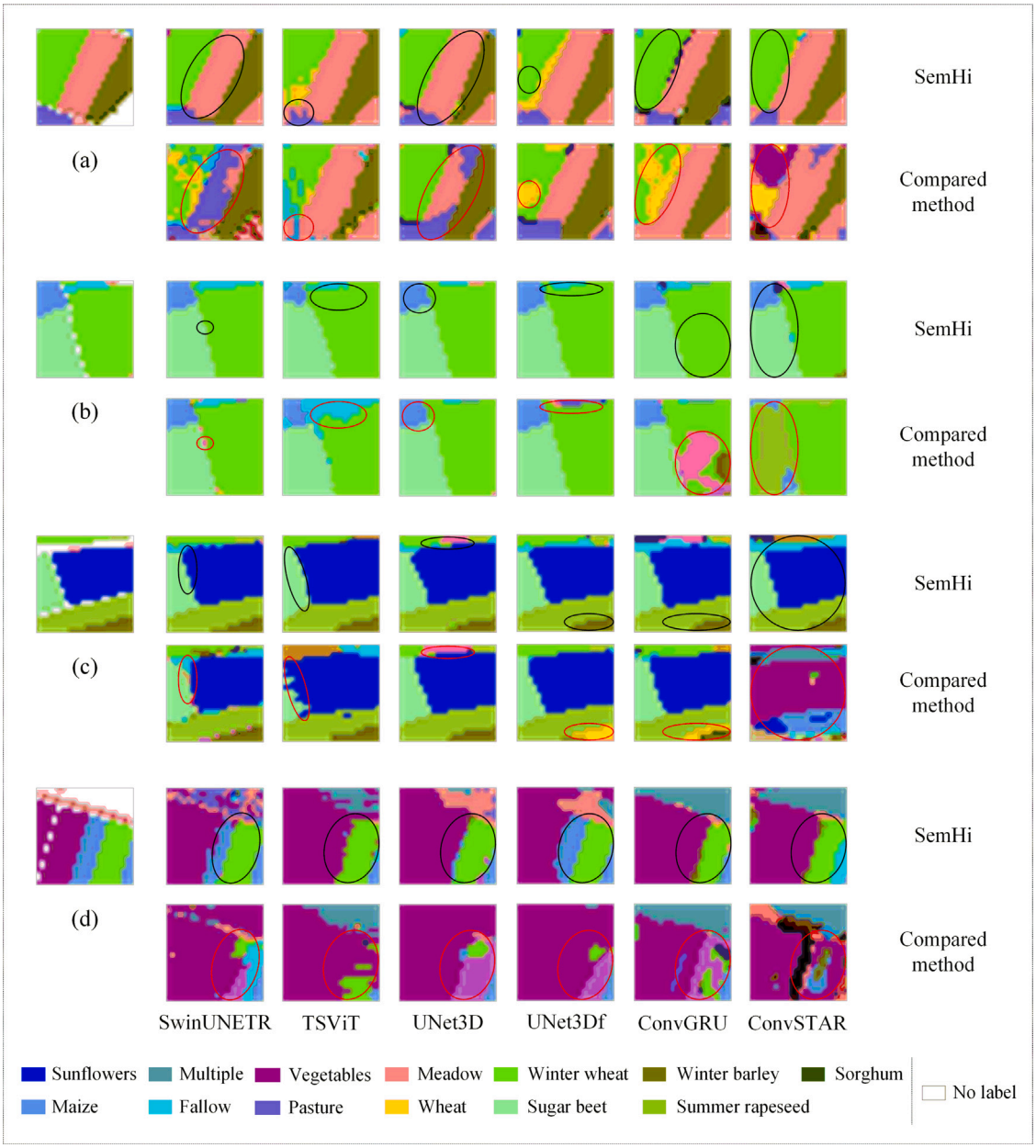


Fig. 13. The prediction results of four test images in ZueriCrop. For the samples (a)–(d), the leftmost column displays the labels of level 3 classes, each column on the right represents different networks. The red circles represent the significant prediction errors of the compared method, and the black circles represent the errors from the SemHi framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Accuracy results of ablation experiments on the two datasets. The best results are highlighted in bold.

Dataset	Model	Level 1 (%)			Level 2 (%)			Level 3 (%)		
		OA	Kappa	MIoU	OA	Kappa	MIoU	OA	Kappa	MIoU
Sen4AgriNet	Baseline	92.18	87.79	79.17	86.61	84.42	67.16	83.33	80.83	60.73
	Baseline+L	91.98	87.58	78.88	86.53	84.38	68.02	86.40	84.35	66.83
	Baseline+L+E	92.30	88.02	79.89	86.49	84.33	68.25	86.46	84.42	67.28
	Baseline+L+E+P	**92.64**	**88.53**	**80.66**	**86.72**	**84.60**	**69.10**	**86.63**	**84.61**	**67.98**
ZueriCrop	Baseline	90.86	82.92	52.16	79.38	72.03	40.34	64.73	55.86	24.46
	Baseline+L	90.70	82.60	54.13	77.69	70.27	40.11	73.13	65.19	31.15
	Baseline+L+E	91.45	83.86	56.74	79.03	71.68	41.44	75.92	68.37	32.70
	Baseline+L+E+P	**91.82**	**84.53**	**58.09**	**80.27**	**73.44**	**43.24**	**77.60**	**70.67**	**39.09**

Overall, the quantitative results indicate that with the deepening of classification levels and the refinement of crop classes, all networks show a trend of gradually decreasing accuracy. However, our SemHi framework can effectively improve the crop classification at the fine-grained scale for all compared methods, verifying the effect of the proposed SemHi framework in hierarchical guiding.

4.5. Module performance experiments

To further investigate the rationality and necessity of the designed modules and optimization objectives, we use SwinUNETR as the baseline network and add each module in sequence for module performance (i.e., ablation) experiments on the two datasets. Baseline+L adds the hierarchical logic regularization module on the basis of SwinUNETR. Baseline+L+E indicates addition of the label embedding module on the basis of Baseline+L. Finally, the prototype distance measurement module is added to Baseline+L+E to achieve the overall framework, i.e., Baseline+L+E+P. Table 4 shows the accuracy of different combinations at different levels on the two datasets, where the ablation models share the same training parameter settings.

(1) Hierarchical logic regularization module. Results in Table 4 indicate that Baseline+L improved the OA, Kappa, and MIoU for level 3 classes on Sen4AgriNet and ZueriCrop. This indicates that the designed logic loss can effectively utilize the information of coarse-grained classes and reduce the confusion of fine-grained classes. However, there is a slight accuracy decrease of Baseline+L at level 1 and level 2 classes. This may result from using the hierarchical logic regularization module alone. This module enforces consistency between fine-grained and coarse-grained classifications, which may lead to adjustments in low-probability predictions at the coarse level, slightly affecting accuracy. Future work could explore adaptive weighting in logical loss computation, adjusting classification instance weights to maintain accuracy across hierarchical levels and improve overall balance.

(2) Label embedding module. Comparing the results of Baseline+L and Baseline+L+E in Table 4, it can be found that for Sen4AgriNet, Baseline+L+E improves the classification accuracy of level 1 and level 3 classes. For ZueriCrop, Baseline+L+E raises the classification accuracy at all levels. In conclusion, the label embedding module effectively mitigates the bias of logical loss on coarse-grained and enhances classification performance at each level. To further investigate the contribution of the name embedding and spatial embedding in label embedding module, we compared them using SwinUNETR as the baseline. As shown in Table 5, name embedding has greater importance than spatial embedding, and their combined effect outperforms each component individually.

(3) Prototype distance measurement module. From Table 4, it can be seen that for both datasets, Baseline+L+E+P has improved classification accuracy at all levels. This indicates that the prototype distance measurement module can effectively strengthen the feature separability by aggregating and separating feature vector prototypes. Additionally, as shown in Table 5, Baseline+P yields less significant accuracy improvements across all levels compared to Baseline+L and Baseline+E, suggesting its relatively lower importance than the hierarchical logic regularization module and label embedding modules.

It can be noticed that compared to the Sen4AgriNet dataset, the accuracy improvement on the ZueriCrop dataset is more significant. This may be due to the fact that ZueriCrop dataset has a more complex hierarchical structure and class system, and has a smaller amount of samples, leaving more significant room for classification accuracy improvement.

To further evaluate the proposed components, we analyze level 3 classification results on ZueriCrop, as it shows the most significant improvement in accuracy. Using Gradient-weighted Class Activation Mapping (Grad CAM) (Selvaraju et al., 2017), we visualize learned feature regions through heatmaps for different ablation models (Fig. 14). Visual analysis shows that Baseline struggles to focus on key feature regions. As modules are added, heatmaps exhibit darker, more defined regions aligned with labels. Baseline+L effectively recovers missing regions, indicating that hierarchical logic regularization reduces fine-grained classification confusion by enforcing logical consistency. Baseline+L+E further enhances attention to key areas by leveraging hierarchical structure and class semantics. Baseline+L+E+P refines region boundaries, demonstrating that the prototype distance measurement module helps delineate crop boundaries by optimizing feature aggregation and separation. Overall, module comprehensiveness correlates with heatmap clarity, validating each component's role in improving feature representation.

4.6. Framework generality experiments

To evaluate whether the SemHi framework has the potential for application in other tasks and domains, we test its applicability on the RSI-CB remote sensing scene classification dataset. Three commonly adopted classification networks – ResNet34 (He et al., 2016), ViT (Dosovitskiy et al., 2020), and SwinT (Liu et al., 2021) – are employed as the backbone. Then, the applicable components of the SemHi framework, namely the label embedding module and the hierarchical logic regularization module, are integrated. It is important to note that, since RSI-CB focuses on scene classification and only involves the encoding process, the prototype distance measurement module based on decoder features is omitted. The comparison of accuracy results between the compared methods and the SemHi framework on fine-grained classes is shown in Table 6.

Table 6 shows that for fine-grained classes, the accuracy of all compared methods improves under the SemHi framework. Among the six networks, SwinT-F achieves the best performance. The performance improvements are due to the effective fusion of text-based label embeddings with scene-level spatial features, allowing the model to leverage both linguistic and visual cues for enhanced classification.

Unlike the Sen4AgriNet and ZueriCrop datasets, the RSI-CB dataset focuses on remote sensing image scene classification and has only a two-level hierarchical structure. These differences highlight two key points: (1) The SemHi framework is not only applicable to crop pixel-level classification tasks (using SITS) but can also be applied to image-level remote sensing classification tasks (using satellite images), demonstrating a degree of task generalization. (2) From the perspective of hierarchical classification, the SemHi framework is adaptable to various hierarchical structures, capable of performing classification at different levels of granularity, thereby exhibiting structural generalization.

Table 5
Comparison of accuracy among name embedding (Name), spatial embedding (Spatial), full label embedding module (E), logic regularization module (L), and prototype distance measurement module (P) on the ZueriCrop dataset. The best results are highlighted in bold.

Model	Level 1 (%)			Level 2 (%)			Level 3 (%)		
	OA	Kappa	MIoU	OA	Kappa	MIoU	OA	Kappa	MIoU
Baseline+Name	87.65	77.66	51.09	78.37	70.49	37.06	73.02	64.66	31.08
Baseline+Spatial	84.81	73.43	46.36	74.08	66.03	34.62	72.81	64.47	30.29
Baseline+E	89.93	79.40	53.15	**78.84**	**71.06**	39.92	**73.82**	65.00	**32.07**
Baseline+L	**90.70**	**82.60**	**54.13**	77.69	70.27	**40.11**	73.13	**65.19**	31.15
Baseline+P	87.06	76.81	48.07	73.66	65.94	35.57	69.88	62.09	30.55

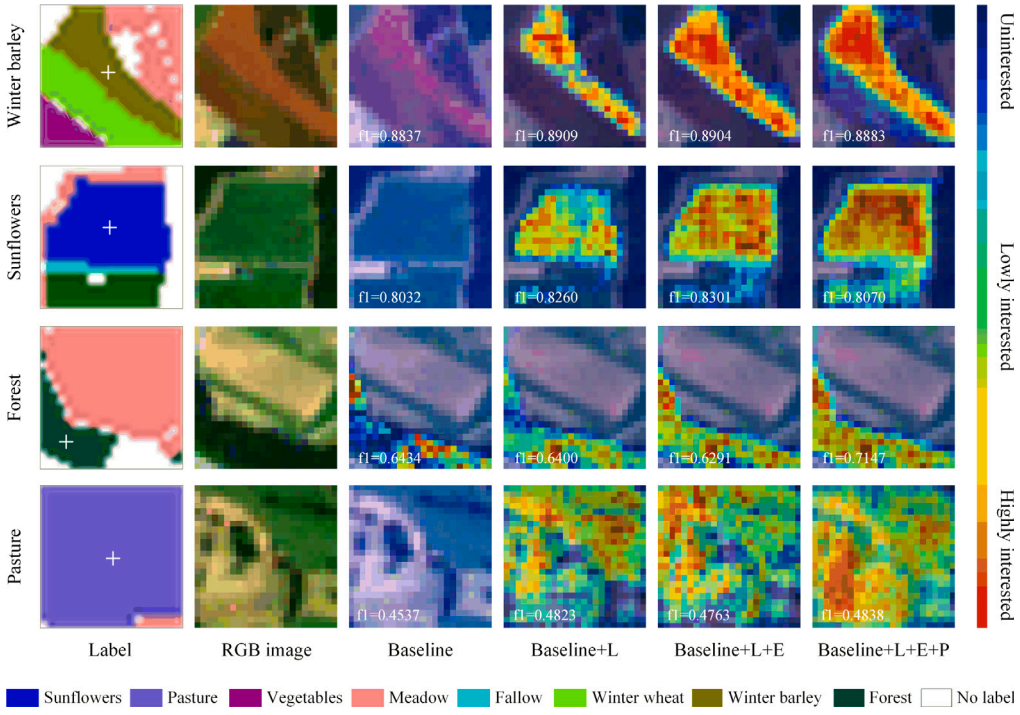


Fig. 14. Grad CAM visualization heatmaps of ablation models on four fine-grained classes in the ZueriCrop dataset. The leftmost text indicating the visual class of the row, which is located by a white cross on the label.

Table 6
Accuracy of the proposed SemHi framework on the RSI-CB test set. The best results are highlighted in bold.

Model	OA (%)	Kappa (%)
ResNet34	98.15	98.08
ViT	97.12	97.01
SwinT	97.64	97.56
ResNet34-F	98.55	98.50
ViT-F	98.02	97.95
SwinT-F	**98.87**	**98.83**

5. Conclusions

In this study, we propose the Class Semantic Guided Hierarchical Segmentation Framework (SemHi framework), leveraging spatio-temporal-text fusion for crop classification. This framework encodes the hierarchical structure and class text, aligning them with the spatio-temporal features from encoder to achieve a comprehensive fusion of textual semantics and spatio-temporal characteristics. Additionally, we perform prototype distance measurement on the deep features extracted by the backbone decoder to enhance intra-class similarity and inter-class separability. Furthermore, a hierarchical logic regularization module is introduced to support multi-granularity predictions while ensuring logical consistency across levels.

The SemHi framework excels on two public crop datasets with fine-grained classes, enabling coarse-to-fine classifications to meet diverse user needs. This framework has potential applications in crop structure optimization, yield prediction and farmland management, offering valuable data support for precision agriculture. Furthermore, the framework is highly scalable and sensor-agnostic, enabling broader applications such as forest type monitoring, ecological assessment, and land use classification. In summary, the multi-dimensional fusion of text, spatial, and temporal information across hierarchical levels is pivotal to our framework's success, enhancing its adaptability and performance across diverse remote sensing tasks and exemplifying the power of information fusion for complex, hierarchical classification.

CRediT authorship contribution statement

Xiyao Li: Writing – original draft, Validation, Software, Methodology, Conceptualization, Writing – review & editing. **Jiayi Li:** Writing – original draft, Supervision, Methodology, Conceptualization, Writing – review & editing. **Jie Jiang:** Funding acquisition, Project administration, Supervision, Visualization, Writing – review & editing. **Xiaofeng Pan:** Writing – original draft, Supervision, Investigation. **Xin Huang:** Writing – original draft, Supervision, Methodology, Conceptualization, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903701 and in part by the National Natural Science Foundation of China under Grant 42471391 and 42271328.

Data availability

The authors do not have permission to share data.

References

Bueno, I.T., Antunes, J.F.G., Dos Reis, A.A., Werner, J.P.S., Toro, A.P.S.G.D.D., Figueiredo, G.K.D.A., Esquerdo, J.C.D.M., Lamparelli, R.A.C., Coutinho, A.C., Magalhães, P.S.G., 2023. Mapping integrated crop-livestock systems in Brazil with planetscope time series and deep learning. Remote Sens. Environ. 299, 113886.

Chen, J., Qian, Y., 2022. Hierarchical multilabel ship classification in remote sensing images using label relation graphs. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv URL https://arxiv.org/abs/2003.10555.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. Vol. 1 (Long Short Pap.). pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv URL https://api.semanticscholar.org/CorpusID:225039882.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 2650–2658.

Futerman, S.I., Laor, Y., Eshel, G., Cohen, Y., 2023. The potential of remote sensing of cover crops to benefit sustainable and precision fertilization. Sci. Total Environ. 891, 164630.

Goel, A., Banerjee, B., Pižurica, A., 2019. Hierarchical metric learning for optical remote sensing scene categorization. IEEE Geosci. Remote. Sens. Lett. 16 (6), 952–956.

Hao, P., Zhan, Y., Wang, L., Niu, Z., Shakir, M., 2015. Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. Remote. Sens. 7 (5), 5347–5369.

He, P., Liu, X., Gao, J., Chen, W., 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv URL https://arxiv.org/abs/2006.03654.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 770–778.

Hu, Q., Yin, H., Friedl, M.A., You, L., Li, Z., Tang, H., Wu, W., 2021. Integrating coarse-resolution images and agricultural statistics to generate sub-pixel crop type maps and reconciled area estimates. Remote Sens. Environ. 258, 112365.

Jiao, L., Sun, W., Yang, G., Ren, G., Liu, Y., 2019. A hierarchical classification framework of satellite multispectral/hyperspectral images for mapping coastal wetlands. Remote. Sens. 11 (19), 2238.

Kang, X., Hong, Y., Duan, P., Li, S., 2024. Fusion of hierarchical class graphs for remote sensing semantic segmentation. Info. Fusion 109, 102409.

Kowsari, K., Brown, D.E., Heidarysafa, M., Jafari Meimandi, K., Gerber, M.S., Barnes, L.E., 2017. HDLTex: Hierarchical deep learning for text classification. In: Proc. IEEE Int. Conf. Mach. Learn. Appl. pp. 364–371.

Li, Z., Bao, W., Zheng, J., Xu, C., 2020a. Deep grouping model for unified perceptual parsing. In: IEEE Conf. Comput. Vis. Pattern Recognit pp. 4052–4062.

Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L., 2020b. RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. Sensors 20 (6), 1594.

Li, Z., Fan, C., Zhao, Y., Jin, X., Casa, R., Huang, W., Song, X., Blasch, G., Yang, G., Taylor, J., Li, Z., 2024. Remote sensing of quality traits in cereal and arable production systems: A review. Crop. J. 12 (1), 45–57.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE Int. Conf. Comput. Vis. pp. 9992–10002.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. URL https://api.semanticscholar.org/CorpusID:198953378.

Ma, Y., Liu, X., Zhao, L., Liang, Y., Zhang, P., Jin, B., 2022. Hybrid embedding-based text representation for hierarchical multi-label text classification. Expert Syst. Appl. 187, 115905.

Nickel, M., Kiela, D., 2017. Poincaré embeddings for learning hierarchical representations. In: Proc. 31st Int. Conf. Neural Inf. Process. Syst.. pp. 6341–6350.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. URL https://api.semanticscholar.org/CorpusID:160025533.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 1–67.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. Med. Image Comput. Comput.-Assist. Intervent., Vol. 9351. pp. 234–241.

Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS Int. J. Geoinf. 7 (4), 129.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE Int. Conf. Comput. Vis. pp. 618–626.

Shang, J., Liu, J., Ma, B., Zhao, T., Jiao, X., Geng, X., Huffman, T., Kovacs, J.M., Walters, D., 2015. Mapping spatial variability of crop growth conditions using RapidEye data in Northern Ontario, Canada. Remote Sens. Environ. 168, 113–125.

Sinha, K., Dong, Y., Cheung, J.C.K., Ruths, D., 2018. A hierarchical neural attention-based text classifier. In: Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. pp. 817–823.

Sulla-Menashe, D., Gray, J.M., Abercrombie, S.P., Friedl, M.A., 2019. Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product. Remote Sens. Environ. 222, 183–194.

Sykas, D., Papoutsis, I., Zografakis, D., 2021. Sen4AgriNet: A harmonized multi-country, multi-temporal benchmark dataset for agricultural earth observation machine learning applications. In: Proc. IEEE Int. Geosci. Remote Sensing Symp. pp. 5830–5833.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3D medical image analysis. In: IEEE Conf. Comput. Vis. Pattern Recognit. pp. 20698–20708.

Tarasiou, M., Chavez, E., Zafeiriou, S., 2023. ViTs for SITS: Vision transformers for satellite image time series. In: IEEE Trans. Geosci. Remote Sens. pp. 10418–10428.

Tarasiou, M., Guler, R.A., Zafeiriou, S., 2022. Context-self contrastive pretraining for crop type semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Turkoglu, M.O., D'Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J.D., 2021. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. Remote Sens. Environ. 264, 112603.

Wang, W., Zhou, T., Qi, S., Shen, J., Zhu, S.-C., 2022. Hierarchical human semantic parsing with comprehensive part-relation modeling. IEEE Trans. Pattern Anal. Mach. Intell. 44 (7), 3508–3522.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 6230–6239.

Zheng, Y., Dong, W., ZhipingYang, Lu, Y., Zhang, X., Dong, Y., Sun, F., 2024. A new attention-based deep metric model for crop type mapping in complex agricultural landscapes using multisource remote sensing data. Int. J. Appl. Earth Obs. 134, 104204.