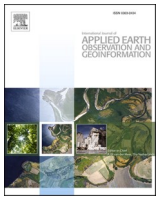
International Journal of Applied Earth Observations and Geoinformation 108 (2022) 102744

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



A hierarchical category structure based convolutional recurrent neural network (HCS-ConvRNN) for Land-Cover classification using dense MODIS Time-Series data

Jiayi Li [a], Ben Zhang [a], Xin Huang [a,b,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China

ARTICLE INFO

Keywords:
Land-cover classification
Hierarchical classification
Convolutional recurrent neural network
Time series
MODIS

ABSTRACT

Hierarchical classification of land cover can be used to describe the Earth's surface with different scales and properties. However, existing studies have rarely considered hierarchical information for land-cover classification, and have ignored dependencies in the hierarchical structure. In this study, we propose a hierarchical category structure-based convolutional recurrent neural network (HCS-ConvRNN). The HCS-ConvRNN method constrains the input through the leaf node of the hierarchical structure based input layer, and then constructs the dependencies among different layers in a top-down manner, in order to classify the pixels into the most relevant classes in a layer-by-layer manner. A total of 219 Moderate Resolution Imaging Spectroradiometer (MODIS) images of China from 2015 to 2017, at a 5-day interval, were used in the reported experiments. It is shown that: 1) the results of HCS-ConvRNN have rich spatial details; 2) the accuracy at each level of HCS-ConvRNN is better than that of MOD12Q1; and 3) generally HCS-ConvRNN can obtain a better classification performance than other networks such as the convolutional neural network (CNN) and gated recurrent unit (GRU). In summary, the proposed HCS-ConvRNN method can effectively achieve hierarchical land cover classification, and has the potential for accurate land cover classification at a large scale.

1. Introduction

Land cover is an important variable when investigating the properties of the Earth's surface. Describing the properties of different land-cover classes and producing accurate land-cover maps is essential for the understanding of global environmental change (Gomez et al., 2016). To date, a series of land-cover products have been produced at different spatial and temporal scales (ESA., 2017; Huang et al., 2021a; Yang and Huang, 2021). In the research related to global climate / environmental change, land-cover products in a wide range of temporal and spatial scales are still very important (Huang et al 2021b).

Time series MODIS images are often used in land cover dynamic monitoring, vegetation dynamics and so on. For example, global land cover types were mapped by MODIS Land Cover Type Product (MOD12Q1) (Sulla-Menashe et al., 2019), thus the information of different dimensions of land cover, land use, and surface hydrology can be hierarchically delineated. The upper layers of the hierarchical category contain information related to the biotic and abiotic land-surface features at a large scale, e.g., land-cover can be classified into barren, vegetation, water, and permanent snow/ice at the top layer of MOD12Q1. On the other hand, the deeper layers of the hierarchical category distinguish the land-cover attributes related to vegetation, landforms, etc. For instance, vegetation can be further classified as shrubs, forests, etc. (Sulla-Menashe et al., 2011). Therefore, classifying land cover under a hierarchical classification system and generating classification maps of different levels is an appropriate way to improve the ecological significance of each class, and can provide more flexible environmental model parameters for global change studies.

The rest of this paper is organized as follows. Section 2 reviews the related classification methods. Section 3 introduces the materials used in this study. Section 4 describes the proposed HCS-ConvRNN method. The results are presented in Section 5. Section 6 includes discussions and comparison with other algorithms. Finally, conclusions are given in Section 7.

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China; State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China.
E-mail address: xhuang@whu.edu.cn (X. Huang).

https://doi.org/10.1016/j.jag.2022.102744
Received 9 January 2022; Received in revised form 1 March 2022; Accepted 10 March 2022
Available online 16 March 2022

2. Related work

Based on whether to consider the hierarchical relationship of land-cover in classification, the existing land-cover classification studies can be divided into two groups. One of the land-cover classifications is regarded as a black box without considering the hierarchical relationship. For instance, the supervised classifiers in machine learning, such as SVM and RF, are used to identify all interested classes independently (Fenske et al. 2020). Forcibly mixing different levels of land-cover can lead to overestimation of high-level abstract land-cover (Gong et al., 2013) and logical contradiction between different levels of land-cover categories (Sulla-Menashe et al., 2011). Alternatively, existing studies concerned hierarchical structure information can be categorized into two groups, i.e., bottom-up and top-down paradigms. The latter directly predicts each parcel with a bottom class label, and then assigns associated labels at higher layers according to the bottom-to-top mapping (Ma et al., 2019b; Sulla-Menashe et al., 2011). Because there are many categories at the bottom level, the problem of samples imbalance is serious, and the performance of classifiers may degrade with the increase of the number of categories, leading to inferior interpretation accuracy (Gong et al., 2013; Zhang et al., 2021). The latter approach first classifies the parcels into the top class level (e.g., Coastal wetland), and then discriminates these parcels into its sub-class labels (e.g., Lagoon). Decision tree and its variants (Mao et al. 2020; Zhao et al. 2021) are the most popular classifiers for the latter approach, since the category hierarchical information as well as class-sensitive rules can be easily modeled. Nevertheless, both paradigms have only focused on the local layers of the category structure, and they have ignored the transmission and dependencies among the different layers of the whole hierarchy. For instance, the determination of a class is not only affected by the parent class, but also influence the child classes. Consequently, it is important to utilize the associations and dependencies among the different layers in the hierarchical structure to improve the accuracy of hierarchical classification. However, little consideration has been given to this in the existing research.

With the rapid development of deep learning in the field of remote sensing, great success has been achieved in remote sensing image analysis tasks such as land-use and land-cover classification and target detection (Li et al., 2019; Ma et al., 2019a; Yuan et al., 2020, Wambugu et al. 2021). However, to date, very few studies have applied deep learning to address the issue of hierarchical classification. Gbodjo et al. (2020) proposed a hierarchical pre-training strategy, which involved training the network from the top layer of the hierarchy and then training the next layer using the previously learned weights. This hierarchical pre-training strategy allowed the model to first focus on the high-level classification problems (such as crop and non-crop), and then gradually adapted to the next layer corresponding to more detailed and complex classification problems. However, as mentioned above, these methods cannot predict all the classes in the whole hierarchy simultaneously, and they ignore the transmission and dependencies among different layers, and also overlook the impact of each layer's result on the overall structure. By assuming that each pixel should have a label at each category level, Turkoglu et al. (2021) designed a data-driven three level hierarchical network. In this network, the sematic feature of each level was solely learned from the samples of current level, and then was copied as input feature to next stage. It is noted that the assumption is not suitable for the hierarchical land cover task. For instance, as seen in Fig, 1, a Tree open sample only has labels at layer 1 and 2 (but does not have labels at the subsequent layers). In this case, this hierarchical network may import wrong information to the subsequent category layers.

Recurrent neural networks (RNNs) have the ability to transfer information between layers. In terms of the network structure, RNNs can memorize the information in the previous layer to affect the output of the subsequent layer, so that the output of the network is not only related to the current input, but also to the output of the previous stage (Connor et al., 1994; LeCun et al., 2015). Benefiting from this merit of RNN, Huang et al. (2019) proposed the hierarchical attention-based recurrent neural network (HARNN) for text classification. The key module of HARNN was the hierarchical attention-based memory (HAM), which was designed to construct the dependencies between text semantic representations and classes, and to transfer the text semantic representations of the corresponding layer to the next layer. In such a way, HARNN classified all the classes of each level in the whole hierarchical structure. However, the RNN structure should be potential but has been not used for the hierarchical land cover classification. There are two reasons that limit the application of HARNN (Huang et al. 2019) for hierarchical land cover classification. First, the core module of HARNN (i.e., the recurrent operation) is only designed for capturing the hierarchical category relationship, but the phenological information, which is important for land cover classification (e.g., vegetation), is ignored. Second, the assumption of HARNN is that each sample has a label at each category layer, which is same to that in Turkoglu et al. (2021), and is not suitable for the hierarchical land cover task.

In this context, the objective of this research is to address the task of land-cover classification under a sophisticated hierarchical classification system. Based on dense MODIS time-series features, a hierarchical category structure based convolutional recurrent neural network (HCS-ConvRNN) is proposed. HCS-ConvRNN can realize information transmission among different layers of the hierarchical category structure through the RNN. Its novelty is that the association and dependencies among the hierarchies are considered for land-cover classification. Furthermore, HCS-ConvRNN can obtain the classification results of all the classes in the hierarchical category structure, which can be projected to different classification schemes for various application requirements.

3. Materials

3.1. Study area

The study area covers China, with an area of approximately 9.6 million square kilometers, spanning tropical, subtropical, warm temperate, middle temperate, and sub-frigid climate zones. China contains almost all the land-cover classes in the existing classification systems, which makes it suitable for the study of land-cover classification with a sophisticated hierarchical classification system.

3.2. MODIS data

The MODIS Nadir Bidirectional Reflectance Distribution Function Adjusted Reflectance (NBAR) product (MCD43A4) is the main input of the land-cover classification. The MCD43A4 provides cloud-screened and atmospherically corrected daily surface reflectance, which can capture more information about seasonal vegetation dynamics and rapid land-surface changes (Wang et al., 2018). Thus, MCD43A4 data from 2015 to 2017 (i.e., 1096 images in total) were used to describe the phenological information. To reduce the computational cost, the proposed work directly chose the daily MCD43A4 data at a 5-day interval, which were then smoothed and gap-filled using penalized splines to obtain high-quality and minimal-omission data. In this way, a total of 219 images were employed. For each period in the dense time series, the mean values of the 3 years (2015, 2016, 2017) were calculated.

3.3. Hierarchical classification system

The hierarchical classification system adopted in this paper is the land-cover classification system of the MOD12Q1 product. The classification system is based on the Land Cover Classification System (LCCS) from the Food and Agricultural Organization (FAO), which can reflect different dimensions of land cover, land use, and surface hydrology (Sulla-Menashe et al., 2019). The hierarchical category structure of the classification system (Fig. 1) consists of four layers, including a total of

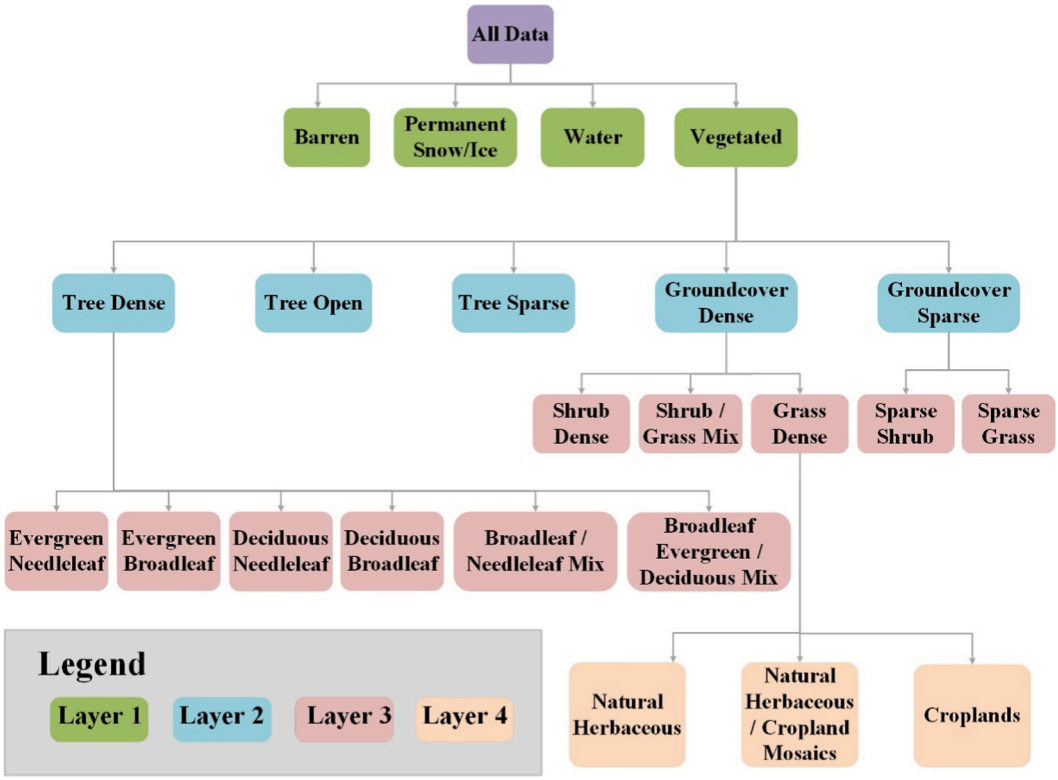


Fig. 1. The hierarchical classification system.

23 land-cover classes. For each layer, we define the label at the bottom layer of the hierarchical category structure as the leaf node class. For example, the Barren, Tree Open, Shrub Dense, and Croplands classes in Fig. 1 are the leaf node classes at the first, second, third, and fourth layers, respectively.

3.4. Training and test samples

Many studies have attempted to extract training samples from existing land-cover products in order to mitigate the workload of sample collection (Xie et al., 2019; Zhang and Roy, 2017). With reference to the methods of Xie et al. (2019), in this study, a series of operations, including temporal constraint, spatial filtering, spectral filtering, and manual checking, were used for MOD12Q1 to control the confidence and reliability of training data.

Based on the LACO-Wiki land-cover validation platform (See et al., 2017), the test samples were selected by referring to the high-resolution Google images, vegetation index curves, geotagged photos, and spectral features. In the experiments, a total of 900 test samples (250 m × 250 m for each) were used in the study area. The number of training and test samples for each category is shown in Table 1, and their distribution is shown in Fig. 2.

4. Methodology

4.1. Hierarchical category structure based convolutional recurrent neural network (HCS-ConvRNN)

The HCS-ConvRNN method proposed in this paper (Fig. 3) mainly consists of three parts:

1) the feature representation layer (FRL) to extract the dense temporal features of MODIS images;

2) the leaf node discriminant function to determine whether the sample is located in the leaf node of the hierarchical category system (LNIL see Fig. 3b);

Table 1
Training samples and test samples.

Layer	Class	Training samples	Test samples
L1	Barren	8732	86
	Permanent Snow /Ice	6335	12
	Water	7938	54
	Vegetated	125,479	462
L2	Tree Dense	43,732	220
	Tree Open	5901	28
	Tree Sparse	15,405	14
	Groundcover Dense	48,732	158
	Groundcover Sparse	11,709	42
L3	Evergreen Needleleaf	4306	9
	Evergreen Broadleaf	15,605	37
	Deciduous Needleleaf	4733	29
	Deciduous Broadleaf	6967	78
	Broadleaf/Needleleaf Mix	7818	27
	Broadleaf Evergreen/Deciduous Mix	4303	40
	Shrub Dense	7286	28
	Shrub /Grass Mix	7223	13
	Grass Dense	34,223	117
	Sparse Shrub	7513	25
	Sparse Grass	4196	17
L4	Natural Herbaceous	16,605	22
	Natural Herbaceous/Croplands Mosaics	7232	34
	Croplands	10,386	61

3) the hierarchical attention-based convolutional recurrent layer (HACRL), and Fig. 3c describes the association between the features and the land-cover classes of each layer in a top-down manner:

4.1.1. The feature representation layer (FRL)

FRL includes three steps: 1) MCD43A4 data are used to extract the dense temporal features; 2) the dense temporal features are enhanced; and 3) the hierarchical category structure is input into the network.

The features extracted for each sample involved: 1) the spectral

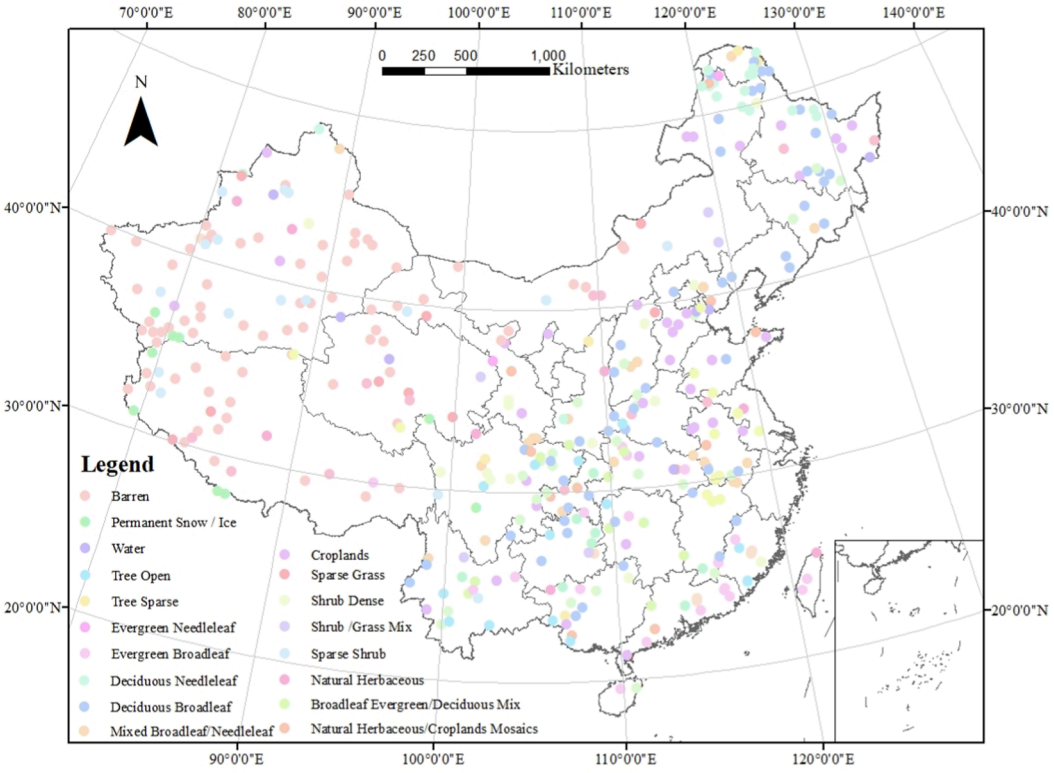


Fig. 2. The distribution of test samples.

values in the B1, B2, and B4-B7 bands; 2) the two-band enhanced vegetation index (EVI2) (Jiang et al., 2008); 3) the NDSI (Salomonson and Appel, 2004); 4) the NDWI (Gao, 1996); and 5) two versions of the normalized difference infrared index (NSII1, NSII2) (Ji et al., 2011). Subsequently, the features of all the time periods were combined into the$V_{data} = (t_1, t_2, \cdots, t_N) \in R^{N\times F}$, where N denotes the total number of time periods ($N = 73$ in this study) and F denotes the dimension of the features at each period ($F = 11$ in this study). An unsupervised neural network, bi-directional long short term memory (Bi-LSTM) (Graves and Schmidhuber, 2005), was conducted to enhance and transfer the dense temporal features $V_{data} = (t_1, t_2, \cdots, t_N) \in R^{N\times F}$ to$V = (\overset{\cdot}{t_1}, \overset{\cdot}{t_2}, \cdots, \overset{\cdot}{t_N}) \in R^{N\times 2u}$, where u represents the hidden neuron dimension of Bi-LSTM. Bi-LSTM not only learns the long temporal dependencies between the input dense temporal features, but also learns the contextual information in both forward and backward directions simultaneously, which is conducive to enhancing the semantic representation of the samples (Graves and Schmidhuber, 2005).

We embedded the land-cover hierarchical category structure γ into the network in the form of a matrix$S = (S^1, S^2, \cdots, S^L)$, to obtain the attention score of different classes in each layer. The matrix S consists of a category embedding matrix S^l for each layer. The way to obtain the attention score of the proposed network $S^l \in R^{C^l\times d_a}$ can be referred to HRANN (Huang et al. 2019), where C^l is the total number of classes in layer l, and d_a is the number of output channels of the convolutional layer in the convolutional category attention module of the HACM module. At the beginning, the attention score of each class in each layer was randomly initialized to a small positive number, and the sum of all attention scores in each layer was 1. Afterwards, during the training process, the attention score of each class in each layer was gradually learned by the back propagation optimization.

4.1.2. The leaf node of the hierarchical structure based input layer (LNIL)

When processing the samples of leaf nodes, transferring their dependency to the subsequent layers may introduce wrong information. Thus, we designed the LNIL layer in HCS-ConvRNN, and in this way, the subsequent layers can be skipped when dealing with the samples of leaf nodes. During the network training, the current loss value is determined by the results of the previous and current layers, and the gradients of these layers are fed back, which is independent of the parameters of the subsequent layers. Therefore, the input is constrained by the LNIL to determine whether the label of the sample is the leaf node class of the current layer. If true, the result is directly output after the HACM module; otherwise, the information is passed to the next layer.

4.1.3. The hierarchical Attention-based convolutional recurrent layer (HACRL)

The HACRL labels the most relevant classes layer by layer in a top-down manner based on the hierarchical classification system. HACRL consists of a stack of HACM modules in each layer. The HACM module was proposed for capturing the association between dense temporal features and classes at different layers, conveying the information among the layers, and predicting the classification results of each layer. Fig. 4 shows the details of the HACM module, which consists of three components: 1) the class convolutional attention module (CCAM); 2) the class prediction module (CPM); and 3) the class transmission module (CTM). HACM module can be considered as an improved version of the HAM unit (Huang et al., 2019) by importing convolution block into CCAM and CPM to further extract time-varying information from dense time series. The benefits of this improvement are described below, and verified in a series of ablation experiments (see Section 6.1).

1) **Class Convolutional Attention Module (CCAM)**: CCAM establishes the association between the dense temporal features and hierarchical classes, and then extracts the current layer information based on this association. Thus, the inputs of CCAM are the enhanced features V and the dependencies ω^{l-1} passed down from the previous layer. The outputs of the CCAM are the associated sample-category representation r^l_{att} and the attention matrixW^l_{att}. The mechanism of CCAM is introduced as follows. Firstly, to highlight the l-th layer featuresV^l, the enhanced

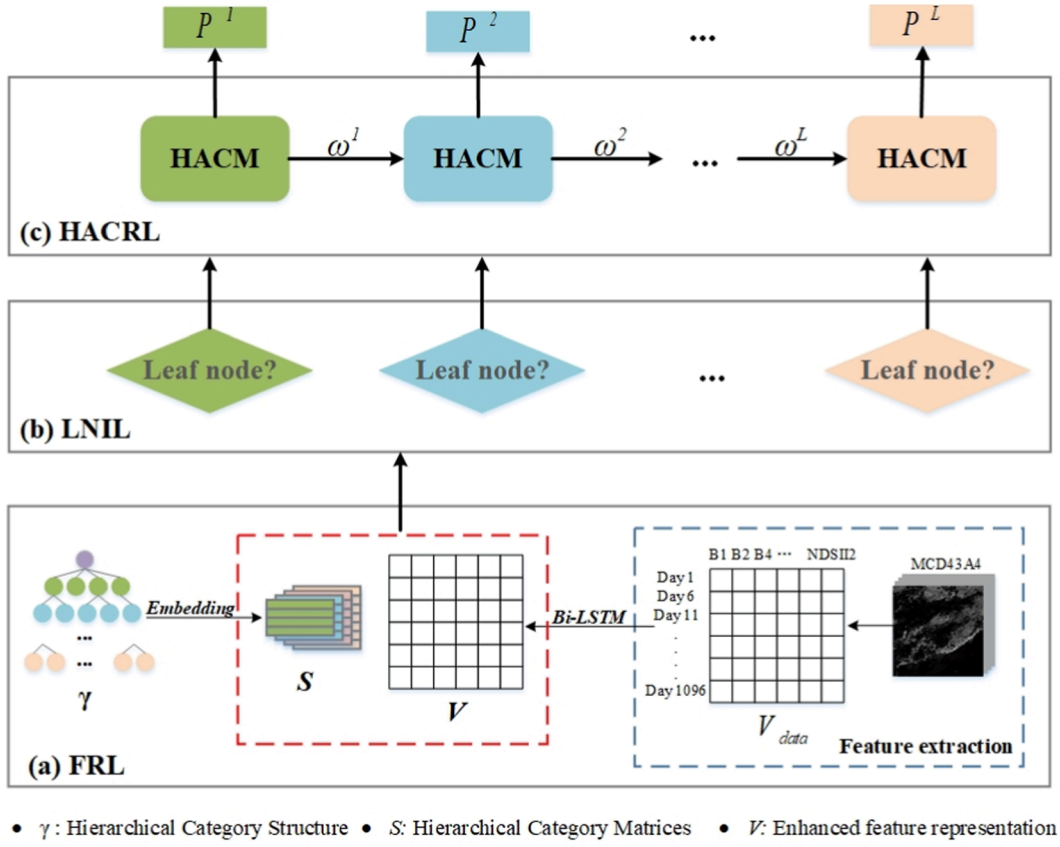


Fig. 3. The hierarchical category structure based convolutional recurrent neural network: (a) The feature representation layer (FRL); (b) the leaf node of the hierarchical structure based input layer (LNIL); and (c) the hierarchical attention-based convolutional recurrent layer (HACRL). V_{data} denotes the extracted dense temporal features, HACM represents the hierarchical attention-based convolutional memory module, ω denotes the dependency information passed between layers, and P is the probability of each layer's output.

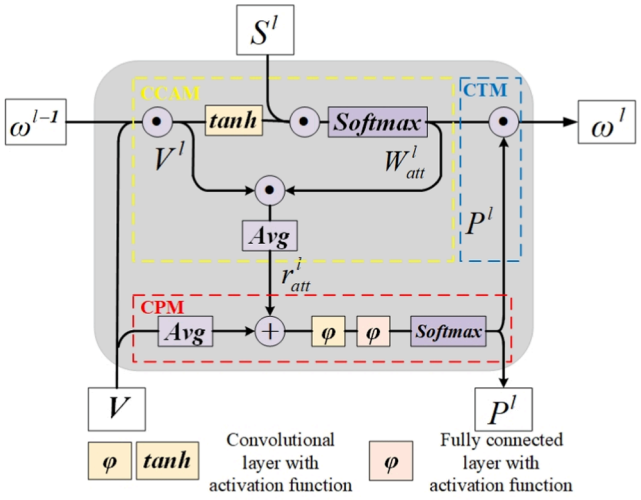


Fig. 4. The details of the proposed HACM module.

features V are multiplied by the dependencies $\omega^{l-1} \in R^{N\times 2u}$:

$$V^l = \omega^{l-1} \bullet \mathrm{V}, \mathrm{V}^l \in \mathrm{R}^{\mathrm{N}\times 2\mathrm{u}} \tag{2}$$

where • indicates the entry-wise product operation. As convolution operation can effectively extract local time-varying feature from the dense time series images, 1D convolution is then performed on V^l. Subsequently, the result is activated by the tanh function to extract the time-varying information O^l:

$$O^l = tanh\left(W_a^l \otimes V^l\right) \tag{3}$$

where $\otimes$ is the 1D convolution operation and W_a^l is a convolution kernel of size k_a and channel d_a.

Different land-covers show different spectral and temporal information. Therefore, assigning different weights through the attention mechanism can effectively highlight the characteristics of each category (Huang et al., 2019; Lin et al., 2017). In the proposed framework, the attention mechanism is used to capture the associations between the sample features and each category in the hierarchy. Please note that S^l is used to obtain the attention weights of the different categories in layer l. In the attention mechanism, the category embedding matrix S^l and the activation O^l are multiplied, and then the attention weights W_{att}^l are computed through the softmax function:

$$W_{att}^l = softmax\left(S^l \bullet \mathrm{O}^l\right) \tag{4}$$

where $W_{att}^l = (W_1^l, W_2^l, \cdots, W_{C^l}^l) \in R^{C^l\times N}$, and W_i^l denotes the attention score of the i-th class in the current layer, where each element in this vector represents its contribution to the i-th class at each period in the dense time series.

Subsequently, the weighted category association representation is obtained by multiplying the attention weight matrix W_{att}^l and the features V^l. The results are then averaged and pooled to obtain the associated sample-category representation for the l-th layer $r_{att}^l \in R^{2u}$:

$$r_{att}^l = avg\left(W_{att}^l \bullet \mathrm{V}^l\right) \tag{5}$$

2) **Class Prediction Module (CPM)**: CPM makes a prediction for the current layer. The features of N periods are combined into an average embedding $\overline{V} = avg(t_1^{'}, t_2^{'}, \cdots, t_N^{'}) \in R^{2u}$ to obtain the overall semantic representation. The purpose of the CPM is to combine the overall semantic representation $\overline{V}$ and the associated sample-category representationr_{att}^{l}, in order to introduce the information from the previous layer and make predictions for the classes of each layer. In current layer, the CPM consists of a convolutional layer, a pooling layer, and a fully connected layer, to better capture the local features from previous feature transformation. The probabilities are then computed after activated by the softmax function:

$$x_c^l = \varphi\left(W_c^l \otimes \left[\overline{V} \oplus r_{att}^l\right] + b_c^l\right) \tag{6}$$

$$f^l = \varphi\left(W_f^l \bullet \mathrm{pooling}_{\mathrm{ps}}\left(\mathrm{x}_{\mathrm{c}}^{\mathrm{l}}\right) + \mathrm{b}_{\mathrm{f}}^{\mathrm{l}}\right) \tag{7}$$

$$P^l = softmax\left(W_s^l \bullet \mathrm{f}^{\mathrm{l}} + \mathrm{b}_{\mathrm{s}}^{\mathrm{l}}\right) \tag{8}$$

where W_c^l is the convolution kernel of size k_c and channeld_c, $b_c^l \in R^{\mathrm{d_c}\times 1}$ is the corresponding bias, φ is the rectified linear unit (RELU) activation function, and pooling denotes the pooling operation. W_f^l is the weight of the fully connected layer, $b_f^l \in R^{v\times 1}$ is the corresponding bias, and v denotes the neuron dimension of the fully connected layer. $W_s^l \in R^{C^l\times v}$ is the weighting matrix that connects the fully connected activation and the class output units, and $b_s^l \in R^{C^l\times 1}$ is the corresponding bias. The category with the highest probability in P^l is the output label at the current layer.

3) **Class Transmission Module (CTM)**: CTM transmits the dependency between the features and the classes to the next layer. In the l-th layer, different classes information contributes differently to the prediction, which can be used as a trade-off parameter to modify the attention matrix weightsW_{att}^l. As shown in (9), W_{att}^l is weighted by P^l to obtain the weighted category attention matrixK^l:

$$K^l = broadcast\left(P^l\right) \bullet \mathrm{W}_{\mathrm{att}}^{\mathrm{l}} \tag{9}$$

where$K^l = (k_1^l, k_2^l, \cdots, k_{C^l}^l) \in R^{C^l\times N}$, k_i^l indicates the weighted attention score of the i-th category, and broadcast(.) enables matrices with different shapes to have compatible shapes for the arithmetic operations (i.e., the entry-wise products).

Average pooling is then performed in dimensionC^l. Next, this average pooling result is broadcast toω^l, which has the same structure as V, to transmit the category-related information of the current layer:

$$\omega^l = broadcast\left(avg\left(K^l\right)\right) \tag{10}$$

where$\omega^l = (\omega_1^l, \omega_2^l, \cdots, \omega_N^l) \in R^{N\times 2u}$, and $\omega_i^l \in R^{2u}$ measures the association between the whole previous layer and the i-th period in the dense time series. Finally, ω^l is passed to the next layer.

4.2. Loss function

We develop a multi-task loss function to efficiently train the proposed network. To reduce the effect of the category imbalance (see Fig. 1), the local loss values of each layer are weighted according to the number of samples of each category. Moreover, in order to reduce the logical errors of the layer, i.e., sample labels that do not strictly obey the relationships of the classes and subclasses at each layer, a logic loss function is added. The logic loss function is based on the prediction labelY_P^{l-1}. The theoretically logical prediction result of this layer is inferred according to the logical transformation relationship matrix$w_{logic} \in R^{C^l\times C^{l-1}}$:

$$w_{logic(i,j)} = \begin{cases} 0, iisnotthesubclassofj \\ 1, iisthesubclassofj \end{cases} \tag{11}$$

When compared with the actual label Y^l of the current layer, if the logical result is the same as the actual one, the logical loss value L_{logic}^l at that layer is set to 0; otherwise, the value of L_{logic}^l is increased, as shown in (12):

$$L_{logic}^l = \begin{cases} 0, Y_P^{l-1} \bullet w_{logic} = Y^l 1, Y_P^{l-1} \bullet w_{logic} \neq Y^l \end{cases} \tag{12}$$

The final loss function in each layer of the classification task is shown in (13):

$$Loss^l = \frac{n_{total}}{n_{class}} \sum_{l=1}^{L} \varepsilon\left(P^l, Y^l\right) + L_{logic}^l\left(Y_P^{l-1} \bullet \mathrm{w}_{\mathrm{logic}}, \mathrm{Y}^{\mathrm{l}}\right) \tag{13}$$

where Y^l is the binary label vector. The cross-entropy $\varepsilon(\hat{\mathrm{A}}\cdot, \hat{\mathrm{A}}\cdot)$ is used to minimize the local loss. n_{total} is the number of total samples, and n_{class} is the number of samples of each class at l-th layer.

In addition, to address the category imbalance issue, a multi-task loss function is used to balance the weights of each layer and combine the classification losses of each layer so as to achieve a better performance (Kendall et al., 2018). The final loss function can be written as:

$$\mathrm{Loss} = \sum_{l=1}^{L} \left(\frac{1}{\sigma_l^2} Loss^l + log\sigma_l\right) \tag{14}$$

where σ_l is the noise parameter of each layer, which is used to characterize the inter-task uncertainty.

5. Experiments and results

5.1. Parameter setting

According to the suggestion by (Huang et al., 2019), we used random search to select the main parameters of the network, where the hidden layer dimension (u) was set to 128 in the Bi-LSTM layer, the size (k_a) of the convolution kernel W_1^l was set to 5, the number of channels (d_a) was set to 100 in the CCAM, the size (k_c) of the convolution kernel W_c^l was set to 5, the number of channels (d_c) was set to 64 in the CPM, and the dimension of all the fully connected layer cells (v) was set to 256. We initialized the parameters in HCSConv-RNN with a truncated normal distribution with a standard deviation of 0.1. For the training of the HCSConv-RNN method, we used the Adam optimizer with a learning rate of 1×10^{-3}, and also used dropout with a dropout rate of 0.5 to prevent overfitting and gradient clipping. All the networks were run on a desktop computer using TensorFlow-GPU-1.14 with an Intel Core i9-7980X CPU (2.60 GHz), 112-GB RAM, and a 11-GB GeForce RTX 1080 Ti GPU.

5.2. Hierarchical classification results

The land-cover classification maps for the four levels are presented in Fig. 5. The OA and kappa coefficients of the land cover classification results of each level were also calculated (see Table 2). It can be seen that with the increase of classification levels, finer grained classes are more difficult to identify. As claimed by Sulla-Menashe et al. (2019), large-scale land cover classification is challenging owing to the spectral similarity of mixed land cover categories, and the classification accuracy for the deep level complex categories is low.

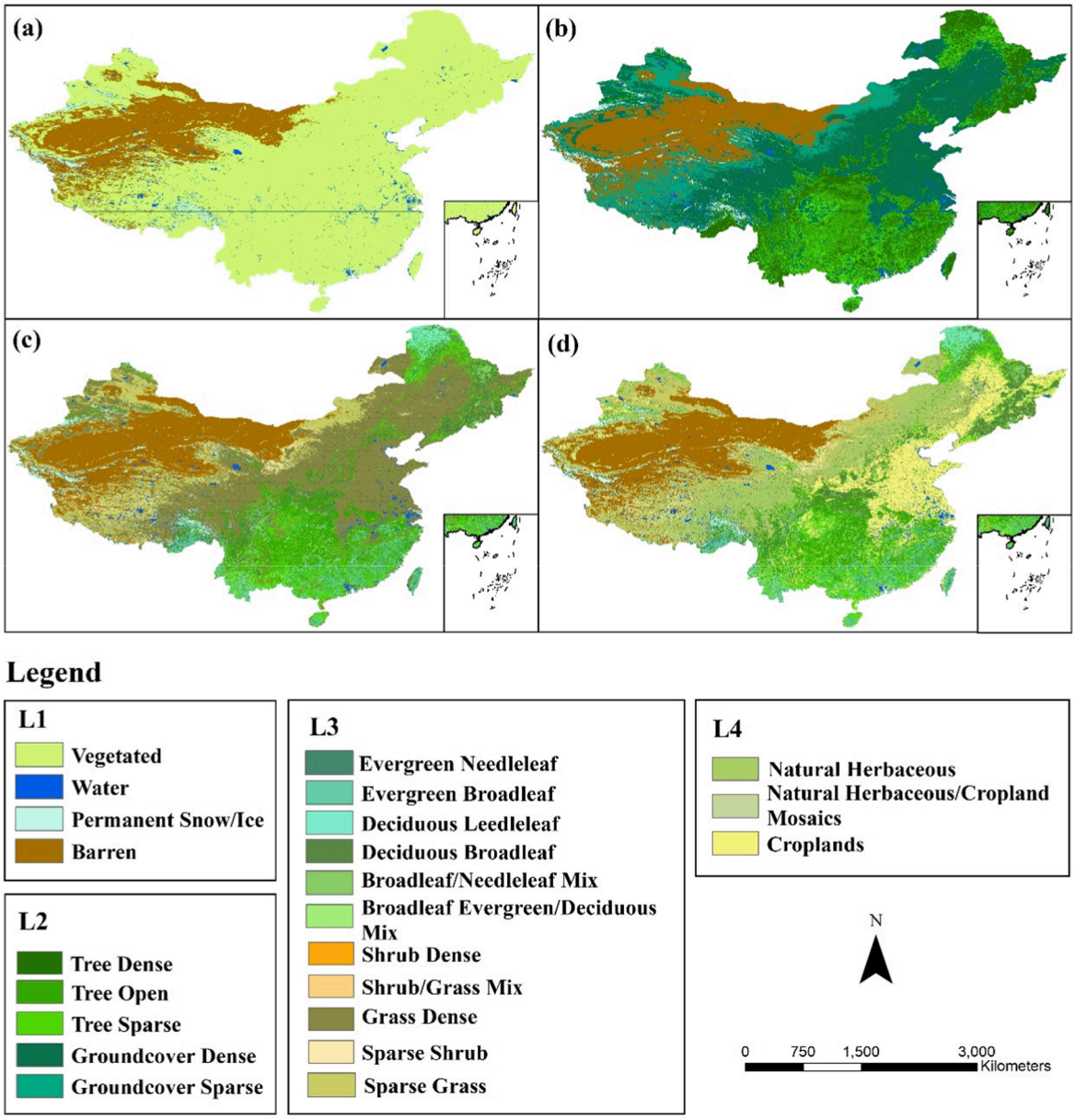


Fig. 5. Results of the land-cover classification: (a) level 1; (b) level 2; (c) level 3; (d) level 4.

Table 2
Accuracy Assessment Results.

Level	Index	HCS-ConvRNN
L1	OA	0.9218
	Kappa	0.8009
L2	OA	0.6172
	Kappa	0.5338
L3	OA	0.4853
	Kappa	0.4343
L4	OA	0.4527
	Kappa	0.4137

Table 3
Accuracy Comparison Between HCS-ConvRNN and MOD12Q1.

Level	Index	HCS-ConvRNN	MOD12Q1
L1	OA	**0.9218**	0.9201
	Kappa	**0.8009**	0.7984
L2	OA	**0.6172**	0.6156
	Kappa	**0.5338**	0.5256
L3	OA	**0.4853**	0.4592
	Kappa	**0.4343**	0.3994
L4	OA	**0.4527**	0.4087
	Kappa	**0.4137**	0.3640

6. Discussions

6.1. Comparison with MOD12Q1

The accuracy of the land-cover classification results is compared with that of the MOD12Q1 (i.e., the most recent product) in Table 3. The MOD12Q1 product classified each layer of its hierarchical classification system in parallel with a RF classifier. Fig. 6 lists the F1 scores (i.e., the geometric average of user and producer accuracy) for each land-cover class under different levels. In general, the accuracy of the proposed method is higher than that of MOD12Q1 at all the four levels. With the increase of the classification level, the classification system becomes more complex, and the difference between MOD12Q1 and HCS-ConvRNN becomes greater. For instance, in L3 and L4, respectively,

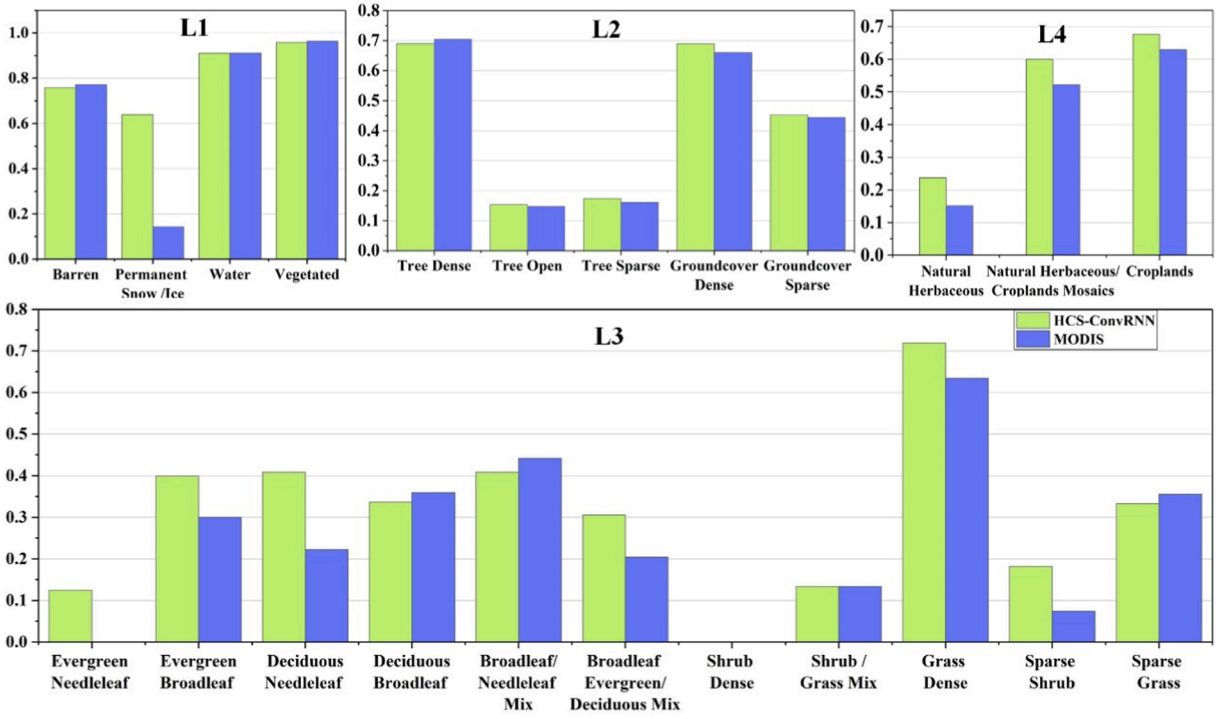


Fig. 6. F1 scores for each land-cover class under the different levels.

HCS-ConvRNN has 2.61% and 4.4% higher OA values and 3.49% and 4.97% higher Kappa values than MOD12Q1.

In addition to the quantitative evaluation, zoomed-in regions under different levels are shown in Fig. 7 to demonstrate more details for the hierarchical classification. In L1, for the Water class, HCS-ConvRNN can extract the small water bodies. In L2, the land-cover classes of both products have similar spatial distribution patterns, but still differ in the details. For example, in Fig. 7(a), HCS-ConvRNN is able to extract Groundcover Sparse, while MOD12Q1 has difficulty in identifying this class. As shown in Fig. 7(b), in the areas containing both forest and farmland, HCS-ConvRNN can effectively distinguish the Tree Sparse class from other groundcover vegetation. However, MOD12Q1 seems insensitive to the distinction between Tree Sparse and low vegetation, and it often misclassifies low vegetation as trees. In L3, there are more omissions for MOD12Q1 in the agroforestry areas (Fig. 7(b)), especially for the class of Grass Dense. MOD12Q1 misclassifies Grass Dense as Tree Sparse in L3 as well. Moreover, as shown in Fig. 7(c), MOD12Q1 has difficulty in detecting the shrub classes (i.e., Shrub Dense, Shrub/Grass Mix, and Sparse Shrub) in L3, but the shrub classes can be identified by HCS-ConvRNN. In L4, MOD12Q1 also present some logical errors. For instance, in Fig. 7(b), some pixels are classified as Croplands by MOD12Q1 in L4 which are identified as Tree Sparse in the previous levels. In contrast, for HCS-ConvRNN, the Tree Sparse class and other low vegetation can be successfully distinguished in the previous levels, which reduces the logical errors and maintains a better logical consistency between the classes in the hierarchical classification system.

6.2. Comparison with related methods

To demonstrate that the HCS-ConvRNN method proposed in this paper shows a superior performance in terms of accuracies in land-cover hierarchical classification, we compared it with other networks. Since the dense temporal feature $V_{data} = (t_1, t_2, \cdots, t_N) \in R^{N \times F}$ used in this study is a two-dimensional feature containing both temporal and spectral information, we considered two deep networks for hyperspectral imagery as comparisons, namely, a convolutional neural network CNN_temp (Hu et al., 2015) and a GRU_temp (Mou et al., 2017). The CNN_temp method used convolution structure to mine local temporal features, and GRU_temp utilized gated recurrent unit structure to mine global temporal features. For a fair comparison, in this experiment, we also trained CNN_temp and GRU_temp with a hierarchical pre-training strategy (Gbodjo et al., 2020), and denoted them as CNN_hiearchy and GRU_hieracrchy, respectively. The OA and the class-average F1 scores of all methods are shown in Fig. 8.

It can be seen that, the classification accuracy of HCS-ConvRNN is higher than that of the CNN_temp method at all levels. With regard to the GRU_temp, except for the OA of L2, HCS-ConvRNN obtains a higher OA than the GRU_temp method in most cases. However, note that OA may be affected or biased by the unbalanced number of samples, and hence, the class-average F1 score for each method is also calculated. In terms of class-average F1 score, HCS-ConvRNN outperforms CNN_temp and GRU_temp. This shows that our method considers the hierarchical relationship between classes and can achieve better accuracy. Similarly, the classification accuracy of HCS-ConvRNN is higher than that of CNN_hierarchy and GRU_hierarchy in most cases. This indicates that our proposed method is better than the existing ones by considering the hierarchical relationship (Gbodjo et al., 2020).

Two additional methods were conducted to analyze the effectiveness of the proposed method (Table 4). Method 1 used the proposed hierarchical deep structure, which was pre-trained on level 1 and then transferred to level 2, 3, and 4. Method 2 directly trained random forest at level 4 using the detailed label. From Table 4, it can be seen that the effectiveness of the proposed HCS-ConvRNN is verified in terms of the accuracy.

6.3. Ablation experiments

To demonstrate the effectiveness of each module of the proposed method, we conducted a series of ablation experiments (Table 5). The ablation experiments include the following parts:

1) HCS-ConvRNN: The proposed method.
2) –MTL: the multi-task loss was replaced by a summation of the cross-entropy loss and the logical loss of each layer.

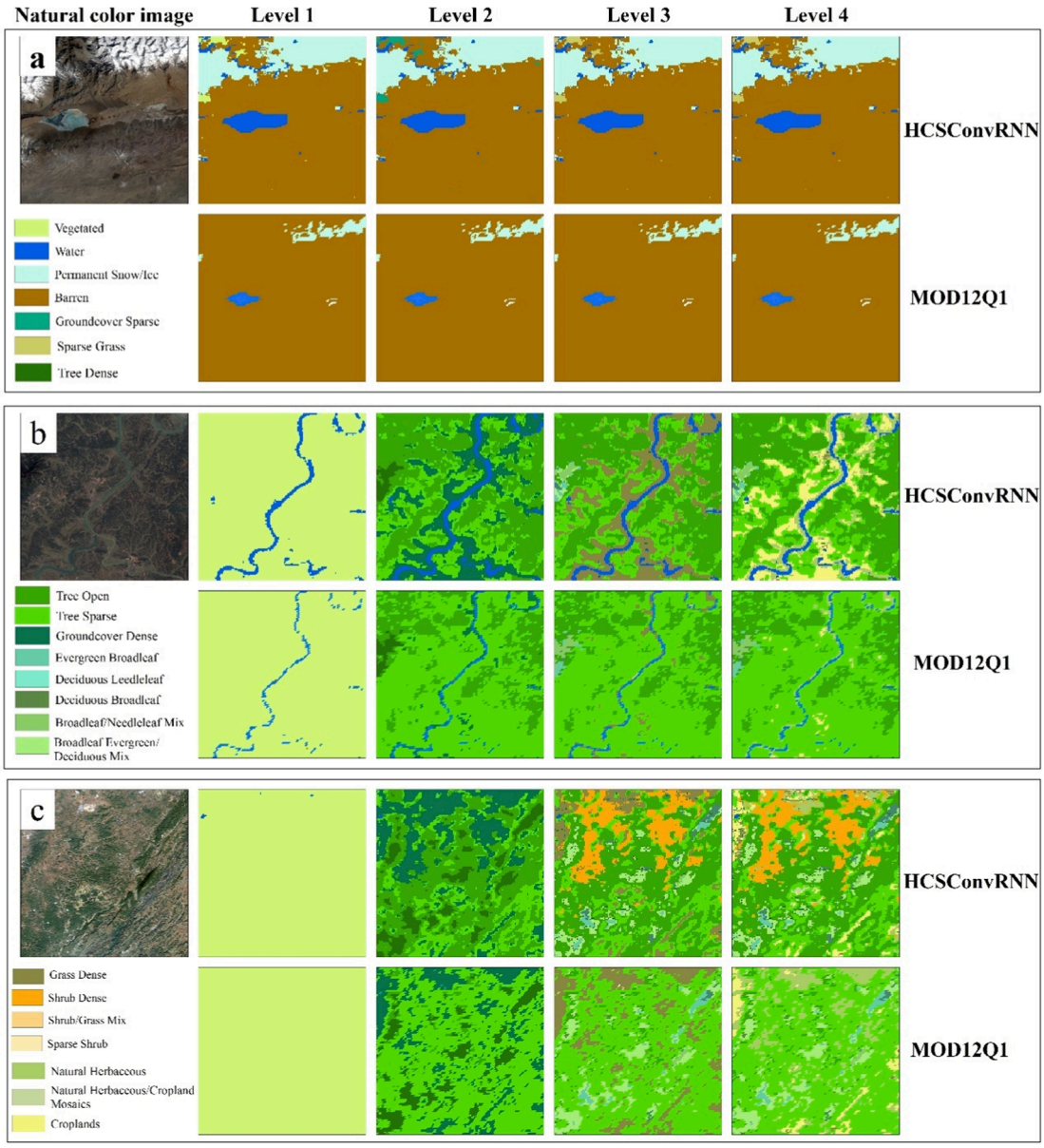


Fig. 7. Zoomed-in regions of the classification results of HCS-ConvRNN and MOD12Q1.

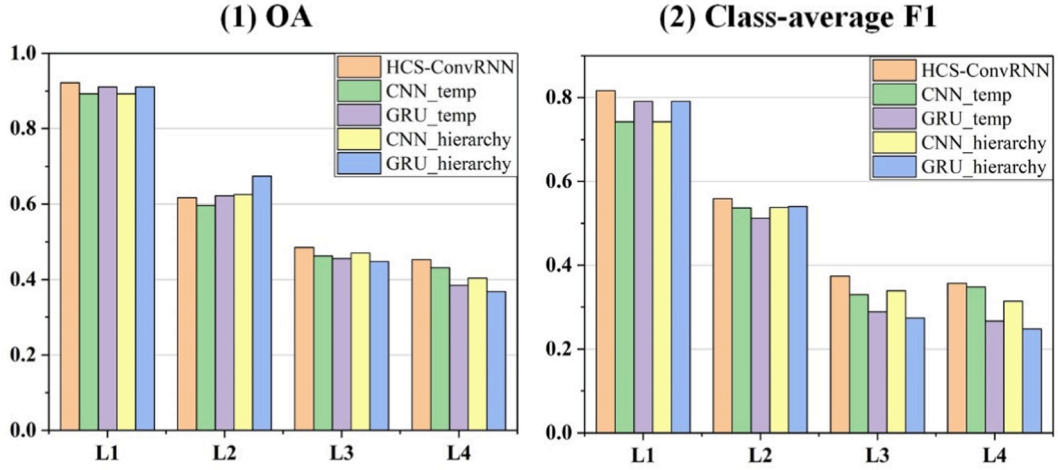


Fig. 8. Accuracy comparison between HCS-ConvRNN and the related networks.

Table 4
Overall Accuracy for HCS-ConvRNN and the Two Comparison methods.

Level	Index	HCS-ConvRNN	Method 1	Method 2
L1	OA	**0.9218**	0.9218	–
	Kappa	**0.8009**	0.8009	–
L2	OA	**0.6172**	0.6140	–
	Kappa	**0.5338**	0.5303	–
L3	OA	**0.4853**	0.4755	–
	Kappa	**0.4343**	0.4241	–
L4	OA	**0.4527**	0.4413	0.4364
	Kappa	**0.4137**	0.4016	0.3956

Table 5
Accuracy Comparison for the Ablation Experiments.

Level	Index	HCS-ConvRNN	-MTL	- CON	-LNI (HARNN)
L1	OA	0.9218	0.9205	0.9141	**0.9335**
	Kappa	0.8009	0.7975	0.7840	**0.8375**
L2	OA	**0.6172**	0.6126	0.5964	0.4440
	Kappa	**0.5338**	0.5307	0.5135	0.3278
L3	OA	**0.4853**	0.4813	0.4732	0.3192
	Kappa	**0.4343**	0.4243	0.4160	0.2435
L4	OA	**0.4527**	0.4489	0.4376	0.1247
	Kappa	**0.4137**	0.4049	0.3923	0.0764

3) –CON: the convolutional and pooling layers in HACM were deleted from 2) and replaced with fully connected layers.
4) –LNI: i.e., HARNN proposed by Huang et al. (2019), the LNIL was further removed from 3)

In general, it can be found that the proposed HCS-ConvRNN method obtains the highest accuracy in most levels, while the recurrent network with the basic hierarchical structure (-LNI) has the lowest accuracy, especially in the higher levels (e.g., L3 and L4).

Specifically, HCS-ConvRNN added a multi-task loss function compared to -MTL, to balance the losses of each layer during the training process, which can improve the classification accuracy at all the layers.

Compared with –CON, -MTL added convolutional layers to improve the accuracy, and the improvement of the classification accuracy in the higher levels is relatively high, indicating that, for hierarchical classification, convolution can better extract the dense temporal features of the classes under deeper layers. The accuracy of –CON is decreased by ~ 1% in the four levels. When compared with HCS-ConvRNN, its OA decreases by 0.77%, 2.08%, 1.21%, and 1.51%, respectively, and Kappa decreases by 1.69%, 2.03%, 1.83%, and 2.14%, respectively.

Lastly, -LNI is the basic hierarchy-based recurrent network, which transmits features directly into the memory units of each layer in the HACRL, without additional filtering, and therefore, it tends to lead to the accumulation of wrong information transmission and misclassification in that layer. For instance, a certain sample only has a label in the first layer, and does not have labels in the next layers, but the features can be still passed into the memory units of the subsequent layers, leading to misclassification. Consequently, for -LNI, the accuracy decreases substantially in the deeper levels. This phenomenon indicates that the LNIL can greatly reduce the misclassification in the hierarchical classification structure, while maintaining the transmission of correct information between layers and reducing the logical errors in the results.

7. Conclusions and outlook

In this study, a hierarchical category structure based convolutional recurrent neural network (HCS-ConvRNN) method is proposed. The HCS-ConvRNN method considers the characteristics of the sample labels under the land-cover class hierarchy, and constrains the input of the network through the leaf node of the hierarchical structure based input layer (LNIL). Subsequently, it constructs dependencies among the different layers in a top-down manner to obtain the classification results for each layer.

Qualitative and quantitative comparisons between HCS-ConvRNN and the MOD12Q1 product showed that HCS-ConvRNN can exhibit richer spatial details. The classification accuracy of HCS-ConvRNN was found better than that of MOD12Q1 at all levels, and the difference became greater as the classification level became deeper. When compared with other networks, i.e., CNN and GRU, HCS-ConvRNN provided better classification performance in most cases. In summary, the HCS-ConvRNN method proposed in this paper can achieve a better performance in hierarchical classification of land cover.

It should be noted that, although the proposed HCS-ConvRNN method can effectively exploit the dependency in a multi-layer category system, the accuracy of the classes (e.g., Shrub Dense and Shrub/Grass Mix) at the deeper layers is still not satisfactory. This suggests that the current data and features are inadequate for a complex classification system, and higher-resolution data could possibly be considered to deal with this issue (Li et al., 2017). Furthermore, the embedding of geographical knowledge and the establishment of a knowledge graph can play a critical role in complex and hierarchical land-cover classification (Lin et al., 2022), which will be considered in our future work.

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.

Acknowledgement

The research was supported by the National Natural Science Foundation of China under Grants 41971295 and 42071311.

References

Connor, J.T., Martin, R.D., Atlas, L.E., 1994. Recurrent neural networks and robust time-series prediction. IEEE Trans. Neural Networks 5, 240–254. https://doi.org/10.1080/01431161.2011.552923.

ESA, 2017. Land cover CCI product user guide version 2.0. Retrieved from: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.

Fenske, K., Feilhauer, H., Förster, M., Stellmes, M., Waske, B., 2020. Hierarchical classification with subsequent aggregation of heathland habitats using an intra-annual RapidEye time-series. Int. J. Appl. Earth Obs. 87, 102036 https://doi.org/10.1016/j.jag.2019.102036.

Gao, B.-C., 1996. NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3.

Gbodjo, Y.J.E., Ienco, D., Leroux, L., Interdonato, R., Gaetano, R., Ndao, B., 2020. Object-based multi-temporal and multi-source land cover mapping leveraging hierarchical class relationships. Remote Sens. 12, 2814. https://doi.org/10.3390/rs12172814.

Gomez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS J. Photogramm. Remote Sens. 116, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, H., Zhu, P., Zhao, Z., Zhang, H., Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, L., Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, J., Chen, J., 2013. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. Int. J. Remote Sens. 34, 2607–2654. https://doi.org/10.1080/01431161.2012.748992.

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18, 602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Hu, W., Huang, Y., Wei, L.i., Zhang, F., Li, H., 2015. Deep convolutional neural networks for hyperspectral image classification. J. Sensors 2015, 1–12.

Huang, W., Chen, E., Liu, Q., Chen, Y., Huang, Z., Liu, Y., Zhao, Z., Zhang, D., Wang, S., 2019. Hierarchical multi-label text classification: an attention-based recurrent network approach. In: Proc. 28th ACM Int. Conf. Info. Knowl. Manage. 1051-1060. Doi: 10.1145/3357384.3357885.

Huang, X., Huang, J., Wen, D., Li, J., 2021a. An updated MODIS global urban extent product (MGUP) from 2001 to 2018 based on an automated mapping approach. Int. J. Appl. Earth Obs. 95, 102255 https://doi.org/10.1016/j.jag.2020.102255.

Huang, X., Li, J., Yang, J., Zhang, Z., Li, D., Liu, X., 2021b. 30 m global impervious surface area dynamics and urban expansion pattern observed by Landsat satellites:

From 1972 to 2019. Sci. China Earth Sci. 64, 1922–1933. https://doi.org/10.1007/s11430-020-9797-9.

Ji, L., Zhang, L., Wylie, B.K., Rover, J., 2011. On the terminology of the spectral vegetation index (NIR - SWIR)/(NIR + SWIR). Int. J. Remote Sens. 32, 6901–6909. https://doi.org/10.1080/01431161.2010.510811.

Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. Remote Sens. Environ. 112, 3833–3845. https://doi.org/10.1016/j.rse.2008.06.006.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. IEEE Conf. CVPR, 7482-7491. doi: 10.1109/CVPR.2018.00781.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, J., Huang, X., Gong, J., 2019. Deep neural network for remote sensing image interpretation: status and perspectives. Natl. Sci. Rev. 6, 1082–1086. https://doi.org/10.1093/nsr/nwz058.

Li, X.D., Ling, F., Foody, G.M., Ge, Y., Zhang, Y.H., Du, Y., 2017. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps. Remote Sens. Environ. 196, 293–311. https://doi.org/10.1016/j.rse.2017.05.011.

Lin, D., Lin, J., Zhao, L., Wang, Z.J., Chen, Z., 2022. Multilabel aerial image classification with a concept attention graph neural network. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2020.3041461.

Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y., 2017. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019a. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Ma, X., Wang, H., Liu, Y., Ji, S., Gao, Q., Wang, J., 2019b. Knowledge guided classification of hyperspectral image on hierarchical class tree. IEEE Int. Geosci. Remote Sens. Sym. 2702–2705 https://doi.org/10.1109/IGARSS.2019.8899885.

Mao, D., Wang, Z., Du, B., Li, L., Tian, Y., Jia, M., Zeng, Y., Song, K., Jiang, M., Wang, Y., 2020. National wetland mapping in China: A new product resulting from object-based and hierarchical classification of Landsat 8 OLI images. ISPRS J. Photogramm. Remote Sens. 164, 11–25. https://doi.org/10.1016/j.isprsjprs.2020.03.020.

Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 55, 3639–3655. https://doi.org/10.1109/TGRS.2016.2636241.

Salomonson, V.V., Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. 89, 351–360. https://doi.org/10.1016/j.rse.2003.10.016.

See, L., Bayas, J.C.L., Schepaschenko, D., Perger, C., Dresel, C., Maus, V., Salk, C., Weichselbaum, J., Lesiv, M., McCallum, I., Moorthy, I., Fritz, S., 2017. LACO-Wiki: A new online land cover validation tool demonstrated using GlobeLand30 for Kenya. Remote Sens. 9, 754. https://doi.org/10.3390/rs9070754.

Sulla-Menashe, D., Friedl, M.A., Krankina, O.N., Baccini, A., Woodcock, C.E., Sibley, A., Sun, G., Kharuk, V., Elsakov, V., 2011. Hierarchical mapping of Northern Eurasian land cover using MODIS data. Remote Sens. Environ. 115, 392–403. https://doi.org/10.1016/j.rse.2010.09.010.

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. https://doi.org/10.1016/j.rse.2018.12.013.

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 https://doi.org/10.1016/j.rse.2021.112603.

Wambugu, N., Chen, Y., Xiao, Z., Wei, M., Bello, S., Junior, J., Li, J., 2021. A hybrid deep convolutional neural network for accurate land cover classification. Int. J. Appl. Earth Obs. 103, 102515 https://doi.org/10.1016/j.jag.2021.102515.

Wang, Z., Schaaf, C.B., Sun, Q., Shuai, Y., Roman, M.O., 2018. Capturing rapid land surface dynamics with collection V006 MODIS BRDF/NBAR/Albedo (MCD43) products. Remote Sens. Environ. 207, 50–64. https://doi.org/10.1016/j.rse.2018.02.001.

Xie, S., Liu, L., Zhang, X., Yang, J., Chen, X., Gao, Y., 2019. Automatic land-cover mapping using Landsat time-series data based on Google Earth Engine. Remote Sens. 11, 3023. https://doi.org/10.3390/rs11243023.

Yang, J., Huang, X., 2021. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. Earth Syst. Sci. Data 13, 3907–3925. https://doi.org/10.5194/essd-13-3907-2021.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241, 111716.

Zhang, H.K., Roy, D.P., 2017. Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. Remote Sens. Environ. 197, 15–34. https://doi.org/10.1016/j.rse.2017.05.024.

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., Mi, J., 2021. GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery. Earth Syst. Sci. Data 13, 2753–2776. https://doi.org/10.5194/essd-13-2753-2021.

Zhao, S., Jiang, X., Li, G., Chen, Y., Lu, D., 2021. Integration of ZiYuan-3 multispectral and stereo imagery for mapping urban vegetation using the hierarchy-based classifier. Int. J. Appl. Earth Obs. 105, 102594 https://doi.org/10.1016/j.jag.2021.102594.